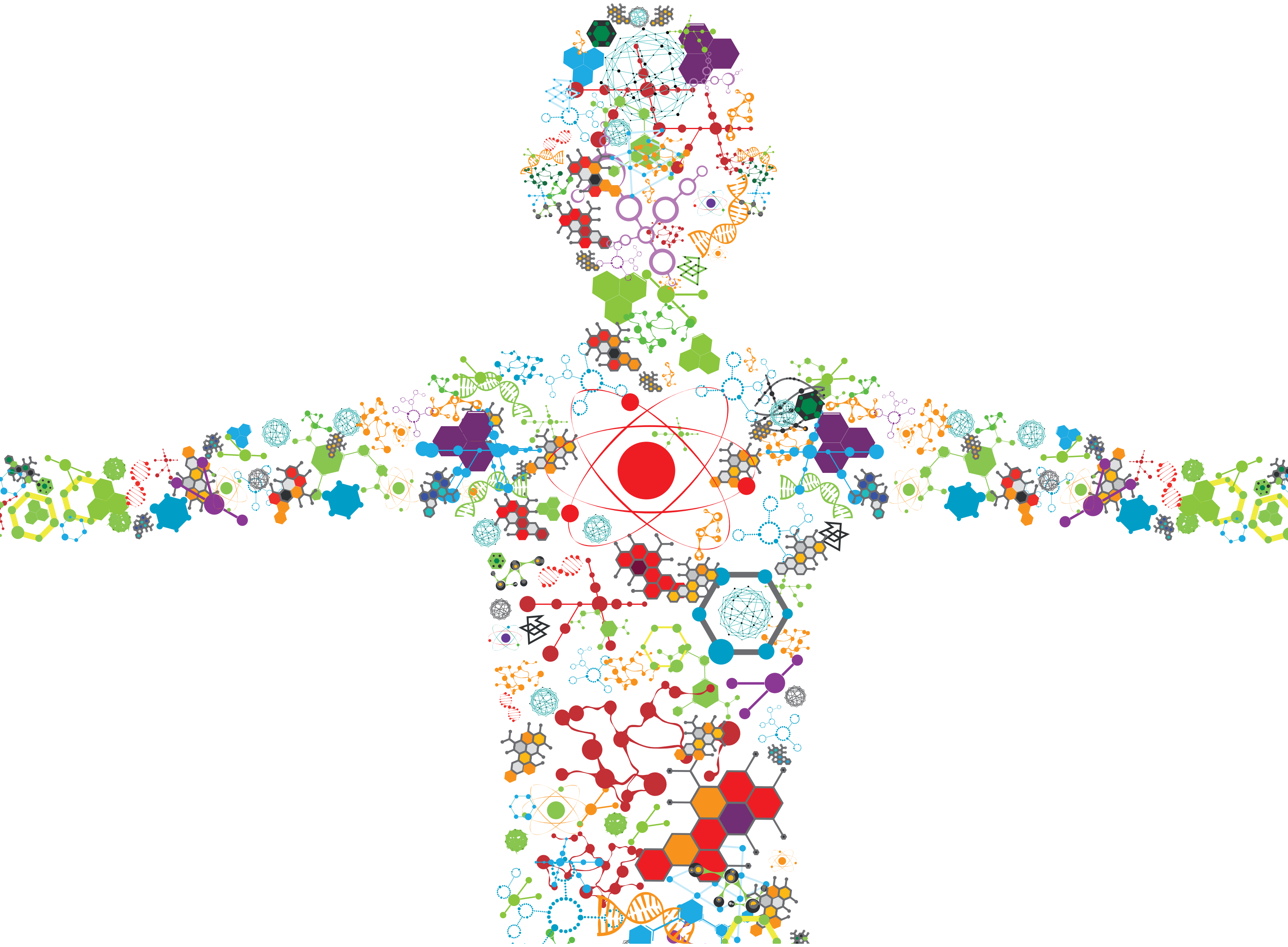
ARTIFICIAL INTELLIGENCE (AI) OPTIMIZED SYSTEMS MODELING FOR THE DEEPER UNDERSTANDING OF HUMAN CANCERS

EDITED BY: Zhiwei Ji, Shu Tao and Bing Wang

PUBLISHED IN: Frontiers in Bioengineering and Biotechnology and
Frontiers in Genetics



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-717-0
DOI 10.3389/978-2-88971-717-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ARTIFICIAL INTELLIGENCE (AI) OPTIMIZED SYSTEMS MODELING FOR THE DEEPER UNDERSTANDING OF HUMAN CANCERS

Topic Editors:
Zhiwei Ji, Nanjing Agricultural University, China
Shu Tao, UCLA Jonsson Comprehensive Cancer Center, United States
Bing Wang, Anhui University of Technology, China

Citation: Ji, Z., Tao, S., Wang, B., eds. (2021). Artificial Intelligence (AI) Optimized Systems Modeling for the Deeper Understanding of Human Cancers. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-717-0

Table of Contents

frontiers
in Bioengineering and Biotechnology

EDITORIAL
published: 11 October 2021
doi: 10.3389/fbioe.2021.756314

Editorial: Artificial Intelligence (AI) Optimized Systems Modeling for the Deeper Understanding of Human Cancers

Zhiwei Ji[1,2]*, *Shu Tao*[3] *and Bing Wang*[4]

[1]*College of Artificial Intelligence, Nanjing Agricultural University, Nanjing, China,* [2]*University of Texas Health Science Center at Houston, Houston, TX, United States,* [3]*UCLA Jonsson Comprehensive Cancer Center, Los Angeles, CA, United States,* [4]*Anhui University of Technology, Ma'anshan, China*

Keywords: multi-scale modeling, multi-omics data, network modeling, single-cell RNA-seq analysis, big clinical data analysis

Editorial on the Research Topic

Artificial Intelligence (AI) Optimized Systems Modeling for the Deeper Understanding of Human Cancers

Cancer research in the field of Computational Systems Biology attempts to address questions that will advance current knowledge in the mechanisms of cancer progression or treatment resistance. By analyzing multi-omics data and developing a predictive mathematical and/or computational model of an unknown biological system, we can systematically understand 1) the mechanisms that tie altered gene expression and downstream molecular mechanisms to functional cancer phenotypes (Colaprico et al., 2020; Menyhárt and Győrffy, 2021); 2) and/or the mechanisms that tie tumor morphology to functional cancer phenotypes (Koutsogiannouli et al., 2013; Suhail et al., 2019); 3) and/or the mechanisms that tie treatment sequence and combination to evolving functional cancer phenotypes (Yalcin et al., 2020). Currently, systems biology still faces some challenges, including model calibration, model validation and generalization, computational efficiency, and the feasibility of clinical transition (Ching et al., 2018). Recent developments in artificial intelligence technologies, e.g., deep learning (DL), allow us to model the hierarchical structure of real biological systems, efficiently converting gene-level data to pathway-level information with an ultimate impact on cell phenotype (Gazestani and Lewis, 2019). Furthermore, such computational models could require fewer training samples, are more generalizable across diverse biological contexts, and can make predictions that are more consistent with the current understanding on the inner-workings of biological systems (Brodland, 2015).

This special issue entitled "*Artificial Intelligence (AI) Optimized Systems Modeling for the Deeper Understanding of Human Cancers*" in *Frontiers in Bioengineering and Biotechnology*, and *Frontiers in Genetics* aims to provide an international forum for:

1) bringing together the greatest research efforts in cancer-specific molecular/network signature identification by integrating multi-omics/multi-level data;
2) exploring future-generation interesting and practical biomedical applications in AI, machine learning, big data sciences, knowledge-based system, *etc.*, to provide novel ideas and solutions in mathematical modeling for tumor growth, drug resistance, and targeting effect prediction;
3) addressing the real-world challenges in the fields of AI-based patient diagnosis or disease progression prediction by utilizing modern machine learning or statistical strategies, and produce a more reliable and promising application environment to develop those technologies.

Edited and reviewed by:
*Richard D. Emes,
University of Nottingham,
United Kingdom*

****Correspondence:***
*Zhiwei Ji
Zhiwei.Ji@njau.edu.cn*

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *10 August 2021*
Accepted: *13 September 2021*
Published: *11 October 2021*

Citation:
Ji Z, Tao S and Wang B (2021) Editorial: Artificial Intelligence (AI) Optimized Systems Modeling for the Deeper Understanding of Human Cancers. Front. Bioeng. Biotechnol. 9:756314. doi: 10.3389/fbioe.2021.756314

Submission for this special issue started from May 2020 and closed in Oct 2020. In nearly one and half years, we received in total 36 paper submissions. All submitted manuscripts had gone through at least two rounds of revision with reviewers in the related fields, including bioinformatics, computational biology, machine learning, and clinical study, *etc*. The final acceptance rate is 50% with 18 accepted papers in this special issue. The summaries of these papers are outlined below.

1) New bioinformatic approaches to Identify key molecular/network signatures for precision diagnosis or treatment of cancers

In the article entitled "*Identification of signatures of prognosis prediction for melanoma using a hypoxia score*" by Shou et al. The authors developed a computational method to identify the gene signatures of melanoma in hypoxic condition for prognosis prediction.

In the article entitled "*Identifying hypoxia characteristics to stratify prognosis and assess the tumor immune microenvironment in renal cell carcinoma*" by Zhang et al. The authors established a hypoxia-related risk model to predict the prognosis of patients.

In the article entitled "*Prediction of Radiosensitivity in Head and Neck Squamous Cell Carcinoma Based on Multiple Omics Data*" by Liu et al. The authors identified 12-gene signature based on multiple omics data achieved the best ability for predicting radiosensitivity in Head and Neck Squamous Cell.

In the article entitled "*An Effective Graph Clustering Method to Identify Cancer Driver Modules*" by Zhang et al. The authors proposed a graph clustering method, called "MCLCluster", to identity cancer driver modules that drive cancer progression.

In the article entitled "*Exploring the differential expression and prognostic significance of the COL11A1 gene in human colorectal carcinoma: an integrated bioinformatics approach*" by Patra et al. The authors developed an integrated bioinformatics approach to reveal the COL11A1 gene as a prognostic biomarker in colorectal carcinoma.

In the article entitled "*MicroRNA-126 Modulates Palmitate-induced Migration in HUVECs by Downregulating Myosin Light Chain Kinase via the ERK/MAPK Pathway*" by Zhu et al. The authors evaluated the effects of miR-126 on the cell migration and uncovered the underlying mechanism in HUVECs treated with palmitate.

In the article entitled "*Integrated analysis of DEAD-box helicase 56: a potential oncogene in osteosarcoma*" by Zhu et al. The authors set up a novel integrated analysis protocol and found that DDX56 is a potential therapeutic target for the treatment of osteosarcoma.

In the article entitled "*A machine learning approach for tracing tumor original sites with gene expression profiles*" by Liang et al. The authors developed a machine learning approach by integrating random forest and Naive Bayesian, to predict the primary origin sites of tumors.

In the article entitled "*A deep learning framework to predict tumor tissue-of-origin based on copy number variation*" by Liang et al. The authors proposed a deep learning framework composed of an autoencoder (AE) and a convolution neural network (CNN) to predict the primary origin sites of tumors.

In the article entitled "*TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression*" by He et al. The authors integrated somatic mutation and gene expressions t infer the primary original sites of tumor and obtained a great accuracy.

2) New studies of clinical informatics for speeding up the development of cancer diagnosis

In the article entitled "*Diagnosis of cervical cancer with parametrial invasion on whole-tumor dynamic contrast-enhanced magnetic resonance imaging combined with whole-lesion texture analysis based on T2-weighted images*" by Li et al. The authors integrated DCE-MRI images and texture analysis for diagnosis cervical cancer.

In the article entitled "*Predictive value of the texture analysis of enhanced computed tomographic images for preoperative pancreatic carcinoma differentiation*" by Zhang et al. The authors extracted 396 features from patient CT images and selected the optimal feature subset to predict the pathological degree of differentiation of pancreatic carcinoma.

In the article entitled "*RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans*" by Jin et al. The authors developed a 3D network model, RA-UNet, to precisely extract the liver region and segment tumors from the liver. Testing on public datasets show that the proposed architecture obtains competitive results.

In the article entitled "*Classification of Infected Necrotizing Pancreatitis for Surgery within or beyond Four Weeks Using Machine Learning*" by Lan et al. The authors applied machine learning models to predict the optimal timing of surgical intervention and identified the key factors associated with surgical timing for infected necrotizing pancreatitis.

In the article entitled "*Prediction of Proximal Junctional Kyphosis after Posterior Scoliosis Surgery with Machine Learning in the Lenke 5 Adolescent Idiopathic Scoliosis Patient*" by Peng et al. The authors developed a machine learning model for proximal junctional kyphosis (PJK) prognostication in Lenke 5 adolescent idiopathic scoliosis (AIS) patients undergoing long posterior instrumentation and fusion surgery.

In the article entitled "*A New Method Based on CEEMD Combined with Iterative Feature Reduction for Aided Diagnosis of Epileptic EEG*" by Peng et al. The authors proposed a computational method based on complementary ensemble empirical mode decomposition (CEEMD) combined with iterative feature reduction for aided diagnosis of epileptic EEG.

3) New strategies for optimizing the data preprocessing and quality control

In the article entitled "*Assessing the impact of data preprocessing on analyzing next generation sequencing data*" by He et al. The authors compared commonly used data preprocessing software and found differences in the detection of hotspot mutations and HLA typing. They also explained the impact of data preprocessing steps on downstream data analysis results.

In the article entitled "*RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute*

Selection and Feature Extraction" by Bian et al. The authors developed a hybrid model RF-PCA, which significantly reduce the time required for the classification, but also improved the accuracy.

The guest editors would like to thank all authors submitting their valuable works to this special section of Frontiers in Bioengineering and Biotechnology, Frontiers in Genetics, as well as all peer-reviewers for their great effort reviewing the submitted articles, providing constructive comments and suggestions and assisting the editors reaching the final decision. Special thanks will be sent to the editor-in-chief (EIC), Ranieri Cancedda and José AG Agúndez, for their precious time and valuable instructions that help us prepare and finalize this special issue.

AUTHOR CONTRIBUTIONS

ZJ coordinated the Research Topic. ZJ and BW coordinated the editorial. ZJ, ST, BW contributed to the development of the Research Topic, suggested and invited the participants, and helped with the peer review process. All authors have approved the final version of the editorial.

FUNDING

This work was supported by National Science Foundation of Jiangsu province (No. BK20211210). This work was also supported by the startup award of new professor at Nanjing Agricultural university (No. 106/804001), and National Science Foundation of Zhejiang province (No. LY20F020003).

ACKNOWLEDGMENTS

We acknowledge the contributions of all the participating authors for this Research Topic.

REFERENCES

Brodland, G. W. (2015). How Computational Models Can Help Unlock Biological Systems. *Semin. Cel Developmental Biol.* 47-48, 62–73. doi:10.1016/j.semcdb.2015.07.001

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and Obstacles for Deep Learning in Biology and Medicine. *J. R. Soc. Interf.* 15 (141), 20170387. doi:10.1098/rsif.2017.0387

Colaprico, A., Olsen, C., Bailey, M. H., Odom, G. J., Terkelsen, T., Silva, T. C., et al. (2020). Interpreting Pathways to Discover Cancer Driver Genes with Moonlight. *Nat. Commun.* 11 (1), 69. doi:10.1038/s41467-019-13803-0

Gazestani, V. H., and Lewis, N. E. (2019). From Genotype to Phenotype: Augmenting Deep Learning with Networks and Systems Biology. *Curr. Opin. Syst. Biol.* 15, 68–73. doi:10.1016/j.coisb.2019.04.001

Koutsogiannouli, E., Papavassiliou, A. G., and Papanikolaou, N. A. (2013). Complexity in Cancer Biology: Is Systems Biology the Answer?. *Cancer Med.* 2 (2), 164–177. doi:10.1002/cam4.62

Menyhárt, O., and Győrffy, B. (2021). Multi-omics Approaches in Cancer Research with Applications in Tumor Subtyping, Prognosis, and Diagnosis. *Comput. Struct. Biotechnol. J.* 19, 949–960. doi:10.1016/j.csbj.2021.01.009

Suhail, Y., Cain, M. P., Vanaja, K., Kurywchak, P. A., Levchenko, A., Kalluri, R., et al. (2019). Systems Biology of Cancer Metastasis. *Cel Syst.* 9 (2), 109–127. doi:10.1016/j.cels.2019.07.003

Yalcin, G. D., Danisik, N., Baygin, R. C., and Acar, A. (2020). Systems Biology and Experimental Model Systems of Cancer. *Jpm* 10 (4), 180. doi:10.3390/jpm10040180

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 07 April 2020
doi: 10.3389/fbioe.2020.00271

An Effective Graph Clustering Method to Identify Cancer Driver Modules

Wei Zhang[1,2], *Yifu Zeng*[1,2], *Lei Wang*[1,3]*, *Yue Liu*[4] *and Yi-nan Cheng*[5]

[1] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China, [2] Hunan Province Key Laboratory of Industrial Internet Technology and Security, Changsha University, Changsha, China, [3] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, China, [4] College of Computer Science and Electronics Engineering, Hunan University, Changsha, China, [5] College of Science, Southern University of Science and Technology, Shenzhen, China

Identifying the molecular modules that drive cancer progression can greatly deepen the understanding of cancer mechanisms and provide useful information for targeted therapies. Most methods currently addressing this issue primarily use mutual exclusivity without making full use of the extra layer of module property. In this paper, we propose MCLCluster to identity cancer driver modules, which use somatic mutation data, Cancer Cell Fraction (CCF) data, gene functional interaction network and protein-protein interaction (PPI) network to derive the module property on mutual exclusivity, connectivity in PPI network and functionally similarity of genes. We have taken three effective measures to ensure the effectiveness of our algorithm. First, we use CCF data to choose stronger signals and more confident mutations. Second, the weighted gene functional interaction network is used to quantify the gene functional similarity in PPI. The third, graph clustering method based on Markov is exploited to extract the candidate module. MCLCluster is tested in the two TCGA datasets (GBM and BRCA), and identifies several well-known oncogenes driver modules and some modules with functionally associated driver genes. Besides, we compare it with Multi-Dendrix, FSME Cluster and RME in simulated dataset with background noise and passenger rate, MCLCluster outperforming all of these methods.

Keywords: driver modules, mutual exclusivity, connectivity, functionally similarity, Markov clustering

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Yuhua Yao,
Hainan Normal University, China
Xing Chen,
China University of Mining and Technology, China

****Correspondence:***
Lei Wang
wanglei@xtu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *17 February 2020*
Accepted: *16 March 2020*
Published: *07 April 2020*

Citation:
Zhang W, Zeng Y, Wang L, Liu Y and Cheng Y (2020) An Effective Graph Clustering Method to Identify Cancer Driver Modules.
Front. Bioeng. Biotechnol. 8:271.
doi: 10.3389/fbioe.2020.00271

INTRODUCTION

Cancer research has shown that gene mutation can disrupt specific cellular pathways that drive cancer development (Weinstein et al., 2013). Recently, the rapid development of next-generation sequencing technologies has increased the generation and availability of high-resolution data related to cancer, providing opportunities for the study of cancer genomes (Wood et al., 2007; Cancer Genome Atlas Research, 2008; Tomczak et al., 2015; Zhao et al., 2019). The key task of cancer genomes research is to identify the molecular mutations or drivers. Functionally related driver mutations in the genome, also known as driver modules or pathways, activate the mechanisms by which cancer occurs, triggering cancer, driving cancer progression and giving cancer cells a selective advantage.

Some computational methods and mathematical models have been developed to detect driver gene sets, pathways and modules by using large-scale sequencing data (Hou et al., 2016; Zheng et al., 2016; Yang et al., 2017; Xi et al., 2018; Ahmed et al., 2019; Deng et al., 2019; Zhang and Wang, 2019a; Pelegrina et al., 2020). Existing research show that the members of cancer driver modules often exhibit specific mutation patterns in cancer samples, the most significant characteristic is mutual exclusivity (mutex) which means once one member mutates, the tumor will gain a significant selection advantage, while later mutations in other members will not give the tumor a selection advantage. Most current methods use only mutex to derive the driver pathway or modules, the other properties of the module are not fully considered, such as functionally similarity of members within a module.

Recently, two types of methods for identifying driver modules or gene sets have been proposed: De novo and knowledge-based methods. The De novo methods usually exploit two characteristics from somatic mutation data: high coverage and mutex (Dees et al., 2012; Vandin et al., 2012; Zhao et al., 2012; Babaei et al., 2013; Leiserson et al., 2013; Paull et al., 2013; Jia et al., 2014; Deng et al., 2019; Zhang and Wang, 2019a,b; Dees et al., 2012; Vandin et al., 2012; Zhao et al., 2012; Babaei et al., 2013; Leiserson et al., 2013; Paull et al., 2013; Jia et al., 2014; Deng et al., 2019; Zhang and Wang, 2019a,b). High coverage means that the driver modules or driver pathway covers a large number of samples. Mutex represents that one of driver gene mutations in a pathway are sufficient to interfere with the pathway. For example, Dendrix (Vandin et al., 2012) identifies driver pathways with high coverage and mutex by transforming the problem into a maximum exclusive sub-matrix. MDPFinder (Wu et al., 2015), Multi-dendrix (Leiserson et al., 2013), ComMDP, and SpeMDP (Zhang and Zhang, 2016) figure out the maximum exclusion sub-matrix problem by utilizing the integer linear programming, focus on identifying mutex gene sets. On the other hand, the knowledge-based approaches, in addition to somatic mutation data, other network- and functional phenotype-based data are combined to detect driver pathway or modules (Hua et al., 2013; Babur et al., 2015; Kim et al., 2015; Leiserson et al., 2015; Nambara et al., 2015; Wang et al., 2015; Reyna et al., 2018; La Vecchia and Sebastian, 2020). These approaches can be subdivided according to the optimization objectives in the computational problem, and they are used to define cancer driver modules identification problems. In the methods of Hotnet (Network, 2012), Hotnet2 (Leiserson et al., 2014), Hierarchical Hotnet (Reyna et al., 2018), thermal diffusion is a common feature. Diffusion values are used to extract modules with high connectivity, which are defined by graph connectivity (usually strong connectivity). Other methods, such as MEMo (Ciriello et al., 2012), RME (Leiserson et al., 2015)and FSME Cluster (Liu et al., 2017), use the interaction network and function relation graph to derive the largest group in the similarity graph, and derive the group with largest mutex. Babur et al. (Babur et al., 2015) proposed a seed growth-based method in the network, which uses TCGA data to identify pan-cancer modules, and the method determines the growth strategy based on mutex scores. Dao et al. (Dao et al., 2017) proposed an ILP method, which combined the definition of interaction density and mutex in the module as the optimization target. MEMCover (Kim et al., 2015) and MEXCOwalk (Ahmed et al., 2019) combined mutation data with interaction data to detect mutually exclusive mutant genomes in the same or different tissues.

In this work, we get inspired by these existed methods and present a novel knowledge-based method to identify cancer driver modules (MCLCluster), which combines mutex, functional similarity and connectivity in PPI network, multiple data type is used. Before we compute the mutex, the Cancer Cell Fraction (CCF) is aided to select stronger signals and more confident mutations, then the weighted gene functional interaction network is used to quantify the gene functional similarity in PPI, exploit graph clustering method based on Markov to extract the candidate module. The similarity measure between a pair of genes is defined as PPI network edge weight through taking into account functional similarity and mutex. Cluster filter and permutation test is used to test which cluster to be driver modules. We compare it with those of three representative approaches [Multi-Dendrix (Leiserson et al., 2013), FSME Cluster (Liu et al., 2017), and RME (Leiserson et al., 2015)] on simulated dataset with background noise, MCLCluster outperform all of these methods. Unlike most of presented approaches to discover driver modules with mutually exclusive between all gene pairs, MCLCluster does not necessarily identify complete exclusivity gene pair, but uses other functional similarity information to complement interaction data for a better identification of modules.

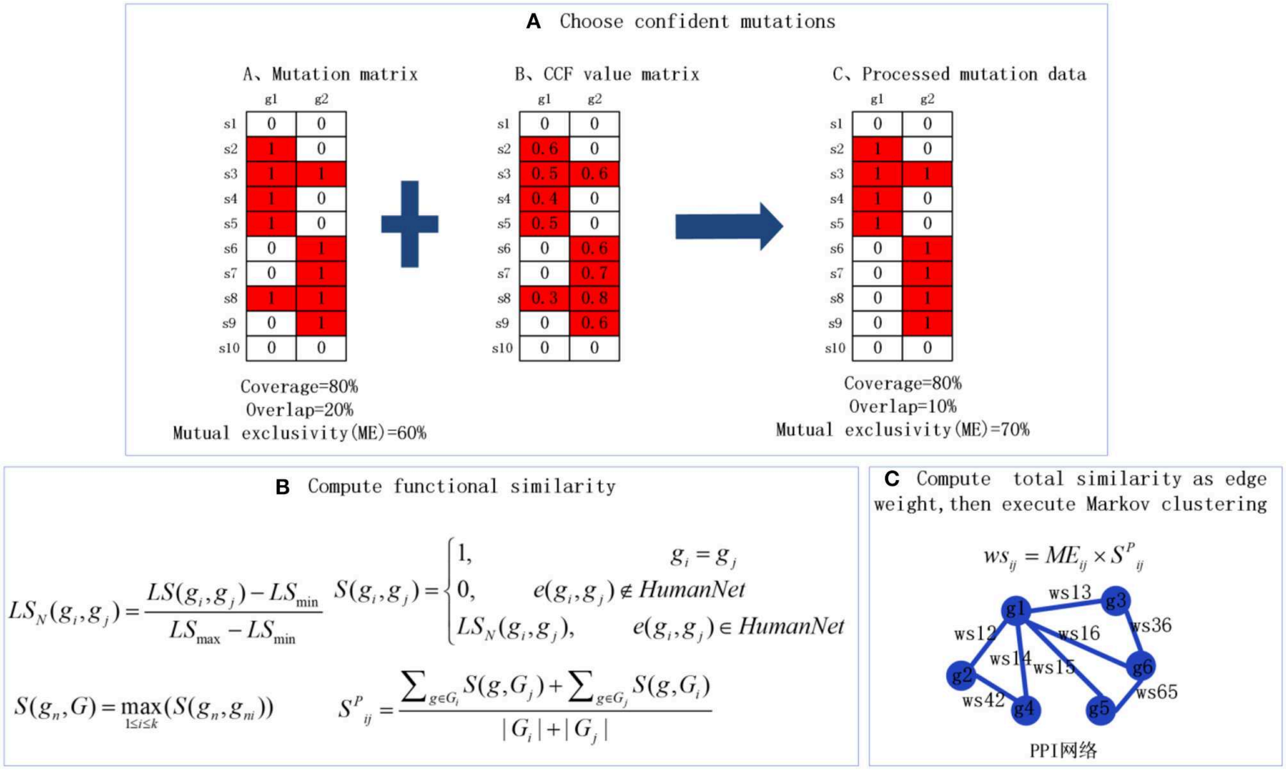
METHODS

The identification of the cancer driver modules based on graph clustering (MCLCluster) is introduced in detail. The schematic flowchart is shown in **Figure 1**.

Datasets

GBM and BRCA datasets which including CNVs and SNVs mutational data are used for testing, which are downloaded from cBioPortal (Cerami et al., 2012). The GBM dataset contains 550 samples, 1,376 mutant genes, and the BRCA dataset contains 1078 samples, and 1463 mutant genes. We combine non-binary data (CCF) to provide more information and prioritize more important mutations (ie, earlier mutations with larger CCF values). The CCF value indicates the proportion of cancer cells in the mutant sample. CCF data is extracted from read count data (Roth et al., 2014). PPI network are derived from Multinet (Khurana et al., 2013), which contains 109599 interactions between 14445 genes.

In order to verify the reliability, we produce various simulation data with random passenger rate and background noise, and the execution of the entire simulation process use the algorithm in RME. MCLCluster is compared with Multi-Dendrix, FSME Cluster and RME in simulation data. Each simulation datasets contains 500 patients and 200 mutant genes. Mutation noise is achieved by converting a value with opposite values (0 for 1 or 1 for 0) in different probability ranges of 0.05 to


A Choose confident mutations
A、Mutation matrix
B、CCF value matrix
C、Processed mutation data
Coverage=80%
Overlap=20%
Mutual exclusivity(ME)=60%
Coverage=80%
Overlap=10%
Mutual exclusivity(ME)=70%
B Compute functional similarity
C Compute total similarity as edge weight, then execute Markov clustering
ws13
ws12
ws16
ws36
ws14
ws15
ws42
ws65
g1
g2
g3
g4
g5
g6
PPI网络

A、Mutation matrix

	g1	g2
s1	0	0
s2	1	0
s3	1	1
s4	1	0
s5	1	0
s6	0	1
s7	0	1
s8	1	1
s9	0	1
s10	0	0

B、CCF value matrix

	g1	g2
s1	0	0
s2	0.6	0
s3	0.5	0.6
s4	0.4	0
s5	0.5	0
s6	0	0.6
s7	0	0.7
s8	0.3	0.8
s9	0	0.6
s10	0	0

C、Processed mutation data

	g1	g2
s1	0	0
s2	1	0
s3	1	1
s4	1	0
s5	1	0
s6	0	1
s7	0	1
s8	0	1
s9	0	1
s10	0	0

$$LS_N(g_i,g_j)=\frac{LS(g_i,g_j)-LS_{\min}}{LS_{\max}-LS_{\min}}\qquad S(g_i,g_j)=\begin{cases}1, & g_i=g_j\\ 0, & e(g_i,g_j)\notin HumanNet\\ LS_N(g_i,g_j), & e(g_i,g_j)\in HumanNet\end{cases}$$

$$S(g_n,G)=\max_{1\le i\le k}(S(g_n,g_{ni}))\qquad S^P{}_{ij}=\frac{\sum_{g\in G_i}S(g,G_j)+\sum_{g\in G_j}S(g,G_i)}{|G_i|+|G_j|}$$

$$ws_{ij}=ME_{ij}\times S^P{}_{ij}$$

FIGURE 1 | The overview of MCLCluster. **(A)** Integrate CCF data to choose stronger signals and more confident mutations, and compute the mutex of each gene pairs. **(B)** The weighted gene functional interaction network is used to quantify the gene functional similarity in PPI. **(C)** Compute total similarity as edge weight, then execute Markov clustering to extract candidate module.

0.11. The remaining genes are considered to be passenger genes and the probability of their mutation uses empirical values.

Similarity Measure

In order to consider the module property on mutex, functional similarity and connectivity in the PPI network, and to facilitate subsequent graph clustering, we define the edge weights of the PPI network as the product of mutex and functional similarity between gene pair.

Functional Similarity

Actually, most of the existing methods widely use cosine coefficient to measure the functional similarity between entities in PPI network, which only consider the network structure and it is too simple to as a functional similarity measurement. So we develop a new metric to measure the entities similarity in PPI with the help of the **w**eighted **g**ene **f**unctional **i**nteraction **n**etwork (**wgfin**), which is downloaded from HumanNet. We use the correlated log-likelihood scores (LS) as a metric of the interaction strength between any two genes in **wgfin**. $LS_N(g_i, g_j)$ represents the normalized value between gene *i* and gene *j* when $LS(g_i, g_j)$ is normalized using min-max normalization, the detail is:

$$LS_N\left(g_i,g_j\right)=\frac{LS\left(g_i,g_j\right)-LS_{min}}{LS_{max}-LS_{min}}\tag{1}$$

Here LS_{min} denotes the minimal *LS* and LS_{max} denotes the maximal *LS* in **wgfin**. As a result, the similarity $S\left(g_i, g_j\right)$ between any two genes that have edges in **wgfin** is calculated:

$$S\left(g_i,g_j\right)=\begin{cases}1, & g_i{=}g_j\\ 0, & e(g_i,g_j)\notin HumanNet\\ LS_N\left(g_i,g_j\right), & (g_i,g_j)\in HumanNet\end{cases}\tag{2}$$

Here $e\left(g_i, g_j\right)$ represents the edge between gene *i* and gene *j*. Then, the similarity of gene g_n and gene set $G = \left\{g_{n1}, g_{n2}, \ldots, g_{np}\right\}$ is calculated as follows:

$$S\left(g_n,G\right)=max_{1\le i\le p}(S\left(g_n,g_{ni}\right))\tag{3}$$

At last, according to the BMA (Best-Match Average) method (Wang et al., 2007; Xiao et al., 2018), the functional similarity of pg_i and pg_j in the PPI network is defined. The detail is as follows:

$$S^P_{ij}=\frac{\sum_{g\in G_i}S(g,G_j)+\sum_{g\in G_j}S(g,G_i)}{|G_i|+|G_j|}\tag{4}$$

Here G_i and G_j respectively denote the a set of gene connected to pg_i and pg_j, and $|G|$ denotes the number of genes in G.

Mutual Exclusivity (Mutex)

To choose stronger signals and more confident mutations, we combine the CCF matrix to process somatic mutation. For each gene, we perform two operations, the one is to delete the mutation with the lowest CCF value, and the other is to delete one mutation when the CCF difference between the two mutations is less than a certain parameter ε (obtain through multiple experiments, usually small than coverage). In this paper, overall consider weighing algorithm efficiency and number of modules, we set the parameter $\varepsilon = 0.1$. The somatic mutation matrix A is filtered by CCF matrix, then it will be used to compute mutex, and the detail of each entry is listed as:

$$\boldsymbol{A_{ab}} = \begin{cases} \mathbf{1}, & \boldsymbol{if\ sample\ a\ mutated\ in\ a\ gene\ b\ and\ it\ CCF} \\ & \boldsymbol{value\ meet\ condition} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad (5)$$

In general, mutations between member genes in a driver module appear to be mutually exclusive. The previous work (Vandin et al., 2011) proposed that a pathway or module is a group of genes characterized by high coverage and low coverage overlap. Coverage represents the patient proportion with at least one gene mutation in a group of gene, and coverage overlap is equal to the patient proportion with more than two gene mutations in a group of gene. The mutex is expressed as:

$$\boldsymbol{ME}\left(\boldsymbol{se}\right) = \boldsymbol{C}\left(\boldsymbol{se}\right) - \boldsymbol{O}\left(\boldsymbol{se}\right) \quad (6)$$

Where ME denotes mutex, se denotes the genes sets, C denotes coverage and O denotes coverage overlap. Here, we calculate the pairwise and group mutex. Pairwise mutex genes help identify all gene pairs which are may take part in the same module, and the group mutex is applied to compute the mutex of all genes in one module. An example in **Figure 1A** shows the computation of coverage, coverage overlap and mutex.

Then combine these two properties (functional similarity and mutex) to calculate the total similarity as the edge weight of the PPI network:

$$\boldsymbol{ws}\left(\boldsymbol{pg_i},\boldsymbol{pg_j}\right) = \boldsymbol{ME}\left(\boldsymbol{pg_i},\boldsymbol{pg_j}\right) \times \boldsymbol{S^P_{pg_i pg_j}} \quad (7)$$

Candidate Module Extraction

Here, we apply Markov clustering (MCL) to identify clusters in the PPI network appling the total similarity matrix ws derived by Equation (7). Markov clustering is an effective biological network clustering algorithm, which is widely used for the identification of functional modules (Brohee and van Helden, 2006; Vlasblom and Wodak, 2009; Shih and Parthasarathy, 2012). After executing the clustering, closely functional related genes will be grouped into the same cliques, which are as candidate modules and will be used for follow-up modules refinement.

The $GR = (N_p, \epsilon_p)$ denotes the undirected graph in the PPI network, in which N_p represents node sets and ϵ_p represents edge set. $pg_i \in N_p$ represents the i-th gene, and $ws\left(pg_i, pg_j\right)$ is the edge weight of $\left(pg_i, pg_j\right)$, $ws\left(pg_i, pg_j\right) > 0$ indicate that pg_i interact with pg_j in the PPI network, $ws\left(pg_i, pg_j\right) = 0$ indicate they are not interaction. $P \in \mathbb{R}^{|N_p|\times|N_p|}$ denotes $GR's$ adjacency matrix, the initialization of P is:

$$P\left(\boldsymbol{i},\boldsymbol{j}\right) = \begin{cases} \boldsymbol{ws}\left(\boldsymbol{pg_i},\boldsymbol{pg_j}\right) & \textbf{if } (\boldsymbol{pg_i},\boldsymbol{pg_j}) \in \epsilon_{\boldsymbol{p}} \\ \textbf{ws}\left(\boldsymbol{pg_i},\boldsymbol{pg_k}\right) & \boldsymbol{if}\ (\boldsymbol{pg_i} = \boldsymbol{pg_j}) \\ \mathbf{0} & \textbf{otherwise} \end{cases}, \quad \boldsymbol{k} \in [\mathbf{1}, \left|\boldsymbol{N_p}\right|] \quad (8)$$

The matrix p can holds the transition probabilities of the Markov chain defined on GR. $p\left(i, j\right)$ denotes the transition probability from pg_i to pg_j. Normalize the matrix P as follow:

$$\tilde{\boldsymbol{P}}\left(\boldsymbol{i},\boldsymbol{j}\right) = \frac{\boldsymbol{P(i,j)}}{\sum_{\boldsymbol{k=1}}^{|\boldsymbol{N_p}|} \boldsymbol{p(k,j)}} \quad (9)$$

Markov clustering contains two processes, which are known as 'Expand' and 'Inflate'. When execute the operation process, the 'Expand' and 'Inflate' respectively are iteratively assigned to the stochastic matrix. The calculation formula of the Expand operation is:

$$\boldsymbol{P}_{\exp} = \tilde{\boldsymbol{P}} * \tilde{\boldsymbol{P}} \quad (10)$$

The inflation parameter rp is used in Inflate process to raise each entry in the matrix $\tilde{p}$. The Inflate process can expand the unevenness of each column. That is to say, flows increase where they are already powerful and decrease when they are weak. The Inflate process is expressed like Equation (9):

$$\boldsymbol{P}_{\inf}\left(\boldsymbol{i},\boldsymbol{j}\right) = \frac{\tilde{\boldsymbol{P}}\boldsymbol{(i,j)}^{\boldsymbol{rp}}}{\sum_{\boldsymbol{k=1}}^{|\mathbf{N_p}|} \tilde{\boldsymbol{P}}\boldsymbol{(k,j)}^{\boldsymbol{rp}}} \quad (11)$$

Markov clustering starts from the matrix **P**, and iteratively uses the Expand and Inflate until convergence. After convergence, there is one non-zero value in each column of the final matrix, and those non-zero value in the same row form a node cluster, we can get them as the candidate modules.

Modules Refinement and Mutex Significant Test

Not all of the clusters (candidate driver modules) obtained by graph clustering can be used as driver modules, nor are all genes in a population members of the module, because it is difficult to obtain the exact size of the module number. Therefore, perform the permutation test on each cluster to evaluate the importance of mutex. However, testing only on the largest cluster may result in the loss of potential subgroups which may pass the test. In order to solve this problem, (Ciriello et al., 2012) proposed the following steps to filter the genes and compute the mutex of the subgroups while limiting the subgroup size. Given a candidate module C containing the r gene, if a significant p value is observed, we will retain the module C, and not consider compute the mutex of all its subgroups. Or else, we list all subgroups of r-1 size, for each member belongs to the C, and executes a permutation test on each subgroup to get a p value. It repeats recursively until one of these two conditions is met (Ciriello et al., 2012): a subgroup is significantly mutually exclusive or $r = 3$ (min_module_size is 3). After testing, only the cluster that gets the most significant p value is reserved as the driver module.

TABLE 1 | Results of GBM.

No	Driver modules	Gene number	ME (Exclusivity)	*P*-value	$\overline{ws}$
1	CDKN2B CDK4 RB1 ERBB2	4	76%	0	0.834
2	TP53 MDM2 MDM4	3	82%	0.001	0.766
3	PTEN PIK3R1 NF1 EGFR	4	78%	0.001	0.741

$\overline{ws}$ is the average value of ws (The sum of the similarities between the pairs divided by the gene number), and ws is the total similarity calculated by Equation (7).

Evaluating Performance

To compare the performance, F1 score is used for evaluating the power of the identification module. F1 score expressed the trade-off between accuracy (abbreviated to Pr) and recall (abbreviated to Re), which can be computed using true positive (abbreviated to TP), false positive (abbreviated to FP), and false negative (abbreviated to FN). The details are:

$$\boldsymbol{Pr} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FP})}, \boldsymbol{Re} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FN})}, \boldsymbol{F1} = \frac{\boldsymbol{2 \bullet Pr \bullet Re}}{\boldsymbol{Pr}+\boldsymbol{Re}} \qquad (12)$$

RESULTS

GBM

We apply MCLCluster to GBM dataset, 3 important driver modules are identified, the detailed information of them are listed in **Table 1**. The interaction among genes within GBM modules are list in **Figure 2**. All the genes in these 3 modules are well-known in the GBM research, they are members of the 3 important signaling pathways and their mutation samples are more than five percent.

The first module contains the mutation of ERBB2, CDK4, and CDKN2B, RB1. The mutation of these four genes cover 78% of the samples, and average functional similarity is 0.834, indicate that the genes in module have similar function. The *p-value* calculated by the permutation test is equal to 0, indicate that the module has significant mutex. Three of these genes (except ERBB2) are from the RB signaling pathway that related to G1/S progression. CDKN2B inhibits CDK4, CDK4 inhibits RB1. CDKN2B and RB1 are core members of the cell cycle and cell cycle mitosis, the over expression of ERBB2 made the proliferation activation, and CDK4 has a strong interaction as a negative regulator of normal cell proliferation (Porta-Pardo et al., 2015; Tang et al., 2016).

The second module includes the mutation of MDM2, MDM4 and TP53. Most of the MDM2 mutation is amplified in the sample. TP53 is an important tumor suppressor gene which is the most common mutant gene in GBM samples. The module is mutated in 85% of the samples, the mutex of the module is 82%, and average functional similarity is 0.766, indicate that the genes in module have similar function, the *p-value* calculated by the permutation test is equal to 0.001, indicate that the module has significant mutex. All the members of this module are well-known members of the p53 signaling pathway (Kim et al., 2015), which is a key and frequently mutated pathway in GBM related to aging and apoptosis (Ciriello et al., 2012). This module contains 3 mutually exclusive gene pairs (all of which are significant), and no gene pair mutates simultaneously (Babur et al., 2015).

The third module consists of deletion of PTEN, the mutation of PIK3R1, NF1, and EGFR. Deletions in PTEN have been linked to the proneural subtype of GBM. Mutations in EGFR and NF1 related to the classical GBM subtype, in addition to the PIK3R1 appearing in the GBM pathway of (Greenman et al., 2007), it has been previously reported to be related to many human cancers (Vandin et al., 2012). The module is mutated in 82% of the samples, the mutex of the module is 78%, and average functional similarity is 0.741, indicate that the genes in module have similar function, the *p-value* calculated by the permutation test is equal to 0.001, indicate that the module has significant mutex. All the members of this module are core members of RTK/RAS/PI(3)K signaling pathway.

BRCA

We apply MCLCluster to BRCA dataset, 4 driver modules are identified, the detailed information of them are listed in **Table 2**. The interaction among genes within BRCA modules are list in **Figure 3**. Most of the genes in these 4 modules are core members of the 4 signaling pathways (p53 signaling, PI(3)K/AKT signaling, ERBB signaling pathway and RB signaling pathway). They are well-known in the BRCA research and their mutation samples are more than five percent. Compared with GBM, these 4 modules cover a smaller percentage of samples, indicate that the mutation heterogeneity or disease heterogeneity of the breast cancer dataset is greater.

The first module contains the mutation of PIK3CA, PIK3R1, AKT1, PTEN. The mutation of these four genes cover 75% samples, and average functional similarity is 0.824, indicate that the genes in module have similar function. The *p-value* calculated by the permutation test is equal to 0, suggesting that the module has significant mutex. All genes in this module are core members of PI(3)K/AKT signaling pathway. AKT1 interact with PTEN, PIK3R1, and PIK3CA, PTEN inhibits PIK3CA and PIK3R1 (Wu et al., 2015; Mandal and Ma, 2016).

The second module includes TRPS1, ZNF217 and FBXO31 gene mutations. The mutation of these 3 genes cover 89% samples, and average functional similarity is 0.811, indicate that the genes in module have similar function. The p-value calculated by the permutation test is equal to 0, suggesting that the module has significant mutex Two third of genes are members of the ERBB signaling pathway, which is an important breast cancer-related pathway. TRPS1 is a common oncogene that plays an important role in controlling cell cycle during breast cancer (Wu et al., 2014). ZNF217 is proved to be a central role in cancer development, and FBXO31 is proved to be a candidate tumor suppressor gene, by generating Skp Cullin F-box containing SCF

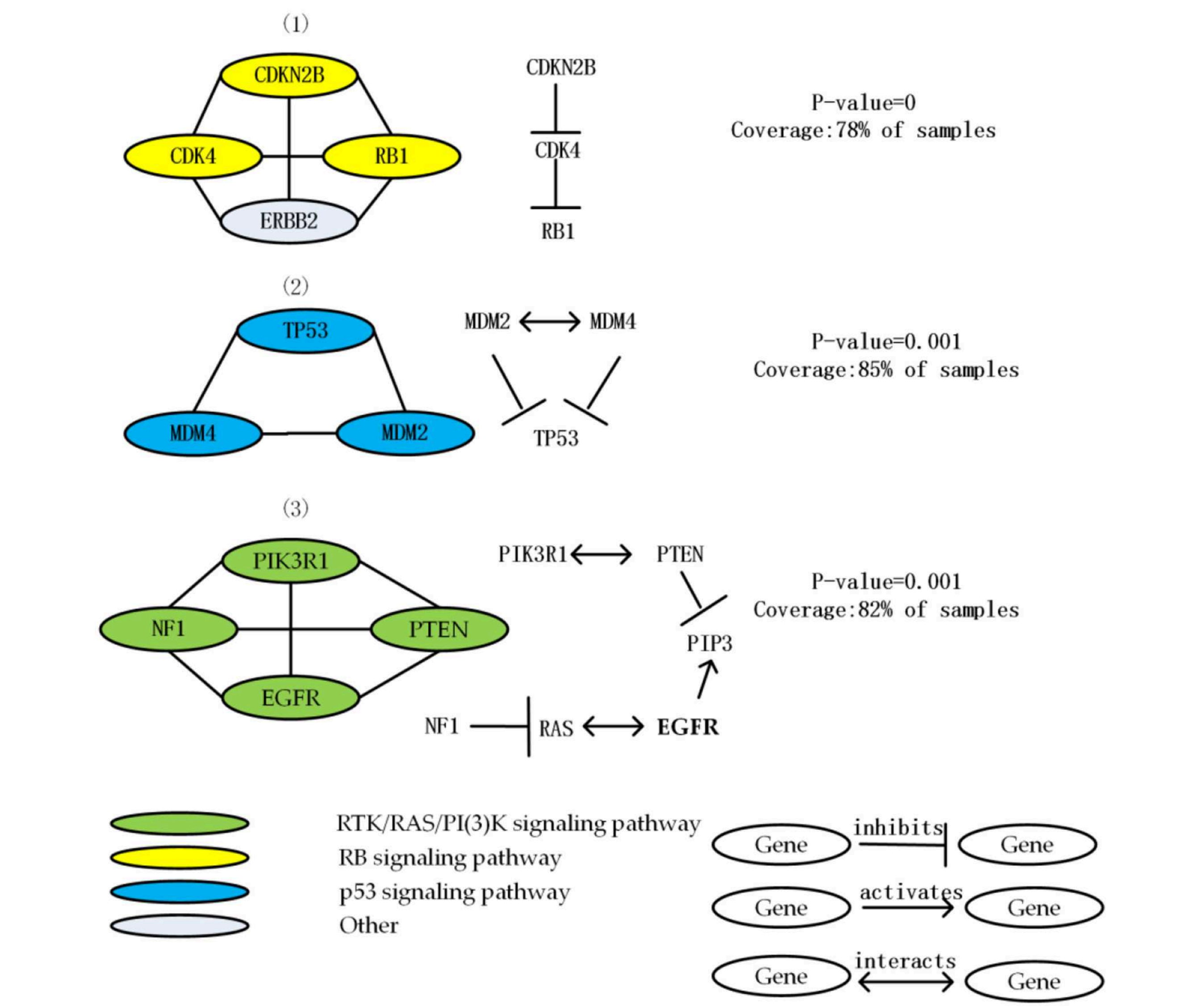


FIGURE 2 | List 3 driver module and the interaction among genes in each driver module in the GBM data. Node color shows the role of GBM in different signal pathways.

TABLE 2 | Results of BRCA.

No	Driver modules	Gene number	ME (Exclusivity)	*P*-value	$\overline{ws}$
1	PTEN PIK3CA PIK3R1 AKT1	4	72%	0	0.824
2	TRPS1 ZNF217 FBXO31	3	74%	0	0.811
3	TP53 CDH1 MYC	3	80%	0.001	0.721
4	FBXO31 RB1 CCDN1	3	70%	0.001	0.714

$\overline{ws}$ is the average value of ws (The sum of the similarities between the pairs divided by the gene number), and ws is the total similarity computed by Equation (7).

complex, it causes cell senescence and has consistent tumor suppressor attributes (Kumar et al., 2005). FBXO31 inhibits TRPS1 and ZNF217.

The third module contains mutations in TP53, CDH1, MYC. The mutation of these 3 genes cover 82% samples, and average functional similarity is 0.721, indicate that the genes in module have similar function. The *p-value* calculated by the permutation test is equal to 0.001, suggesting that the module has significant mutex. Two third of genes are core members of the p53 signaling pathway. Loss or down-regulation of the Ecadherin gene CDH1 at 16q22.1 is associated with breast cancer proliferation and invasion, MYC is an effective tumorigenic

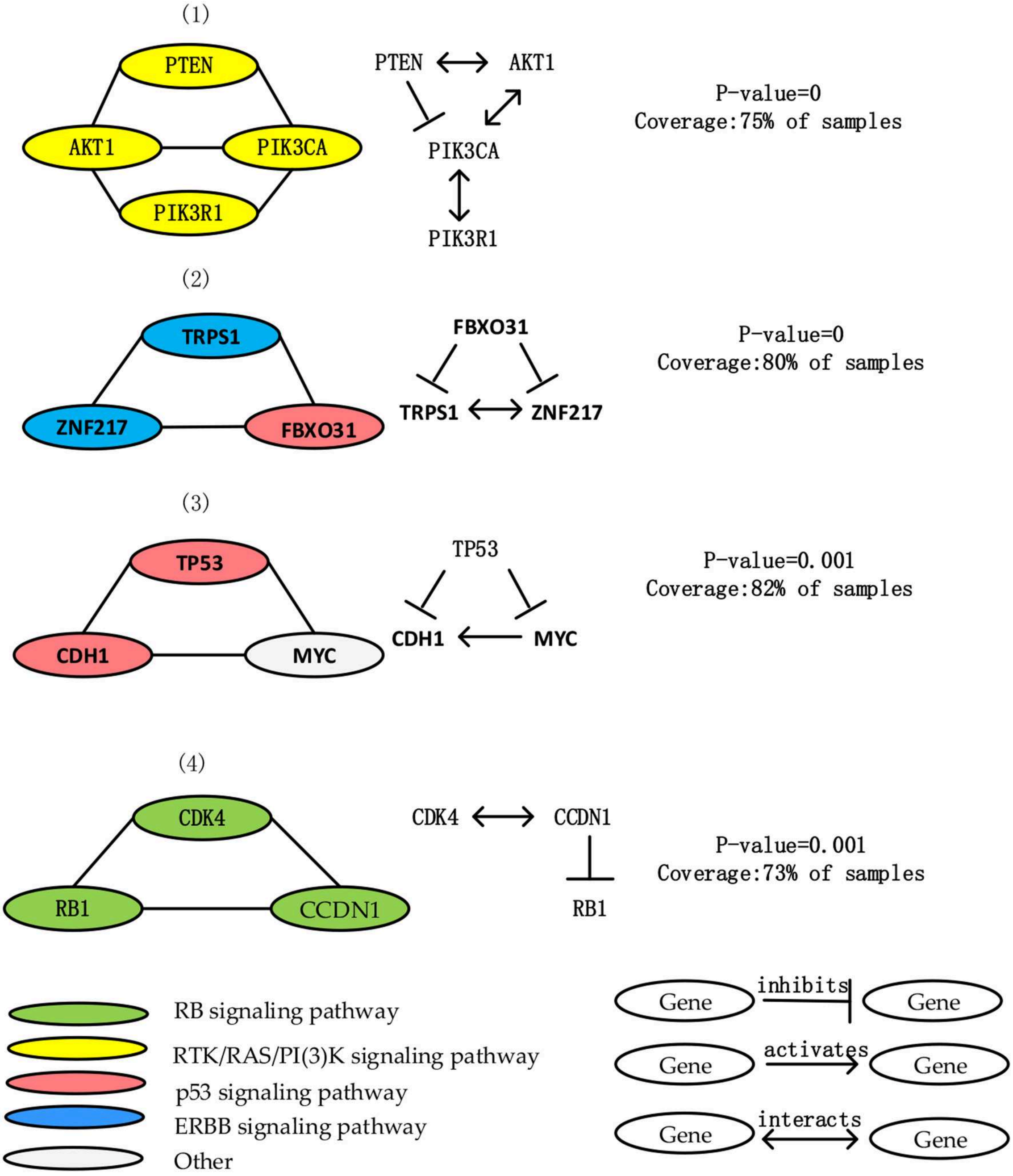


FIGURE 3 | List 4 driver module and the interaction among genes in each driver module in the BRCA data. Node color shows the important role of BRCA in different signal pathways.

activator, a transcription factor, and a key regulator of cell growth, differentiation, and apoptosis (Amgalan and Lee, 2015; Nangalia et al., 2015).

The forth module contains mutations in CCND1, RB1 and CDK4. The mutation of these three genes cover 73% samples, and average functional similarity is 0.714, indicate that the genes in module have similar function. The *p-val*ue calculated by the permutation test is equal to 0.001, suggesting that the module has significant mutex. All of genes in this module are important members of the RB signaling pathway. CDK4 interacts with CCND1, CCND1 inhibits RB1. CCND1 and RB1 encode interact proteins that have an important effect in cell cycle (Placke et al., 2014). CCND1 encodes the cyclind1 protein, it affect the retinoblastoma protein which encoded through overphosphorylation by RB1 (Rozenchan et al., 2014). Hyperphosphorylation of RB inactivates its role as a tumor suppressor gene, so mutations targeting CCND1 or RB1 are of great significance for tumor proliferation (Salgia et al., 2017).

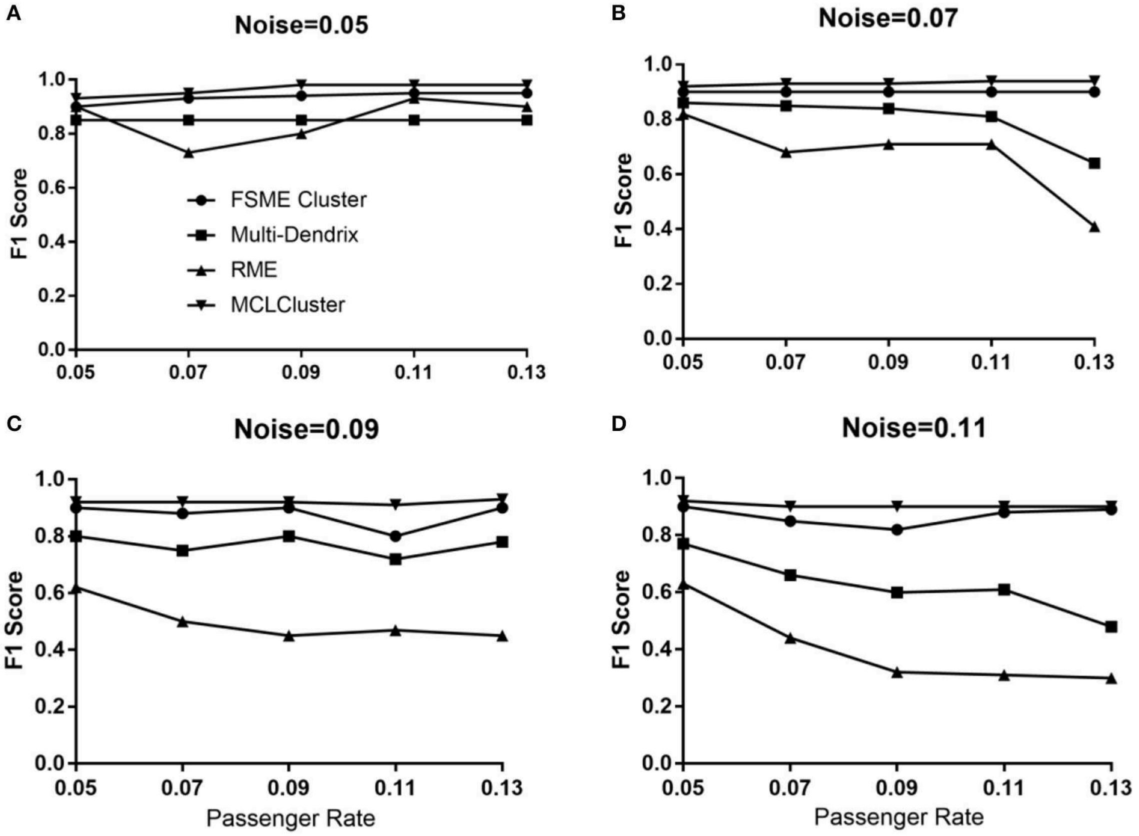


FIGURE 4 | The F1 score of MCLCluster, Multi-Dendrix, FSME Cluster and RME in simulation data for 1 module. **(A)** When noise = 0.05, the F1 score of the four methods with different passenger rate. **(B)** When noise = 0.07, the F1 score of the four methods with different passenger rate. **(C)** When noise = 0.09, the F1 score of the four methods with different passenger rate. **(D)** When noise = 0.11, the F1 score of the four methods with different passenger rate.

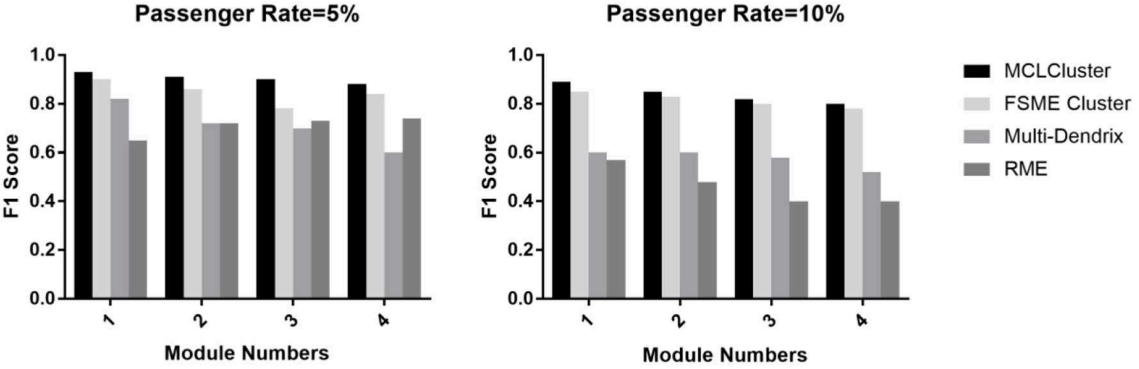


FIGURE 5 | The F1 score of MCLCluster, Multi-Dendrix, FSME Cluster and RME in simulation data for multiply modules.

Simulated Data

Identifying Top One Module

To comparing the four methods (MCLCluster, Multi-Dendrix, FSME Cluster and RME), we generated simulation samples considering two parameters (passenger rate and background noise). The Multi-Dendrix need to input the module size, and it is difficult to obtain, so considering fairness, Multi-Dendrix is applied three times for each data set, the module sizes are set to three, four, and five, respectively. The remaining parameter used in other three approaches is set to the default value. By default, MCLCluster will identify multiple modules, the module with the highest *ws* and the lowest *p-value* will be selected. It's worth noting that in simulation experiment, we cannot consider the CCF value.

As shown in **Figure 4**, when the noise is 0.05, the four methods all achieve high F1 score under different passenger rates. Among them, MCLCluster received F1 scores above 0.94. In general, when the noise is greater than 0.07, the F1 scores decrease with the increase of passenger rate in Multi-Dendrix and RME. In addition, when noise and passenger rates all greater than or equal to 0.09, the F1 scores of RME are all less than 0.6. MCLCluster and FSME Cluster also faces a decline in F1 score, when the noise is greater than 0.09. MCLCluster have better performance than the others in all cases, which shows that MCLCluster have a strong ability to detect mutually exclusive drive modules. Compared with the other three methods, under different noise environments, as the passenger rate increases, the MCLCluster shows good stability.

Identifying Multiply Modules

We identify one to four modules to compare MCLCluster, Multi-Dendrix, FSME Cluster and RME. The passenger rate is set to 0.05 and 0.10, and the module noise is set to 0.10. We can see from **Figure 5**, the F1 scores of the four methods have a slight downward trend. When the passenger rate is 0.05, the RME showed a high F1 score relative to Multi-Dendrix in most cases, and when the passenger mutation rate increased to 0.10, Multi-Dendrix performed better than RME. The MCLCluster can outperform all other methods in any cases, both the increased module numbers and the two different passenger rates.

CONCLUSIONS AND DISCUSSIONS

We develop a new approach named MCLCluster, which uses somatic mutation data, Cancer Cell Fraction (CCF) data, gene functional interaction network and protein-protein interaction (PPI) network to detect multiple driver modules that simultaneously display functional similarity and mutation mutex in cancer. The reliability of MCLCluster is verified using GBM and BRCA cancer datasets and simulation samples. Taking GBM as an example, MCLCluster successfully identified 3 driver modules, which include some important and common driver genes, like CDKN2B, CDK4, RB1, ERBB2, TP53, EGFR etc., which provided important verification for this method. In the simulation dataset, the MCLCluster can maintain higher performance than Multi-Dendrix, FSME Cluster and RME in F1 scores. With the increase of noise, passenger rate and the module numbers in the simulation data, our method keeps a stable and sufficiently high F1 score, indicate that the MCLCluster can accurately identify modules in complex cases. BRCA and GBM are used as examples to prove the effectiveness of the method, and actually it is universal and can be applied to other type of interest cancer. In this paper, we use a general method to preprocess the real data set and construct the simulated data set, which is a feasible method verified by a lot of experiments. In addition, some parts of our method are general and can be used to solve other bioinformatics problems, such as the similarity measure method, which can be used to identify cancer-related microRNA modules based on microRNA-disease associations.

However, like previous researches of Multi-Dendrix, FSME Cluster and RME, MCLCluster is also designed for large sample sets to achieve statistical significance. Therefore, applying MCLCluster to a small number of samples may have some limitations. Some extensions can be used to further improve the MCLCluster method, for example, we can integrate the methylation and mRNA expression data, and use well-researched pathways reported in many literatures as a priori information. As the genome sequencing dataset in TCGA expands to more than 20 types of cancer, MCLCluster will be an important approach to identify new driver modules in different cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

AUTHOR CONTRIBUTIONS

LW conceived the study and supervised the study. WZ and YL developed the method. YL and YZ implemented the algorithms. WZ and YZ analyzed the data. WZ and LW wrote the manuscript. YC reviewed and improved the manuscript. All authors read and approved the final manuscript.

FUNDING

This project is partly sponsored by the National Natural Science Foundation of China (Nos. 61873221, 61672447), the Natural Science Foundation of Hunan Province (Nos. 2018JJ4058, 2019JJ70010), and the Hunan Province Key Laboratory of Industrial Internet Technology and Security (No. 2019TP1011).

REFERENCES

Ahmed, R., Baali, I., Erten, C., Hoxha, E., and., Kazan, H. (2019). MEXCOWalk:Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules. *Bioinformatics* 36, 872–879. doi: 10.1093/bioinformatics/btz655

Amgalan, B., and Lee, H. (2015). DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method. *Bioinformatics* 31, 52–60. doi: 10.1093/bioinformatics/btv175

Babaei, S., Hulsman, M., Reinders, M., and de Ridder, J. (2013). Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC Bioinform.* 14, 345–357. doi: 10.1186/1471-2105-14-29

Babur, O., Gonen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., et al. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* 16, 34–45. doi: 10.1186/s13059-015-0612-6

Brohee, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *Bmc Bioinformatics*. 7, 2944–2952. doi: 10.1186/1471-2105-7-488

Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature 07385

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Dao, P., Kim, Y. A., Wojtowicz, D., Madan, S., Sharan, R., and Przytycka, T. M. (2017). Bewith: a between-within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Comput. Biol.* 13:e1005695. doi: 10.1371/journal.pcbi.1005695

Dees, N. D., Zhang, Q. Y., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Deng, Y. L., Luo, S. Y., Deng, C. Y., Luo, T., Yin, W. K., Zhang, H. Y., et al. (2019). Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Brief Bioinform.* 20, 254–266. doi: 10.1093/bib/bbx109

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610

Hou, J. P., Emad, A., Puleo, G. J., Ma, J., and Milenkovic, O. (2016). A new correlation clustering method for cancer mutation analysis. *Bioinformatics* 32, 3717–3728. doi: 10.1093/bioinformatics/btw546

Hua, X., Xu, H. M., Yang, Y. N., Zhu, J., Liu, P. Y., and Lu, Y. (2013). DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Hum. Genet.* 93, 439–451. doi: 10.1016/j.ajhg.2013.07.003

Jia, P. L., Zhao, Z. M., and VarWalker. (2014). Personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput. Biol.* 10, 342–353. doi: 10.1371/journal.pcbi.1003460

Khurana, E., Fu, Y., Chen, J. M., and Gerstein, M. (2013). Interpretation of genomic variants using a unified biological network approach. *Plos Comput. Biol.* 9:e1002886. doi: 10.1371/journal.pcbi.1002886

Kim, Y. A., Cho, D. Y., Dao, P., and Przytycka, T. M. (2015). MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 31, 84–92. doi: 10.1093/bioinformatics/btv247

Kumar, R., Neilsen, P. M., Crawford, J., McKirdy, R., Lee, J., Powell, J. A., et al. (2005). FBXO31 is the chromosome 16q24.3 senescence gene, a candidate breast tumor suppressor, and a component of an SCF complex. *Cancer Res.* 65, 11304–11313. doi: 10.1158/0008-5472.CAN-05-0936

La Vecchia, S., and Sebastian, C. (2020). Metabolic pathways regulating colorectal cancer initiation and progression. *Semin. Cell Dev. Biol.* 98, 63–70. doi: 10.1016/j.semcdb.2019.05.018

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., and Raphael, B. R. (2014). Pan-cancer identification of mutated pathways and protein complexes. *Cancer Res.* 74, 112–123. doi: 10.1158/1538-7445.AM2014-5324

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9, 23–34. doi: 10.1371/journal.pcbi.1003054

Leiserson, M. D. M., Wu, H. T., Vandin, F., and Raphael, B. J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 16:160. doi: 10.1186/s13059-015-0700-7

Liu, X., Xi, J., Zhang, C., Feng, H., Li, A., and Wang, M. (2017). "Identification of driver network modules in protein-protein interaction network using patient mutation profiles," in *2017. 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (Shanghai: IEEE), 1–6. doi: 10.1109/CISP-BMEI.2017.8302274

Mandal, B. N., and Ma, J. (2016). l(1) regularized multiplicative iterative path algorithm for non-negative generalized linear models. *Comput. Stat. Data Anal.* 101, 289–299. doi: 10.1016/j.csda.2016.03.009

Nambara, S., Kurashige, J., Saito, T., Komatsu, H., Ueda, M., Sakimura, S., et al. (2015). Omics approach to identify driver genes for peritoneal dissemination of gastric cancer cells. *Cancer Res.* 75:5169. doi: 10.1158/1538-7445.AM2015-5169

Nangalia, J., Nice, F. L., Wedge, D. C., Godfrey, A. L., Grinfeld, J., Thakker, C., et al. (2015). DNMT3A mutations occur early or late in patients with myeloproliferative neoplasms and mutation order influences phenotype. *Haematologica* 100, E438–E442. doi: 10.3324/haematol.2015.129510

Network, C. G. A. R. (2012). Comprehensive genomic characterization of squamous cell lung cancers the cancer genome atlas research network. *Nature* 489, 519–525. doi: 10.1038/nature11666

Paull, E. O., Carlin, D. E., Niepel. M., Sorger, P. K, Haussler, D., and., Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). *Bioinformatics* 29, 2757–2564. doi: 10.1093/bioinformatics/btt471

Pelegrina, L. T., Sanhueza, M. D., Caceres, A. R. R., Cuello-Carrion, D., Rodriguez, C. E., and Laconi, M. R. (2020). Effect of progesterone and first evidence about allopregnanolone action on the progression of epithelial human ovarian cancer cell lines. *J. Steroid Biochem. Mol. Biol.* 196:105492. doi: 10.1016/j.jsbmb.2019.105492

Placke, T., Faber, K., Nonami, A., Putwain, S. L., Salih, H. R., Heidel, F. H., et al. (2014). Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. *Blood* 124, 13–23. doi: 10.1182/blood-2014-02-558114

Porta-Pardo, E., Hrabe, T., and Godzik, A. (2015). Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* 43, 968–973. doi: 10.1093/nar/gku1140

Reyna, M. A., Leiserson, M. D. M., and Raphael, B. J. (2018). Hierarchical hotnet: identifying hierarchies of altered subnetworks. *Bioinformatics* 34, 972–980. doi: 10.1093/bioinformatics/bty613

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., et al. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods.* 11, 396–398. doi: 10.1038/nmeth.2883

Rozenchan, P. B., Mundim, F. G., Roela, R. A., Katayama, M. L., Pasini, F. S., Brentani, H., et al. (2014). RHOA, RAC1 and PAK1 evaluation in paired stromal fibroblasts of breast cancer primary and of lymph node metastasis: importance of these biomarkers in lymph node invasion. *Cancer Res.* 74, 213–224. doi: 10.1158/1538-7445.AM2014-186

Salgia, R., Weaver, R. W., McCleod, M., Stille, J. R., Yan, S. B., Roberson, S., et al. (2017). Prognostic and predictive value of circulating tumor cells and CXCR4 expression as biomarkers for a CXCR4 peptide antagonist in combination with carboplatin-etoposide in small cell lung cancer: exploratory analysis of a phase II study. *Invest. New Drugs* 35, 334–344. doi: 10.1007/s10637-017-0446-z

Shih, Y. K., and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics* 28, I473–I479. doi: 10.1093/bioinformatics/bts370

Tang, C., Jiang, Y. S., Shao, W. W., Shi, W., Gao, X. S., Qin, W. Y., et al. (2016). Abnormal expression of FOSB correlates with tumor progression and poor survival in patients with gastric cancer. *Int. J. Oncol.* 49, 1489–1496. doi: 10.3892/ijo.2016.3661

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1007/978-3-642-12683-3_33

Vandin, F., Upfal, E., and Raphael, B. J. (2012). *De novo* discovery of mutated driver pathways in cancer. *Genome Res.* 22, 375–385. doi: 10.1101/gr.120477.111

Vlasblom, J., and Wodak, S. J. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10:99. doi: 10.1186/1471-2105-10-99

Wang, J., Zuo, Y., Man, Y. G., Avital, I., Stojadinovic, A., Liu, M., et al. (2015). Pathway and network approaches for identification of cancer signature markers from omics data. *J. Cancer* 6, 54–65. doi: 10.7150/jca. 10631

Wang, J. Z., Du, Z. D., Payattakool, R., Yu, P. S., and Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108–1113. doi: 10.1126/science.1145720

Wu, H., Gao, L., Li, F., Song, F., Yang, X. F., and Kasabov, N. (2015). Identifying overlapping mutated driver pathways by constructing gene networks in cancer. *BMC Bioinformatics.* 16, 334–345. doi: 10.1186/1471-2105-16-S5-S3

Wu, L. L., Wang, Y. Z., Liu, Y., Yu, S. Y., Xie, H., Shi, X. J., et al. (2014). A central role for TRPS1 in the control of cell cycle and cancer development. *Oncotarget* 5, 7677–7690. doi: 10.18632/oncotarget.2291

Xi, J. N., Wang, M. H., and Li, A. (2018). Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC Bioinform.* 19:214. doi: 10.1186/s12859-018-2218-y

Xiao, Q., Luo, J. W., Liang, C., Cai, J., and Ding, P. J. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Yang, H., Wei, Q., Zhong, X., Yang, H., and Li, B. (2017). Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework. *Bioinformatics* 33, 483–490. doi: 10.1093/bioinformatics/btw662

Zhang, J., and Zhang, S. (2016). The discovery of mutated driver pathways in cancer: models and algorithms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 988–998. doi: 10.1109/TCBB.2016.2640963

Zhang, W., and Wang, S. L. (2019a). An integrated framework for identifying mutated driver pathway and cancer progression. *Ieee/Acm Trans. Comput. Biol. Bioinform*. 16, 455–464. doi: 10.1109/T. C. B. B.2017.2788016

Zhang, W., and Wang, S. L. (2019b). A novel method for identifying the potential cancer driver genes based on molecular data integration. *Biochem. Genet.* doi: 10.1007/s10528-019-09924-2

Zhao, B. H., Zhao, Y. L., Zhang, X. X., Zhang, Z. H., Zhang, F., and Wang, L. (2019). An iteration method for identifying yeast essential proteins from heterogeneous network. *BMC Bioinformatics* 20:355. doi: 10.1186/s12859-019-2930-2

Zhao, J., Zhang, S., Wu, L. Y., and Zhang, X. S. (2012). Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* 28, 2940–2947. doi: 10.1093/bioinformatics/bts564

Zheng, C. H., Yang, W., Chong, Y. W., and Xia, J. F. (2016). Identification of mutated driver pathways in cancer using a multi-objective optimization model. *Comput. Biol. Med.* 72, 22–29. doi: 10.1016/j.compbiomed.2016.03.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 19 May 2020
doi: 10.3389/fbioe.2020.00394

TOOme: A Novel Computational Framework to Infer Cancer Tissue-of-Origin by Integrating Both Gene Mutation and Expression

Binsheng He[1]*†, Jidong Lang[2]†, Bo Wang[2], Xiaojun Liu[2], Qingqing Lu[2], Jianjun He[1], Wei Gao[3], Pingping Bing[1]*, Geng Tian[2]* and Jialiang Yang[2]*

[1] Academician Workstation, Changsha Medical University, Changsha, China, [2] Geneis Beijing Co., Ltd., Beijing, China, [3] Fujian Provincial Cancer Hospital, Fuzhou, China

Metastatic cancers require further diagnosis to determine their primary tumor sites. However, the tissue-of-origin for around 5% tumors could not be identified by routine medical diagnosis according to a statistics in the United States. With the development of machine learning techniques and the accumulation of big cancer data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), it is now feasible to predict cancer tissue-of-origin by computational tools. Metastatic tumor inherits characteristics from its tissue-of-origin, and both gene expression profile and somatic mutation have tissue specificity. Thus, we developed a computational framework to infer tumor tissue-of-origin by integrating both gene mutation and expression (TOOme). Specifically, we first perform feature selection on both gene expressions and mutations by a random forest method. The selected features are then used to build up a multi-label classification model to infer cancer tissue-of-origin. We adopt a few popular multiple-label classification methods, which are compared by the 10-fold cross validation process. We applied TOOme to the TCGA data containing 7,008 non-metastatic samples across 20 solid tumors. Seventy four genes by gene expression profile and six genes by gene mutation are selected by the random forest process, which can be divided into two categories: (1) cancer type specific genes and (2) those expressed or mutated in several cancers with different levels of expression or mutation rates. Function analysis indicates that the selected genes are significantly enriched in gland development, urogenital system development, hormone metabolic process, thyroid hormone generation prostate hormone generation and so on. According to the multiple-label classification method, random forest performs the best with a 10-fold cross-validation prediction accuracy of 96%. We also use the 19 metastatic samples from TCGA and 256 cancer samples downloaded from GEO as independent testing data, for which TOOme achieves a prediction accuracy of 89%. The cross-validation validation accuracy is better than those using gene expression (i.e., 95%) and gene

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Taigang Liu,
Shanghai Ocean University, China
Wenhui Wang,
Icahn School of Medicine at Mount Sinai, United States
Min Chen,
Hunan Institute of Technology, China

****Correspondence:***
Binsheng He
hbscsmu@163.com
Pingping Bing
bpping@163.com
Geng Tian
tiang@geneis.cn
Jialiang Yang
yangjl@geneis.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *19 February 2020*
Accepted: *08 April 2020*
Published: *19 May 2020*

Citation:
He B, Lang J, Wang B, Liu X, Lu Q, He J, Gao W, Bing P, Tian G and Yang J (2020) TOOme: A Novel Computational Framework to Infer Cancer Tissue-of-Origin by Integrating Both Gene Mutation and Expression. Front. Bioeng. Biotechnol. 8:394. doi: 10.3389/fbioe.2020.00394

mutation (53%) alone. In conclusion, TOOme provides a quick yet accurate alternative to traditional medical methods in inferring cancer tissue-of-origin. In addition, the methods combining somatic mutation and gene expressions outperform those using gene expression or mutation alone.

Keywords: tissue-of-origin, somatic mutation, gene expression, random forest, cross-validation

INTRODUCTION

Metastatic cancer is a common clinical challenge for limited evidence to determine its primary origin. Patients with carcinoma of unknown primary (CUP) account for about 5% of total cancer patients (Shaw et al., 2007). CUP are usually heterogeneous, and can lead to dilemmas in diagnosing and treatment since the original tumor site is unknown (Rizwan and Zulfiqar, 2010). Clinically, CUP patients are generally treated with non-selective empirical chemotherapy, which usually leads to low survival rates (Kurahashi et al., 2013). Thus, identifying cancer tissue-of-origin (TOO) is critical in improving the treatment of cancer patients and extending their surviving time (Hudis, 2007; Varadhachary et al., 2008; Hyphantis et al., 2013).

There are several ancillary examinations in CUP identification, among which immunohistochemistry (IHC) is an important one. However, this method relies on the experiences of pathologists and is labor-intensive. As a result, it is inaccurate in most of the times (Huebner et al., 2007; Voigt, 2008; Centeno et al., 2010; Kandalaft and Gown, 2015; Janick et al., 2018). Positron emission tomography (PET) and computed tomography (CT) are also commonly used in the identification of CUP (Fencl et al., 2007; Kwee et al., 2010; Fu et al., 2019). The detection rate of conventional radiological imaging on primary carcinoma reach 20–27%, and that of PET reach 24–40% (Ambrosini et al., 2006). The detection accuracy of PET/CT is awfully low that it rarely brings help to identify the primary origin. Obstacles in image technology cause much difficulty of effective use of relative Carcinoma image to help tracing cancer tissue origin.

Molecular profiling of tissue-specific genes is also being used in CUP work-up. Quantities of large-scale profiles of different tumors have been used for diagnose. Molecular profiling is as well as or better than IHC, in terms of poorly differentiated or undifferentiated tumors (Oien and Dennis, 2012). Therefore, making use of molecular profiling has become a popular way for diagnosis of unknown origin. Comprehensive molecular profiles displayed in The Cancer Genome Atlas (TCGA) including copy number variation, somatic mutation, gene expression, microRNA expression, DNA methylation, and protein expression, are used to identifying human tumor types (Li et al., 2017). By analysis of tumor types from data of methylation and copy number variation, tissue of origin and molecular classification can be revealed (Hoadley et al., 2014). The methylation profile of metastasis in a meningeal melanocytic tumor is similar to that of primary tumor, and it is suggest that particular copy number variations may be associated with metastatic behavior (Küsters-Vandevelde et al., 2017). Methylation and copy number variation are DNA-level molecular profiling, which brought great help to identify tumor origins.

The copy number profile and gain or loss in specific chromosome regions have been researched by hybridization and cytogenetic-based methods (Baudis, 2007; Beroukhim et al., 2007). An *IDH1* somatic mutation in genomic profiling was revealed to bring great benefit to the diagnosis of cholangiocarcinoma and trace the primary origin in a malignancy (Sheffield et al., 2016). Marquard et al. (2016) obtained classification accuracy of 69% and 85% on 6 and 10 primary sited with somatic mutation, respectively, based on PM and CN classifier (classifiers with both point mutations and copy number aberrations) with cross-validation. Mutation of tumor-specific enrichment in certain genes, has been utilized to infer tumor localization, and Dietlein and Eschner (2014) developed a tool with mutation spectra to infer cancer origins with a prediction specificity of 79% (Lawrence et al., 2014). As a DNA-level molecular profiling, SNP, that is somatic mutation, can be used as a very useful tool to infer the tissue of origins.

A lot of RNA-level gene expression profile have been explored to identify the cancer tissue of origin (Erlander et al., 2004; Qu et al., 2007; Gross-Goupil et al., 2012; Greco, 2013; Hainsworth et al., 2013). Erlander et al. (2011) have demonstrated that the gene expression value of samples detected in metastatic tumor is similar to that in the original tumor under condition of CUP. Centeno et al. (2010) developed a hybrid model by integrating expression profiling and IHC for microRNA-based qRT-PCR test on identification of cancer tissue origin, with 85% of the cases correctly identified (Rosenwald et al., 2010). Bloom et al. (2004) utilized artificial neural networks (ANNs) to predict the unknown cancer tissue origin with mean accuracy of 83–88% in different platforms.

Numerous researches have utilized molecular profiles, such as copy number variation, somatic mutation, gene expression, and so on for predicting cancer tissue origin. However, the accuracy of prediction was not satisfying. Identifying cancer tissue origin by combining somatic mutation and gene expression profiling on DNA level and RNA level, respectively, is first proposed in this study. Firstly, we obtained the data of somatic mutation and gene expression profiling from International Cancer Genome Consortium (ICGC) Database. Machine learning methods can help to improve the performance on prediction of cancer tissue origin. We aim to obtain better performance in predicting cancer tissue origin, by the combination of somatic mutation and gene expression profiling, based on random forest. Machine learning algorithm, such as logistic regression can be used to select gene (Kao et al., 2006). However, random forest algorithm (Sandri and Zuccolotto, 2006) was chosen as the gene selection algorithm in this study due to its advantage, good robustness and easy to use. Finally, we used random forest algorithm for classification of cancers. Experiment results showed that higher accuracy can be obtained by using the method proposed in this study.

MATERIALS AND METHODS

Gene Expression Data

Gene expression profile was downloaded from ICGC Database version release-26[1]. Each gene is named by Gene Symbol ID. The value of gene expression in each labeled sample is normalized by TPM. After deduplication, samples were extracted for combination with SNP samples.

Somatic Mutation Data

The somatic mutation data was downloaded from ICGC Database version release-28[2]. Each gene is named by Ensembl Gene ID. For Gene Symbol ID is most widely used in paper, the Ensembl Gene ID of gene name in somatic mutation data was converted to Gene Symbol ID. The samples are deduplicated according to information of ICGC-donor-ID, chromosome, and

[1] https://dcc.icgc.org/releases/release_26/

[2] https://dcc.icgc.org/releases/release_28/

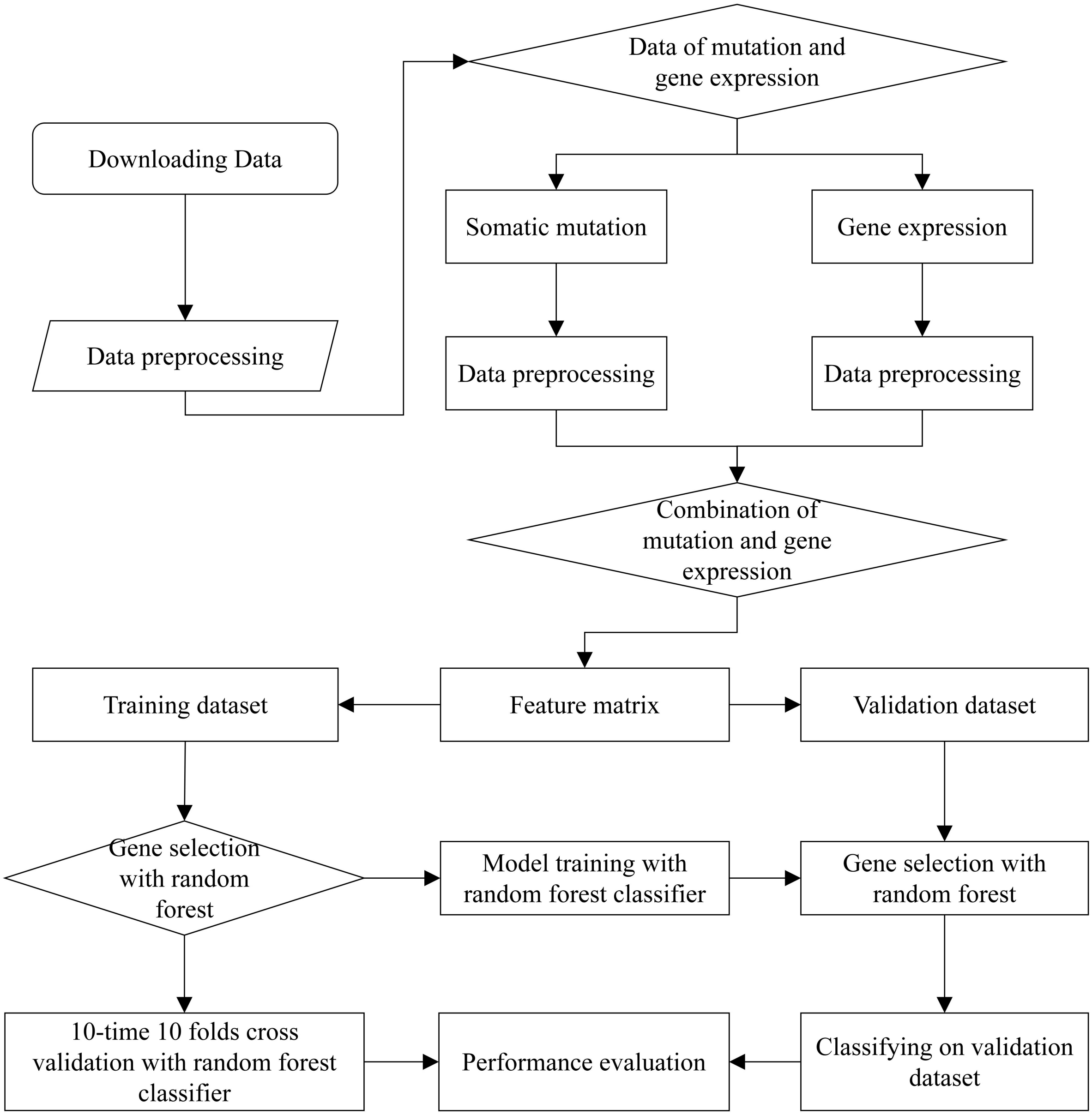

Downloading Data
Data preprocessing
Data of mutation and gene expression
Somatic mutation
Gene expression
Data preprocessing
Data preprocessing
Combination of mutation and gene expression
Training dataset
Feature matrix
Validation dataset
Gene selection with random forest
Model training with random forest classifier
Gene selection with random forest
10-time 10 folds cross validation with random forest classifier
Performance evaluation
Classifying on validation dataset

FIGURE 1 | The complete workflow of prediction on cancer tissue origin.

locus in chromosome and gene-affected. Each sample was labeled by its type of cancer.

Data Combination

The gene expression and somatic mutation data were merged into one feature matrix. For labeled samples with gene expression array data only involves in 21 cancer types, and samples with Skin Cutaneous Melanoma (SKCM) were removed for it contributes to the major metastasis cancers. The sample with somatic mutation data whose label was not included in these 20 cancer types was removed. Then, the shared sample data was chosen, therefore the samples data after filtering is obtained from 20 different cancer types. An M*N matrix was generated, where M and N represents the number of sample and gene, respectively.

Gene Selection

Because gene sequencing and mutation detection are costly and time consuming, a scale reduction of gene number is necessary. There are many feature selection algorithms, like Lasso, PCA (Malhi and Gao, 2005; Muthukrishnan and Rohini, 2016) and etc. The Random forest (Breiman, 2001; Sandri and Zuccolotto, 2006) was a supervised learning algorithm, which is an ensemble learning algorithm based on decision tree and was used to select genes. Best performance was obtained by using 80 selected genes. $\sqrt{n}$ genes were used in a tree, where n represents the number of genes. At the process of splitting node, Gini index was used, which is calculated by formula:

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} {p_k}^2 \tag{1}$$

Where p represents the weight referring to frequencies of cancers in a node, k represents the number of cancers and p_k represents the weight of the kth cancer. The variable importance measures of ith gene in node m, that is the Gini index variation after splitting of node m, is calculated by formula:

$$VIM_{im}^{(Gini)} = GI_m - GI_l - GI_r \tag{2}$$

Where m is a node in M, which is a set of nodes, $VIM_{im}^{(Gini)}$ represents variable importance measures of ith gene in node m, the GI_m represents the Gini index before splitting, GI_l and GI_r represents the Gini index of two new node after splitting, respectively. The importance of the ith gene, in the tth tree is calculated by formula:

$$VIM_{ti}^{(Gini)} = \sum_{m \in M} VIM_{im}^{(Gini)} \tag{3}$$

Where $VIM_{ti}^{(Gini)}$ represents the importance of the ith gene in the tth tree. If the set of trees is T, the importance of the ith gene in all the tree is calculated by formula:

$$VIM_{i}^{(Gini)} = \sum_{t=1}^{T} VIM_{ti}^{(Gini)} \tag{4}$$

TABLE 1 | Sample distribution of each cancer from ICGC database.

Available cancer types	Abbreviation	Samples	
		Amount	**Percentage**
Bladder urothelial carcinoma	BLCA	294	4.20%
Breast invasive carcinoma	BRCA	970	13.84%
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	241	3.44%
Colon adenocarcinoma	COAD	390	5.57%
Glioblastoma multiforme	GBM	148	2.11%
Head and neck squamous cell carcinoma	HNSC	460	6.56%
Kidney renal clear cell carcinoma	KIRC	345	4.92%
Kidney renal papillary cell carcinoma	KIRP	216	3.08%
Acute myeloid leukemia	LAML	121	1.73%
Brain lower grade glioma	LGG	433	6.18%
Liver hepatocellular carcinoma	LIHC	282	4.02%
Lung adenocarcinoma	LUAD	475	6.78%
Lung squamous cell carcinoma	LUSC	411	5.87%
Ovarian serous cystadenocarcinoma	OV	185	2.64%
Pancreatic adenocarcinoma	PAAD	134	1.91%
Prostate adenocarcinoma	PRAD	374	5.34%
Rectum adenocarcinoma	READ	137	1.95%
Stomach adenocarcinoma	STAD	412	5.88%
Thyroid carcinoma	THCA	486	6.93%
Uterine corpus endometrial carcinoma	UCEC	494	7.05%
Total		7008	100%

TABLE 2 | Performance of classification of combination of somatic mutation and gene expression by using 80 genes.

Cancer type	Precision	Recall	F1-score	Support	Specificity
BLCA	0.8906	0.9354	0.9124	294.0000	0.9950
BRCA	0.9987	0.9947	0.9967	970.0000	0.9998
CESC	0.9148	0.8859	0.9001	241.0000	0.9971
COAD	0.7548	0.9644	0.8468	390.0000	0.9815
GBM	0.9940	1.0000	0.9970	148.0000	0.9999
HNSC	0.9916	1.0000	0.9958	460.0000	0.9994
KIRC	0.9850	0.9516	0.9680	345.0000	0.9992
KIRP	0.9344	0.9630	0.9485	216.0000	0.9979
LAML	1.0000	1.0000	1.0000	121.0000	1.0000
LGG	0.9926	0.9977	0.9952	433.0000	0.9995
LIHC	0.9925	0.9844	0.9884	282.0000	0.9997
LUAD	0.9358	0.9448	0.9403	475.0000	0.9953
LUSC	0.9408	0.9000	0.9199	411.0000	0.9965
OV	1.0000	0.9946	0.9973	185.0000	1.0000
PAAD	0.9378	0.9552	0.9464	134.0000	0.9988
PRAD	0.9973	1.0000	0.9987	374.0000	0.9998
READ	0.7569	0.1591	0.2627	137.0000	0.9990
STAD	0.9947	0.9976	0.9961	412.0000	0.9997
THCA	1.0000	0.9979	0.9990	486.0000	1.0000
UCEC	0.9673	0.9816	0.9744	494.0000	0.9975
Accuracy	0.9577	0.9577	0.9577	0.0000	

Where $VIM_i^{(Gini)}$ is the importance of the ith gene in all trees. We sorted the importance scores of all genes, then the top H genes were selected, where H is the variable number of genes that can be set to find the best result.

Multi-Classifier Random Forest

The random forest is actually a special method of bagging that using the decision tree as a model in bagging (Breiman, 2001; Meyer et al., 2019). First, the bootstrap method is used to generate m training sets, which is a set of samples. Then, each training set is used to construct a tree. $\sqrt{n}$ genes are used in a tree, where n represents the number of selected genes. When splitting a node, not all the genes are used to optimize the metric Gini index used in this study, a part of genes is randomly extracted instead. An optimal solution can be found among the extracted genes, and applied to node splitting. Leaf node in the tree records which gene is used to determine the cancer type, and each leaf node represents the last judged cancer type. The predicted cancer type is given by maximum votes from decision tree.

Statistical Analysis

The metric of precision, recall and F1 score were used to evaluate the performance of the model. True-positive, false-positive, true-negative and false-negative are abbreviated as TP, FP, TN, and FN, respectively. Precision is calculated by $(TP)/(TP + FP)$, which indicates the ability of classifier to differentiate positive from negative cases. Recall is calculated by $(TP)/(TP + FN)$, which indicates the ability of classifier to recognize all positive cases. The $F1$ score is calculated by $(2 * recall * precision)/(recall + precision)$. Each individual cancer type is calculated by these metrics, and the cohort metric adopt the mean report. The entire cohort is calculated by accuracy, reported as $(TP + TN)/(total\,cases)$. Ten times 10-fold cross validation is used to obtain the metric report, whose average is treated as the result metric.

Gene Annotation

The functions annotation of specific gene set was given. Geno ontology (Ye et al., 2006; Waardenberg et al., 2016) was used as enrichment analysis database. Gene clustering and visualization was realized by R package cluaterProfiler and gogadget (Yu et al., 2012; Nota, 2016).

RESULTS

The Workflow of TOOme

The complete workflow of prediction on cancer tissue origin is shown in **Figure 1**. The process can be split into three steps. At

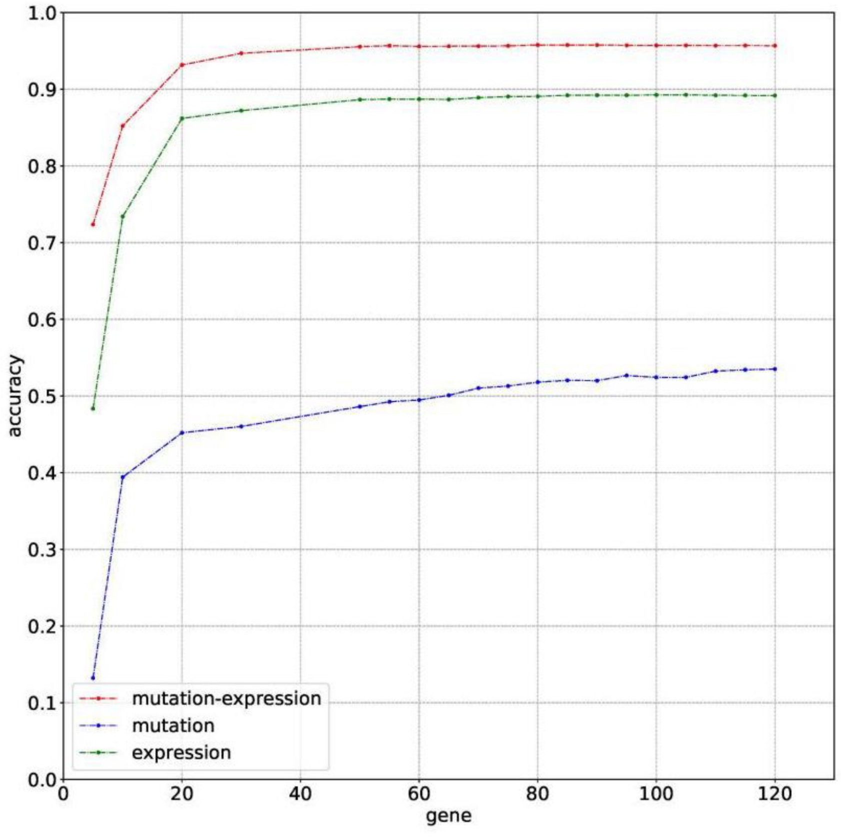


FIGURE 2 | The classification accuracy of using somatic mutation, gene expression and combination of somatic mutation and gene expression, respectively.

TABLE 3 | Prediction probabilities of each samples on each cancer.

Cancer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
BLCA	0.0005	0.0015	0.0005	0	0.1825	0.162	0.0665	0.0155	0.002	0.001	0.034	0	0	0	0	0.0015	0.0005	0	0
BRCA	0.993	0.9675	0.9995	0.999	0.6375	0.1195	0.045	0.066	0.0015	0.0005	0.0085	0.001	0.0005	0	0	0	0	0	0
CESC	0.0005	0.004	0	0	0.047	0.101	0.8	0.086	0.0275	0.002	0.1115	0	0	0	0	0.0015	0	0	0.001
COAD	0	0.001	0	0.0005	0.005	0.01	0.008	0.002	0.7015	0.001	0.009	0	0	0	0	0.001	0	0	0
GBM	0	0	0	0	0.001	0.0035	0	0	0	0	0.001	0	0	0	0	0	0.0005	0	0
HNSC	0.0005	0	0	0	0.0065	0.011	0.0055	0.0015	0	0.993	0.754	0	0	0	0	0	0	0	0.001
KIRC	0	0	0	0	0.0015	0.0535	0.001	0.003	0.0005	0	0.001	0	0.0005	0	0	0.0015	0.001	0	0
KIRP	0	0	0	0	0.004	0.038	0.001	0.0045	0.0005	0	0	0	0	0	0	0.0005	0.0015	0	0
LAML	0	0.006	0	0	0.0155	0.0055	0	0.005	0.001	0	0.0005	0	0	0	0	0	0	0	0
LGG	0	0	0	0	0.0125	0.165	0.0055	0.01	0.0005	0.0005	0.0035	0	0	0	0	0.001	0	0	0.0005
LIHC	0	0.0005	0	0	0.003	0.0365	0.0045	0.0045	0.0095	0	0.001	0	0	0	0	0	0	0	0
LUAD	0.0025	0.006	0	0	0.011	0.0225	0.009	0.012	0.001	0	0.0055	0.0065	0	0	0	0.0025	0.001	0.001	0.001
LUSC	0.001	0.008	0	0.0005	0.017	0.0735	0.0375	0.008	0	0	0.024	0.001	0.0005	0	0	0.0015	0.0005	0.0005	0.002
OV	0	0	0	0	0.002	0.0005	0	0.001	0	0	0	0	0	0	0	0	0.002	0	0
PAAD	0	0.0005	0	0	0.0095	0.0775	0.004	0.0045	0.0075	0	0.001	0	0	0	0	0.0005	0	0	0
PRAD	0	0.0005	0	0	0.003	0.004	0.002	0.001	0	0	0.0005	0	0	0	0	0	0	0	0.001
READ	0	0.002	0	0	0.0005	0.001	0.003	0.0005	0.242	0.0005	0.0065	0	0	0	0	0	0	0	0
STAD	0	0	0	0	0.0055	0.0025	0.0005	0.0005	0.0045	0	0.004	0	0	0	0	0	0	0	0
THCA	0	0	0	0	0.0015	0.0035	0	0.0065	0	0	0.0005	0.991	0.9985	1	1	0.9875	0.9925	0.9985	0.992
UCEC	0.002	0.0025	0	0	0.034	0.1095	0.007	0.768	0.0005	0.0015	0.034	0.0005	0	0	0	0.001	0.0005	0	0.0015
LOW_CONFIDENCE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Predicted_label	BRCA	BRCA	BRCA	BRCA	BRCA	LGG	CESC	UCEC	COAD	HNSC	HNSC	THCA	THCA	THCA	THCA	THCA	THCA	THCA	THCA
True_label	BRCA	BRCA	BRCA	BRCA	BRCA	BRCA	CESC	CESC	COAD	HNSC	HNSC	THCA	THCA	THCA	THCA	THCA	THCA	THCA	THCA
Correct	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1

the first step, we download the raw data from ICGC Database, and extracted the effective information to obtain preliminary data of somatic mutation and gene expression profiling. At the second step, we filtered the data of somatic mutation and gene expression profiling, respectively. Then, samples with both somatic mutation data and gene expression proofing were used to form feature matrix. As a result, the generated feature matrix was used for gene selection. At the third step, most of the samples were utilized to train the model with 10-time 10 folds cross validation by using random forest classification algorithm. We carried out numerous experiments to evaluate the performance of the proposed method.

Data Used in This Study

We used ICGC version 26 and 28 databases, with Gene expression profile and somatic mutation information to classify tumor samples. The allele mutation in somatic mutation data can be A/G, C/T, C/A, and etc. For it is hard to distinguish mutation types with limited relative information and tools, we consider all kinds of allele mutation as gene mutation and count the number of gene mutation of each sample. Different from somatic mutation data, gene expression profile array data is directly used. The sample distribution of each cancer is showed in **Table 1**, where samples suffer from BRCA are much more than from other cancers. Considerable prediction results can be obtained by our model. The precision, recall and F_1 score, showed in **Table 2**, reach 99.86%, 99.47% and 99.67%, respectively.

In this study, there are 371 samples with metastasis, where 352 samples are SKCM. To avoid unbalanced distribution of samples, we removed all the SKCM samples with metastasis. Only 19 samples with metastasis were used as test dataset.

Performance Evaluation

The classification accuracies obtained by using data of somatic mutation, gene expression profiling and both of them, under condition of using different number of genes, have been compared in **Figure 2**. Motivated by Ma et al. (2006) that five genes can be used to solve a 32-type classification problem, five was chosen as the minimum number of genes. For gene sequencing and mutation detection are costly and time consuming, 120 was chosen as the maximum number of genes. A lot of experiments have been done using the prepared data between the interval from 5 to 120. For using small number of genes did not obtain satisfying classification performance, the interval between number of genes was set to 10 or even larger until the number of genes equals to 50. Then the interval was set to 5 for fine tuning, based on small fluctuation by changed number of genes.

Results with 10-time 10 folds cross validation on training dataset are shown in **Figure 2** that accuracy of using data of both somatic mutation and gene expression profiling is always higher than that of only using one of it. The best result of them are 95.77%, 53.51%, and 89.28%, obtained by using 80, 120, and 105 genes, respectively. Results shows that gene expression can make much contribution to obtain higher accuracy than data of somatic mutation. However, a combination of them achieved best classification performance.

As for the test dataset, we conducted experiments by using the chosen 80 genes in training model. The overall classification accuracy is 89.47%. **Table 3** shows the prediction probabilities of each sample on each cancer. The value on the table highlighted by color of green, yellow, and pink presents high, middle, and low probabilities, respectively, of predicting a sample to

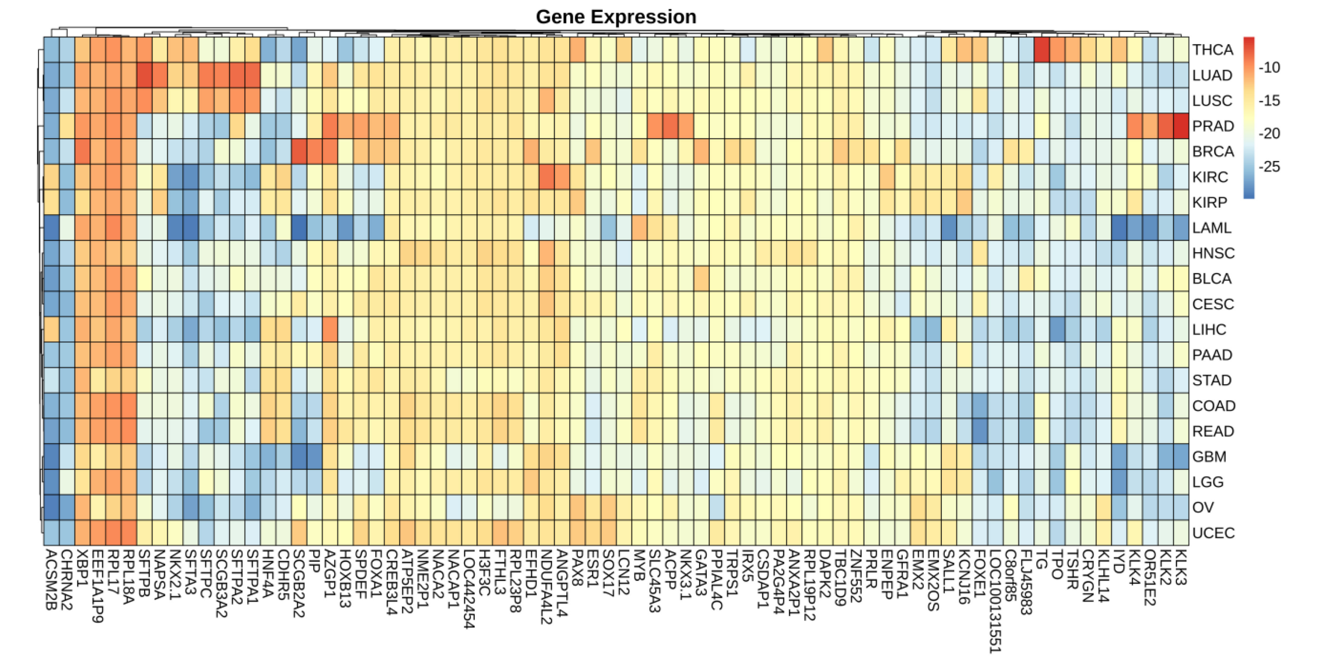


FIGURE 3 | Heatmap of mean value of gene expression on each cancer.

a cancer type. We obtained considerable prediction accuracy on sample with BRCA and THCA. Each sample was correctly predicted to the same as the true label. A sample whose true label is CESC was predicted to UCEC. A sample whose true label is BRCA was predicted to LGG with a terrible probability 1.65%. In this condition, we considered that little error on classification is tolerable.

Mean Value of Gene Expression and Somatic Mutations on Each Cancer

We plotted the heatmap of mean value of gene expression and somatic mutations on each cancer. In **Figure 3**, the rows represent 74 genes of gene expression and columns denote the cancers. In **Figure 4**, the rows represent six genes of somatic mutation and columns represent the cancers. The mean value of gene expression and somatic mutation on a logarithmic scale was plotted with relative color. A color bar was used to display the value difference. Cancers that fell into cluster at horizontal axis had a similar value between gene expression or mutation number. The genes were also clustered at vertical axis based on the similarity between cancers.

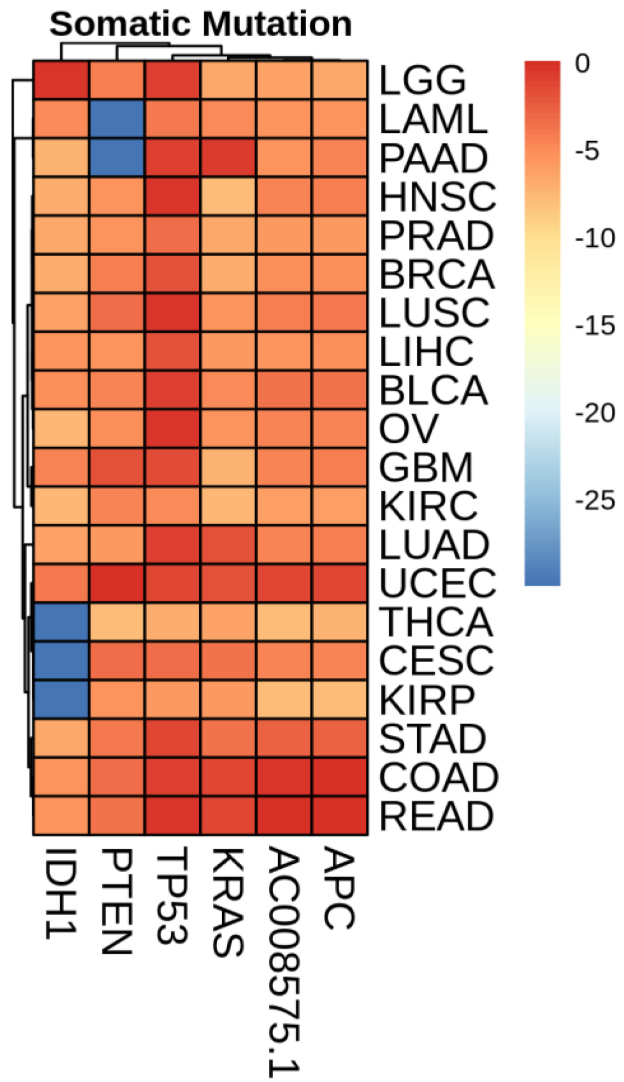


FIGURE 4 | Heatmap of mean value of somatic mutations on each cancer.

DISCUSSION

Data of somatic mutation and gene expression profiling can be used to identify the primary site of tumors. However, it was the first time to identify the cancer tissue origin by using both data of somatic mutation and gene expression profiling. We carried out experiments by using 7008 samples with combination of data of somatic and gene expression profiling among 20 cancers. By comparing the performance of them, we obtained highest accuracy by leveraging both of the data of somatic mutation and gene expression profiling.

The primary analysis tool we used was random forest (Breiman, 2001; Sandri and Zuccolotto, 2006), a machine learning algorithm that can be used for gene selection and tumor classification. We chose top-rank 80 genes, where 6 genes and 74 genes are corresponding to mutation and expression, respectively, for classification. Therefore, it showed that data of somatic mutation performs worse than gene expression profiling on prediction of cancer tissue origin. Our method obtained 96% overall accuracy on the training dataset. The performance is maintained considerably on the external cohorts, and the overall accuracy on sample with metastatic disease is 89%. Our model cannot provide good performance on physiologically proximal cancers, such as uterine corpus endometrial carcinoma and cervical squamous cell carcinoma and endocervical adenocarcinoma. The endometrial and ovarian endometrioid carcinomas evolve from similar precursor endometrial epithelial cells; many researches are involved in the molecular pathogenesis of the endometrial and ovarian endometrioid carcinomas (McConechy et al., 2014).

We studied the role that gene plays in cellular component, biological process and molecular function. **Figure 5** shows the top-rank 80 genes selected by random forest algorithm. The selected genes were enriched in hormone metabolic process, tissue and organ development and hormone-mediated signaling pathway, specifically in gland development, urogenital system development, hormone metabolic process, morphogenesis of a branching epithelium, morphogenesis of a branching structure, endocrine system development, branching morphogenesis of an epithelial tube, thyroid hormone metabolic process, thyroid hormone generation and prostate gland development. For example, *APC* plays a significant role in discovering pathogenesis of soft tissue tumors (Kuhnen et al., 2000). Birnbaum et al. (2012) investigated what role the *APC* gene play in colorectal cancer, at the investigation of 183 colon adenocarcinomas, point mutations were found in 73% of cases. We obtained the similar conclusion that mutation of *APC* gene may be the important impact of colorectal cancer, as heatmap shown in **Figure 4** that the mean number of *APC* gene mutation in colorectal cancer is more than that in other cancers except rectum adenocarcinoma. It can be

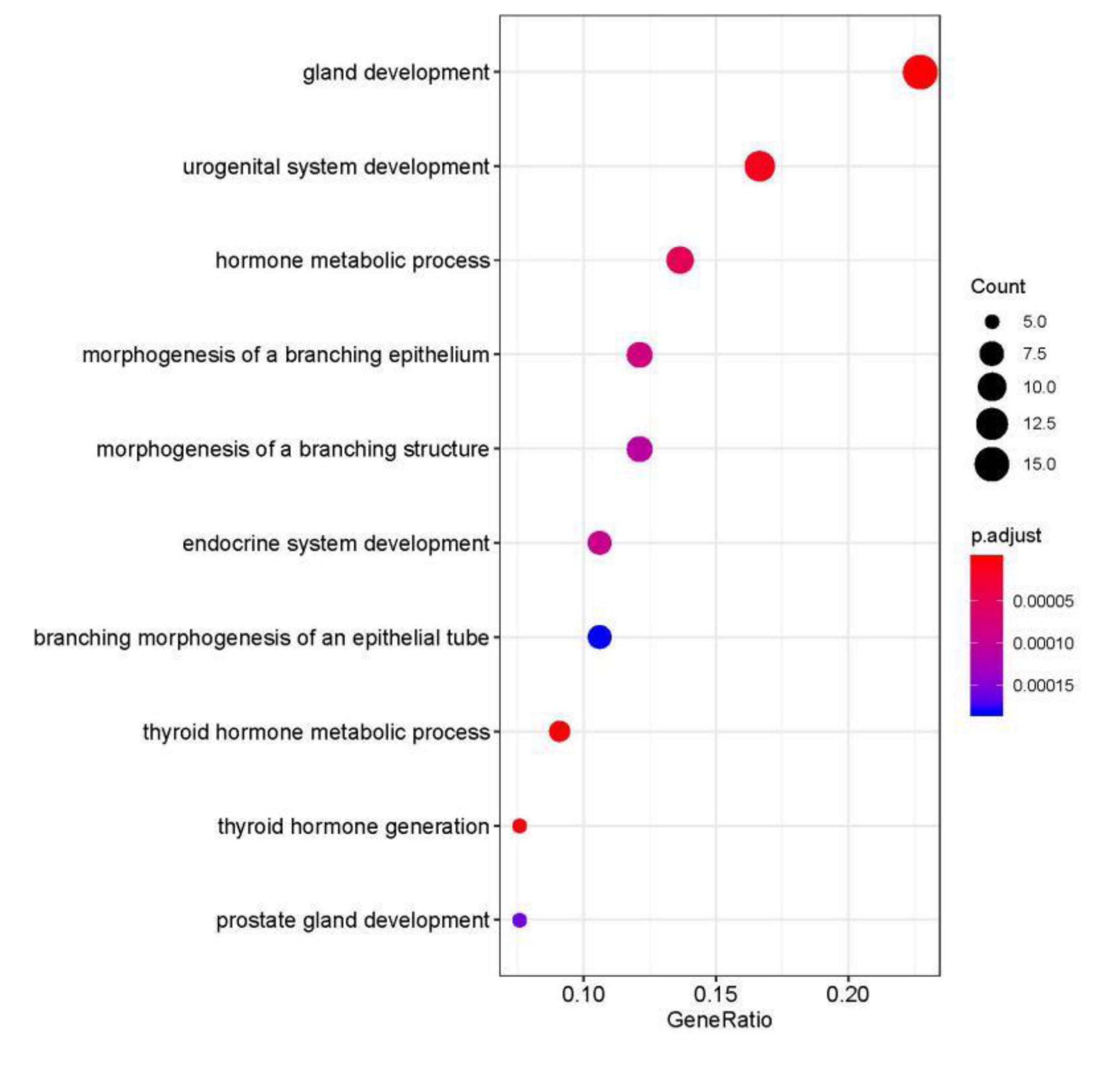


FIGURE 5 | Selected top-rank 80 genes enriched in cellular component, biological process and molecular function.

explained that they are two physiologically proximal cancers. Mutation in *IDH1* gene can reduce cell survival, proliferation and invasion of human glioma (Cui et al., 2016). Mutation in *IDH1* gene is an oncogenic driver in a majority of lower-grade gliomas and have an impact on brain lower grade glioma with different genetic pathway (Ohno et al., 2013; Pieper et al., 2014; Ohka et al., 2017). The same conclusion was acquired in **Figure 4** that the mean number of *IDH1* gene mutation in Brain lower grade glioma is more than that in other cancers.

ACPP gene plays a vital key in prostate adenocarcinoma (Maatman et al., 1984; Drago et al., 1989; Vihko et al., 2005). From the heatmap, it is clear that the level of *ACPP* gene expression in prostate adenocarcinoma is higher than that in other cancers. The expression levels of TG were found to be altered in all kinds of thyroid carcinomas (Makhlouf et al., 2016). From **Figure 3**, we obtained similar results that the level of *TG* gene expression in thyroid carcinomas is higher than that in other cancers.

Molecular profiling of tissue-specific genes can be utilized to identify the primary site of tumor. Combination of data of somatic mutation and gene expression profiling were first proposed in this study to predict the primary origin. We obtained considerable prediction performance, and therefore this research can bring great help to the identification of cancer tissue origin. However, we did not carry out research to discover the relationship between data of gene expression and somatic mutation. Our method cannot classify physiologically proximal cancers yet. And it is also a future work to employing other machine learning algorithms that can improve the classification performance.

CONCLUSION

Identification of cancer tissue origin is a challenging work recently and in the future. With a lot of molecular profiling available, we can make use of them alone and combine some of them to improve performance of identification primary site of tumor. Machine learning algorithm is also an effective tool to help classifying the cancers. The prediction performance can be tremendously affected by the number of features used.

In this study, we used both molecular data of somatic mutation and gene expression profiling to generate a feature matrix. Then the optimal number of genes was obtained and the data was trained, based on random forest algorithm. The performance of using our method was also compared to only by using data of somatic mutation or gene expression profiling. Our method achieved highest accuracy. Experiment results shows that our method can be an effective tool for primary origin tracing.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/releases/release_26/, https://dcc.icgc.org/releases/release_28/.

AUTHOR CONTRIBUTIONS

JY, GT, and PB conceived the concept of the work. BH, XL, BW, and JL performed the experiments. BH and XL wrote the manuscript. QL, WG, and JH reviewed the manuscript. All authors approved the final version of this manuscript.

FUNDING

This research was funded by Hunan Provincial Innovation Platform and Talents Program (Grant No. 2018RS3105), the Natural Science Foundation of China (Grant No. 61803151), the Natural Science Foundation of Hunan Province (Grant No. 2018JJ3570), and the Project of Scientific Research Fund of Hunan Provincial Education Department (Grant Nos. 19A060 and 19C0185).

REFERENCES

Ambrosini, V., Nanni, C., Rubello, D., Moretti, A., Battista, G., Castellucci, P., et al. (2006). 18F-FDG PET/CT in the assessment of carcinoma of unknown primary origin. *Radiol. Med.* 111, 1146–1155. doi: 10.1007/s11547-006-0112-6

Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* 7:226. doi: 10.1186/1471-2407-7-226

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl.Acad. Sci. U.S.A.* 104, 20007–20012. doi: 10.1073/pnas.0710052104

Birnbaum, D. J., Laibe, S., Ferrari, A., Lagarde, A., Fabre, A. J., Monges, G., et al. (2012). Expression profiles in stage II colon cancer according to APC gene status. *Transl. Oncol.* 5, 72–76. doi: 10.1593/tlo.11325

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., et al. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am. J. Pathol.* 164, 9–16. doi: 10.1016/S0002-9440(10)63090-8

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Centeno, B. A., Bloom, G., Chen, D.-T., Chen, Z., Gruidl, M., Nasir, A., et al. (2010). Hybrid model integrating immunohistochemistry and expression profiling for the classification of carcinomas of unknown primary site. *J. Mol. Diagn.* 12, 476–486. doi: 10.2353/jmoldx.2010.090197

Cui, D., Ren, J., Shi, J., Feng, L., Wang, K., Zeng, T., et al. (2016). R132H mutation in IDH1 gene reduces proliferation, cell survival and invasion of human glioma by downregulating Wnt/β-catenin signaling. *Int. J. Biochem.Cell Biol.* 73, 72–81. doi: 10.1016/j.biocel.2016.02.007

Dietlein, F., and Eschner, W. (2014). Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. *Hum. Mol. Genet.* 23, 1527–1537. doi: 10.1093/hmg/ddt539

Drago, J. R., Badalament, R. A., Wientjes, M. G., Smith, J. J., Nesbitt, J. A., York, J. P., et al. (1989). Relative value of prostate-specific antigen and prosttic acid phosphatase in diagnosis and management of adenocarcinoma of prostate ohio state university experience. *Urology* 34, 187–192. doi: 10.1016/0090-4295(89)90369-5

Erlander, M. G., Ma, X.-J., Kesty, N. C., Bao, L., Salunga, R., and Schnabel, C. A. (2011). Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification. *J. Mol. Diagnost.* 13, 493–503. doi: 10.1016/j.jmoldx.2011.04.004

Erlander, M. G., Moore, M. W., Cotter, P., Reyes, M., Stahl, R., Hamati, H., et al. (2004). Molecular classification of carcinoma of unknown primary by gene expression profiling from formalin-fixed paraffin-embedded tissues. *J. Clin. Oncol.* 22:9545. doi: 10.1200/JCO.2007.14.4378

Fencl, P., Belohlavek, O., Skopalova, M., Jaruskova, M., Kantorova, I., and Simonova, K. (2007). Prognostic and diagnostic accuracy of [18F]FDG-PET/CT in 190 patients with carcinoma of unknown primary. *Eur. J. Nucl. Med. Mol. Imaging* 34, 1783–1792. doi: 10.1007/s00259-007-0456-8

Fu, Z., Chen, X., Yang, X., and Li, Q. (2019). Diagnosis of primary clear cell carcinoma of the vagina by 18F-FDG PET/CT. *Clin. Nucl. Med.* 44, 493–494. doi: 10.1097/RLU.0000000000002463

Greco, A. F. (2013). Cancer of unknown primary or unrecognized adnexal skin primary carcinoma? Limitations of gene expression profiling diagnosis. *J. Clin. Oncol.* 31, 1479–1481. doi: 10.1200/JCO.2012.47.1615

Gross-Goupil, M., Massard, C., Lesimple, T., Merrouche, Y., Blot, E., Loriot, Y., et al. (2012). Identifying the primary site using gene expression profiling in patients with carcinoma of an unknown primary (CUP): a feasibility study from the GEFCAPI. *Onkologie* 35, 54–55. doi: 10.1159/000336300

Hainsworth, J. D., Rubin, M. S., Spigel, D. R., Boccia, R. V., Raby, S., Quinn, R., et al. (2013). Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the sarah cannon research institute. *J. Clin. Oncol.* 31, 217–223. doi: 10.1200/JCO.2012.43.3755

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929–944. doi: 10.1016/j.cell.2014.06.049

Hudis, C. A. (2007). Trastuzumab–mechanism of action and use in clinical practice. *N. Engl. J. Medi.* 357, 39–51.

Huebner, G., Morawietz, L., Floore, A., Buettner, R., Folprecht, G., Stork-Sloots, L., et al. (2007). Comparative analysis of microarray testing and immunohistochemistry in patients with carcinoma of unknown primary – CUP syndrome. *Eur. J. Cancer Suppl.* 5, 90–91.

Hyphantis, T., Papadimitriou, I., Petrakis, D., Fountzilas, G., Repana, D., Assimakopoulos, K., et al. (2013). Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site. *PsychoOncol.* 22, 2009–2015. doi: 10.1002/pon.3244

Janick, S., Elodie, L.-M., Marie-Christine, M., Philippe, R., and Marius, I. (2018). Immunohistochemistry for diagnosis of metastatic carcinomas of unknown primary site. *Cancers* 10:108. doi: 10.3390/cancers10040108

Kandalaft, P. L., and Gown, A. M. (2015). Practical applications in immunohistochemistry: carcinomas of unknown primary site. *Arch. Pathol. Lab. Med.* 140, 508–523. doi: 10.5858/arpa.2015-0173-CP

Kao, K. J., Cheng, S. H., and Huang, A. T. (2006). Gene expression profiling for prediction of distant metastasis and survival in primary nasopharyngeal carcinoma. *J. Cli. Oncol.* 24, 5503–5503.

Kuhnen, C., Herter, P., Monse, H., Kahmann, S., Muehlberger, T., Vogt, P. M., et al. (2000). APC and β-catenin in alveolar soft part sarcoma (ASPS) - immunohistochemical and molecular genetic analysis. *Pathol. Res. Pract.* 196, 299–304. doi: 10.1016/s0344-0338(00)80059-x

Kurahashi, I., Fujita, Y., Arao, T., Kurata, T., Koh, Y., Sakai, K., et al. (2013). A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. *PloS One* 8:e63249. doi: 10.1371/journal.pone.0063249

Küsters-Vandevelde, H. V. N., Kruse, V., Van Maerken, T., Boterberg, T., Pfundt, R., Creytens, D., et al. (2017). Copy number variation analysis and methylome profiling of a GNAQ-mutant primary meningeal melanocytic tumor and its liver metastasis. *Exp. Mol. Pathol.* 102, 25–31. doi: 10.1016/j.yexmp.2016.12.006

Kwee, T. C., Basu, S., Cheng, G., and Alavi, A. (2010). FDG PET/CT in carcinoma of unknown primary. *Eur. J. Nucl. Med. Mol. Imaging* 37, 635–644. doi: 10.1007/s00259-009-1295-6

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., et al. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics* 18:508. doi: 10.1186/s12864-017-3906-0

Ma, X. J., Patel, R., Wang, X., Salunga, R., Murage, J., Desai, R., et al. (2006). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch. Pathol. Lab. Med.* 130, 465–473. doi: 10.1043/1543-2165(2006)130[465:MCOHCU]2.0.CO;2

Maatman, T. J., Gupta, M. K., and Montie, J. E. (1984). The role of serum prostatic acid phosphatase as a tumor marker in men with advanced adenocarcinoma of the prostate. *J. Urolo.* 132, 58–60. doi: 10.1016/s0022-5347(17)49463-8

Makhlouf, A. M., Chitikova, Z., Pusztaszeri, M., Berczy, M., and Dibner, C. (2016). Identification of CHEK1, SLC26A4, c-KIT, TPO and TG as new biomarkers for human follicular thyroid carcinoma. *Oncotarget* 7, 45776–45788. doi: 10.18632/oncotarget.10166

Malhi, A., and Gao, R. (2005). PCA-based feature selection scheme for machine defect classification. *Instrument. Meas. IEEE Trans.* 53, 1517–1525.

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2016). TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med. Genomics* 8:58. doi: 10.1186/s12920-015-0130-0

McConechy, M. K., Ding, J., Senz, J., Yang, W., Melnyk, N., Tone, A. A., et al. (2014). Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Modern Pathol.* 27, 128–134. doi: 10.1038/modpathol.2013.107

Meyer, J. G., Liu, S., Miller, I. J., Coon, J. J., and Gitter, A. (2019). Learning drug function from chemical structure with convolutional neural networks and random forests. *J. Chem. Inform. Model.* 59, 4438–4449. doi: 10.1021/acs.jcim.9b00236

Muthukrishnan, M., and Rohini, R. (2016). "LASSO: a feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore.

Nota, B. (2016). Gogadget: an R Package for interpretation and visualization of go enrichment results. *Mol. Inform.* 36:1600132. doi: 10.1002/minf.201600132

Ohka, F., Yamamichi, A., Kurimoto, M., Motomura, K., Tanahashi, K., Suzuki, H., et al. (2017). A novel all-in-one intraoperative genotyping system for IDH1-mutant glioma. *Brain Tumor Pathol.* 34, 91–97. doi: 10.1007/s10014-017-0281-0

Ohno, M., Narita, Y., Miyakita, Y., Matsushita, Y., and Shibui, S. (2013). Secondary glioblastomas with IDH1/2 mutations have longer glioma history from preceding lower-grade gliomas. *Brain Tumor Pathol.* 30, 224–232. doi: 10.1007/s10014-013-0140-6

Oien, K. A., and Dennis, J. L. (2012). Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling. *Ann. Oncol.* 23(Suppl._10), x271–x277. doi: 10.1093/annonc/mds357

Pieper, R. O., Ohba, S., and Mukherjee, J. (2014). Mutant idh1-driven cellular transformation increases rad51-mediated homologous recombination and Temozolomide (Tmz) resistance. *Cancer Res.* 74, 4836–4844. doi: 10.1158/0008-5472.CAN-14-0924

Qu, K. Z., Li, H., Whetstone, J. D., Sferruzza, A. D., and Bender, R. A. (2007). Molecular identification of carcinoma of unknown primary (CUP) with gene expression profiling. *J. Clin. Oncol.* 25, 21024–21024.

Rizwan, M., and Zulfiqar, M. (2010). Carcinoma of unknown primary. *J. Pakistan Med. Assoc.* 60, 598–599.

Rosenwald, S., Gilad, S., Benjamin, S., Lebanony, D., Dromi, N., Faerman, A., et al. (2010). Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. *Mod. Pathol.* 23, 814–823. doi: 10.1038/modpathol.2010.57

Sandri, M., and Zuccolotto, P. (eds) (2006). *Variable Selection Using Random Forests. Data Analysis, Classification and the Forward Search*. Berlin: Springer.

Shaw, P. H. S., Adams, R., Jordan, C., and Crosby, T. D. L. (2007). A clinical review of the investigation and management of carcinoma of unknown primary in a single cancer network. *Clin. Oncol.* 19, 87–95. doi: 10.1016/j.clon.2006.09.009

Sheffield, B. S., Tessier-Cloutier, B., Li-Chang, H., Shen, Y., Pleasance, E., Kasaian, K., et al. (2016). Personalized oncogenomics in the management of gastrointestinal carcinomas-early experiences from a pilot study. *Curr. Oncol.* 23, 68–73. doi: 10.3747/co.23.3165

Varadhachary, G. R., Raber, M. N., Matamoros, A., and Abbruzzese, J. L. (2008). Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions. *Lancet Oncol.* 9, 596–599. doi: 10.1016/S1470-2045(08)70151-7

Vihko, P. T., Quintero, I., Rönkä, A. E., Herrala, A., Jäntti, P., Porvari, K., et al. (2005). Prostatic acid phosphatase (PAP) is PI(3)P-phosphatase and its inactivation leads to change of cell polarity and invasive prostate cancer. *Cancer Res.* 65, 62–78.

Voigt, J. J. (2008). Immunohistochemistry: a major progress in the classification of carcinoma of unknown primary. *Oncologie* 10, 693–697.

Waardenberg, A. J., Bassett, S. D., Bouveret, R., and Harvey, R. P. (2016). Erratum to: 'CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments'. *BMC Bioinform.* 17:179. doi: 10.1186/s12859-015-0701-2

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, 293–312. doi: 10.1093/nar/gky400

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for comparing biological themes among gene clusters. *Omics J. Integra. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Conflict of Interest: BW, JL, XL, QL, GT, and JY were employed by the company Geneis Beijing Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 June 2020
doi: 10.3389/fbioe.2020.00541

Classification of Infected Necrotizing Pancreatitis for Surgery Within or Beyond 4 Weeks Using Machine Learning

Lan Lan[1]*, *Qiang Guo*[2], *Zhigang Zhang*[3,4], *Weiling Zhao*[4], *Xiaoyan Yang*[1], *Huimin Lu*[5], *Zongguang Zhou*[6] *and Xiaobo Zhou*[4]

[1] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China, [2] Vascular Surgery, West China Hospital, Sichuan University, Chengdu, China, [3] School of Information Management and Statistics, Hubei University of Economics, Wuhan, China, [4] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [5] Pancreatic Surgery, West China Hospital, Sichuan University, Chengdu, China, [6] Institute of Digest Surgery, West China Hospital, Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:
*Shu Tao,
UCLA Jonsson Comprehensive Cancer Center, United States*

Reviewed by:
*Khanh N. Q. Le,
Taipei Medical University, Taiwan
Ximiao He,
Huazhong University of Science and Technology, China*

***Correspondence:**
*Lan Lan
lanl@scu.edu.cn*

Specialty section:
*This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Bioengineering and Biotechnology*

Received: *02 March 2020*
Accepted: *05 May 2020*
Published: *04 June 2020*

Citation:
*Lan L, Guo Q, Zhang Z, Zhao W, Yang X, Lu H, Zhou Z and Zhou X (2020) Classification of Infected Necrotizing Pancreatitis for Surgery Within or Beyond 4 Weeks Using Machine Learning.
Front. Bioeng. Biotechnol. 8:541.
doi: 10.3389/fbioe.2020.00541*

Background: The timing of surgery for necrotizing pancreatitis remains a matter of controversial debate, which has not been resolved by randomized controlled trial (RCT). This study aims to classify surgical timing within or beyond 4 weeks for patients with infected necrotizing pancreatitis by using machine learning methods.

Methods: This study analyzed 223 patients who underwent surgery for infected pancreatic necrosis at West China Hospital of Sichuan University. We used logistic regression, support vector machine, and random forest with/without the simulation of generative adversarial networks to classify the surgical intervention within or beyond 4 weeks in the patients with infected necrotizing pancreatitis.

Results: Our analyses showed that interleukin 6, infected necrosis, the onset of fever and C-reactive protein were important factors in determining the timing of surgical intervention ($<$ 4 or $\geq$ 4 weeks) for the patients with infected necrotizing pancreatitis. The main factors associated with postoperative mortality in patients who underwent early surgery ($<$ 4 weeks) included modified Marshall score on admission and preoperational modified Marshall score. Preoperational modified Marshall score, time of surgery, duration of organ failure and onset of renal failure were important predictive factors for the postoperative mortality of patients who underwent delayed surgery ($\geq$ 4 weeks).

Conclusions: Machine learning models can be used to predict timing of surgical intervention effectively and key factors associated with surgical timing and postoperative survival are identified for infected necrotizing pancreatitis.

Keywords: classification, surgery, timing, machine learning, necrotizing pancreatitis

INTRODUCTION

Necrotizing pancreatitis occurs in about 20% of patients suffering from acute pancreatitis (AP) (Banks, 1997). The current management guideline for necrotizing pancreatitis from IAP/APA (Working Group IAP/APA Acute Pancreatitis Guidelines, 2013) recommends delaying the timing of surgery until 4 or more weeks after initial necrotizing presentation to become walled-off shown in an addition file (**Supplementary Figure 1**). However, some patients with necrotizing pancreatitis will die before 4 weeks from the onset of AP. Therefore, how to identify those patients is an urgent problem to be solved. In addition to IAP/APA guideline, recommendations for surgical timing of necrotizing pancreatitis in the United States, United Kingdom, Italy, and Japan are also delayed as far as possible, without recommendations for individuals (Association et al., 2005; Tenner et al., 2013; Pezzilli et al., 2015; Yokoe et al., 2015).

Guo et al. concluded that the postoperative mortality of patients in 2 weeks with necrotizing pancreatitis was much higher than that after 2 weeks, and the prognosis of patients who did surgery before 4 weeks in necrotizing pancreatitis without persistent organ failure (POF) was same with that of patients who did surgery after 4 weeks in necrotizing pancreatitis without POF (Guo et al., 2014). A systematic review suggested that debridement should be done at least 12 days later for adult patients with necrotizing pancreatitis (Mowery et al., 2017). The first drainage time in step-up approach was 3.5–75.5 days from the onset of AP (Mowery et al., 2017). The timing of surgical intervention in necrotizing pancreatitis is controversial. A randomized controlled trial (RCT) which was established to optimize timings of surgery following PCD in patients with infected pancreatic necrosis was forced to stop early due to practical difficulties (Shenvi et al., 2016). The surgical timing problem has not been resolved by RCT.

What's more, infection and organ failure have long been used as key factors in determining whether or not to undergo surgery and are considered as the determinants of mortality for the patients with necrotizing pancreatitis. Surgical indications for the patients with necrotic pancreatitis are determined empirically among clinicians (Gomatos et al., 2015; Van Grinsven et al., 2016). A prospective study observed that POF in the first week was more likely to determine mortality than infection in patients with necrotizing pancreatitis (Guo et al., 2013). While a prospective cohort study from the Netherlands showed that there were no associations between infection, onset of organ failure, duration of organ failure and mortality in the patients with necrotizing pancreatitis (Schepers et al., 2018). These findings are inconsistent. Additionally, current studies cannot explain the relationship between the suggested surgical indications of necrotizing pancreatitis, mortality and surgical timing (Van Grinsven et al., 2016).

Nowadays, artificial intelligence (AI) is increasingly used in medicine (Nature Medicine, 2019). Therefore, we applied the machine learning and deep learning methods in AI to extract the clinical features from the patients with infected necrotizing pancreatitis who received early surgery in West China Hospital of Sichuan University and analyzed the associations between early surgical treatment, organ failure, infection and clinical predictors. We also identified the key factors associated with patients' mortality following early (<4 weeks) or late (≥4 weeks) surgery.

MATERIALS AND METHODS

Patients and Treatment Protocol

A total of 223 patients (median age: 43 years old, male: 60.99%) were analyzed in this study. Those patients were hospitalized and operated due to infected necrotizing pancreatitis in West China Hospital of Sichuan University from January 2009 to June 2012. The AP was diagnosed according to the classification system of 2012 revision of the Atlanta edition, and pancreatic necrosis or peripancreatic necrosis was determined by contrast-enhanced computed tomography (CECT). Its treatment protocol was reported previously (Guo et al., 2013, 2014). The patients with severe clinical signs of persistent degeneration were operated before 4 weeks and the remaining patients were operated after 4 weeks from the onset of AP. This study was approved by the ethics review board of West China Hospital of Sichuan University, and the need for informed consent was waived owing to the retrospective nature of the study.

Clinical Data Collection

The clinical data related above patients were collected, including infection, organ failure, operation time, postoperative mortality, postoperative complications, during hospitalization of those patients, etc. The collecting procedure and definitions of the indicators were described previously (Guo et al., 2013, 2014).

Statistical Analysis

To classify surgical timing, there was nothing worthwhile to learn about a failed surgery. For example, if a patient died after surgery, we regarded this kind of case as a failed one. So, the successful surgery needed to be learned. We assumed that the best surgical timing was the actual time of a successful surgery. Based on the time from the onset of AP to surgical intervention, the patients were divided into the early (<4 weeks) and delayed (≥4 weeks) surgery groups. The baseline conditions of these patients were analyzed, including organ failure, infection, etc. T-test, and Chi-square test were used to evaluate the difference between the two groups. We then analyzed the factors that affect surgical timing and the factors associated with postoperative mortality by feature selection. Finally, we used multiple classifiers to classify the patients and compared the classifiers' performance. Variables with a $p < 0.05$ were considered to be statistically significant.

Three classifiers were used in this study, including logistic regression (LR), support vector machine (SVM) and random forest (RF) (Le, 2019; Le et al., 2019b). The LR is a commonly used statistic model in the healthcare industry and SVM is a popular machine learning approach. RF is a classifier that uses multiple trees to train and predict and has both features of high accuracy and balancing errors when analyzing unbalanced classification data sets. In order to find predictors of postoperative mortality at different surgical timings, in addition to feature selection and classification of surgical timing in survived patients after surgery, we performed feature selection and classification of postoperative death in the early and delayed surgery. Finally, we divided the patients into three groups based on the surgical time and mortality for classification analyses.

The survived patients after surgery ($n = 186$) were divided into the early group ($n = 73$) and the delayed group ($n = 113$), to predict whether surgical treatment should be performed early;

The patients received early surgery ($n = 106$) were divided into the death group ($n = 33$) and survival group ($n = 73$), to predict the death rate of patients after receiving an early surgery.

The patients with delayed surgery ($n = 117$) were divided into death group ($n = 4$) and survival group ($n = 113$), to predict the death rate of patients after delayed surgery.

To solve the problem of positive and negative sample imbalance and small sample size, which will severely affect the performance of classifiers, we used generative adversarial networks (GAN) to generate simulated samples, which had the same distributions as the real samples (Creswell et al., 2017). GAN, a recently developed deep learning approach (Goodfellow et al., 2014), shows promising simulation performances in many fields (Deshpande, 2013; Santana and Hotz, 2016; Li et al., 2017; Pascual et al., 2017), such as image synthesis, language processing, etc. Douzas and Bacao (2017) used a conditional version (referring to each category) of GAN to approximate the true data distribution and generated data for the minority class of various imbalanced datasets. To improve the effectiveness of a classifier, Fiore et al. (2017) trained a GAN model to mimic the original minority class examples and then merged the synthetic examples with training data into an augmented training set. More importantly, by using variant of GAN, Baowaly et al. (2019) have proved that GAN can adequately learn the data distribution of real electronic health records and efficiently generate realistic synthetic electronic health records. GAN is a powerful generation model (Goodfellow et al., 2014; Douzas and Bacao, 2017; Fiore et al., 2017; Wang et al., 2017). Therefore, we applied GAN to electronic medical records to investigate the timing of surgical intervention for the patients with infected necrotizing pancreatitis. In this study, the data was randomly divided into training dataset and testing dataset according to the ratio of 4:1. The real training dataset were used to train the simulated samples to optimize GAN parameters. The simulated samples generated by the GAN generator were filtered by the GAN discriminator. The simulated samples after filtration were tested by LR, SVM, and RF (**Figure 1**).

We used several classification indicators to determine the classification performance of our models, including accuracy, precision, recall, F1-measure and area under curve (AUC) (Le et al., 2017, 2019a). Accuracy provides a percentage of correct classification. Precision is a measurement of how many positive classifications are actual positive observations. Recall, a proportion of all real positive observations that are correct, is a measure of how many actual positive observations are classified correctly. F1-measure, the harmonic mean of precision and recall, is an "average" of both precision and recall. AUC is the area under the ROC curve. The greater the value of the indicators, the better the model performance. We combined multiple evaluation indicators to evaluate the performance of the models. The simulation for GAN was calculated in Python software and others were conducted using R software.

RESULTS

Characteristics of Survived Patients After Early or Delayed Surgery

We compared the major organ failure in the early and delayed surgery groups as shown in an additional file (**Supplementary Table 1**). In general, there were no differences in POF and the number of organ failure systems between the two groups. The proportion of renal failure in the early group was higher than that in the delayed group. Onsets of renal failure and multiple organ failure in the delayed group were earlier, but the duration of organ failure was shorter. In terms of the preoperative POF, more than half of patients with POF were recovered before surgery, and the proportion of POF in the early group was higher than that in the delayed group. As shown in another additional file (**Supplementary Table 2**), the median time and interquartile range of surgery for the early group were 21 and 6 days and the delayed group was 37 and 21 days, respectively. More patients received continuous renal replacement therapy (CRRT) in the early group than those in the delayed group. The delayed group had a higher proportion of infected necrosis. The onset of fever in the early group was earlier than that in the delayed group. The proportion of abnormal interleukin 6 (IL-6) level in the delayed group was higher than that in the early group. There was no difference in the modified Marshall score between the two groups, but preoperational modified Marshall score was higher in the early group. The proportions of intra-abdominal bleeding and re-intervention were higher in early group. Age, length of hospital stay and gender composition ratio were similar between the two groups. There were no differences between the two groups in blood culture, sputum, white blood cell (WBC), C-reactive protein (CRP), procalcitonin (PCT), enterocutaneous fistula, and new-onset organ failure.

Predictors of Surgical Timing and Postoperative Mortality and Classification Performance

In order to classify the surgical timing (<4 or ≥4 weeks), the patients were divided into three groups as shown in the Methods based on the time of surgery and postoperative mortality, including the patients surviving after surgery (group 1), the patients received earlier surgery (group 2) and the patients underwent delayed surgery (group 3). LR, SVM, RF w/wo GAN were adopted to predict surgical timing from the three groups of patients, respectively, where LR was used as a statistical model and SVM and RF were used as machine learning models. The first group of patients was used to assess the predictors of surgical timing. The second group of patients was used for evaluating the predictors of mortality from early surgery and the third group of patients for the predictors of mortality from delayed. We used the stepwise selection procedures for the selection of independent variables (predictors) in LR. The Boruta function in R was applied

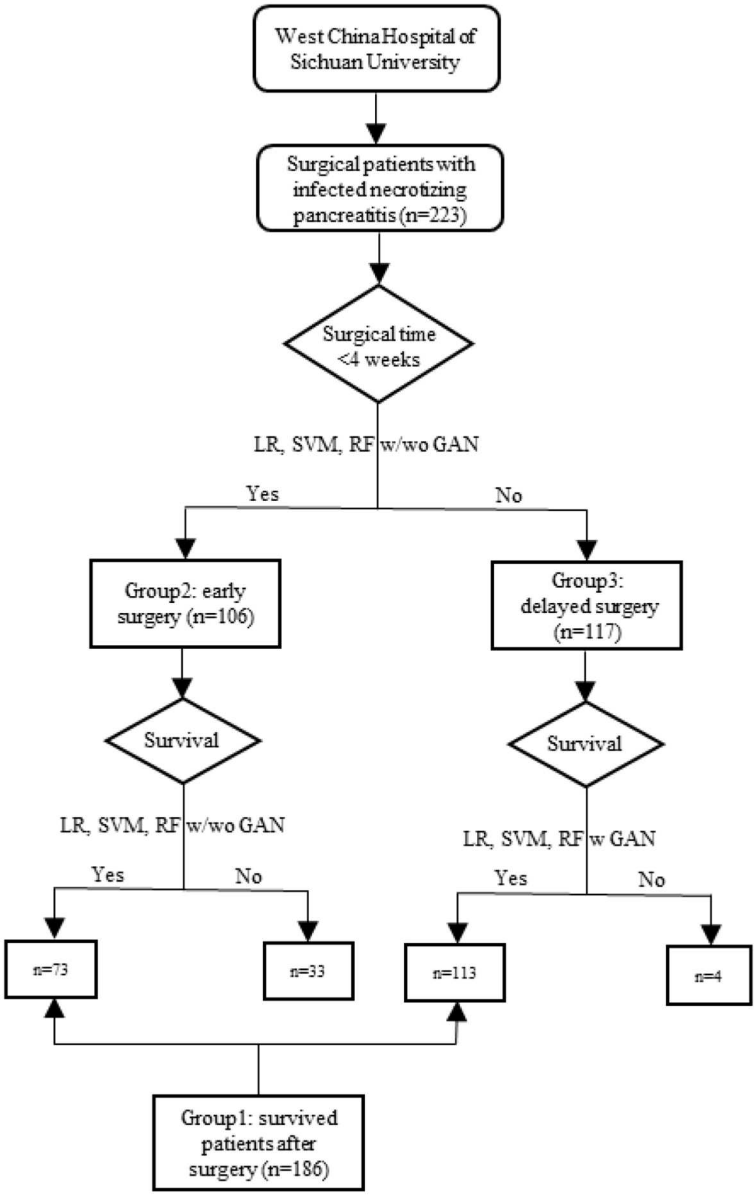

West China Hospital of Sichuan University
Surgical patients with infected necrotizing pancreatitis (n=223)
Surgical time <4 weeks
LR, SVM, RF w/wo GAN
Yes
No
Group2: early surgery (n=106)
Group3: delayed surgery (n=117)
Survival
LR, SVM, RF w/wo GAN
Yes
No
Survival
LR, SVM, RF w GAN
Yes
No
n=73
n=33
n=113
n=4
Group1: survived patients after surgery (n=186)

FIGURE 1 | Flowchart of the study.

to select important features in SVM, where the value of *meanImp* indicates the importance of a predictor. RF itself comes with a feature selection function, where the value of *MeanDecreaseGini* represents the importance of a feature. The larger the value, the more important it is.

The analysis results for group 1 indicated that IL-6, infected necrosis, the onset of fever and CRP were the important factors for surgical timing of patients with necrotizing pancreatitis (**Table 1**). The results of the following models are derived from the testing dataset. We also assessed the classification accuracy of the three models. The classified accuracy of RF (0.80) was higher than SVM (0.78) and LR (0.71). With the simulation by the GAN, the classification accuracies for all of the three models were improved. GAN-RF (accuracy: 0.89) had a better performance than GAN-SVM (0.84) and GAN-LR (0.83). The recall rates also reached 1 (**Table 2**).

We assessed the key factors affecting patient mortality in the group 2 patients using LR, SVM, and FR models, respectively. As shown in the **Table 3**, top-ranked factors associated with patient mortality include the modified Marshall score on admission and preoperational modified Marshall score. By combining with GAN, the classification accuracies of the three models

TABLE 1 | Top important features for survived patients after early surgery (<4 weeks) compared with survived patients after delayed surgery (≥4 weeks).

LR		SVM		RF	
Significant variable	$\left\|\beta\right\|$	**Confirmed variable**	***meanImp***	**Variable**	***MeanDecreaseGini***
Pulmonary failure	21.89	Onset of fever	21.40	Onset of fever	14.25
POF	19.77	Infected necrosis	13.37	Age	9.18
Renal failure	5.11	IL-6	10.82	Infected necrosis	5.10
IL-6	2.71	Modified Marshall score pre-operation	9.58	Modified Marshall score on admission	4.35
PCT	1.36	Modified Marshall score on admission	7.99	CRP	3.54
Duration of organ failure	1.20	Sputum	7.63	IL-6	3.40
Infected necrosis	1.11	CRRT	6.79	Modified Marshall score pre-operation	3.31
WBC	0.78	CRP	6.12	Duration of organ failure	2.66
Onset of fever	0.68	Onset of renal failure	5.49	WBC	1.63
CRP	0.31	Renal failure	5.42	Sputum	1.62

LR, logistic regression; SVM, support vector machine; RF, random forest; POF, persistent organ failure; IL-6, interleukin 6; PCT, procalcitonin; WBC, white blood cell; CRP, C-reactive protein; CRRT, continuous renal replacement therapy.

TABLE 2 | Classification performance for survived patients after early (<4 weeks) or delayed surgery (≥4 weeks).

Model	Accuracy	Precision	Recall	F1-Measure	AUC
LR	0.71	0.70	0.53	0.58	0.71
SVM	0.78	0.77	0.63	0.67	0.75
RF	0.80	0.75	0.70	0.71	0.78
GAN-LR	0.83	0.62.	1.00	0.76	0.90
GAN-SVM	0.84	0.72	1.00	0.84	0.86
GAN-RF	0.89	0.80	1.00	0.88	0.90

LR, logistic regression; SVM, support vector machine; RF, random forest; GAN, generative adversarial networks.

for mortality in early surgery patients were largely improved. GAN-RF (0.99) and GAN-SVM (0.99) had a better performance in evaluating the key factors than GAN-LR (0.90) (**Table 4**).

As shown in the **Table 5**, the modified Marshall score pre-operation was predicted by all three models as an important factor for the mortality of patients who underwent delayed surgery. The time of surgery, duration of organ failure and onset of renal failure were top 5-ranked features predicted by SVM and RF models (**Table 6**). Due to the unbalanced positive and negative samples, we simulated this group of samples using GAN first and then did classification analysis for the postoperative mortality using three classifiers. The classification accuracies of GAN-LR, GAN-SVM, and GAN-RF were 0.97, 0.99, and 0.99, respectivel (**Table 6**).

DISCUSSION

This study has two main highlights. (1) We compared the performance of machine learning models with a common statistic model (LR) and the performance of machine learning models were better. (2) We identified the key factors associated with surgical timing (<4 or ≥4 weeks) and postoperative survival for infected necrotizing pancreatitis and predicted the surgical timing by applying machine learning models.

An international survey shows that 55% of pancreatic specialists would wait for the effect of antibodies and postpone surgical management for the patients with infected pancreatic necrosis, whereas 45% of specialists would take an immediately action of surgical treatment after diagnosis (Abdelhafez et al., 2015). The time of operation varies greatly. Therefore, it is necessary to demonstrate if the patient with necrotizing pancreatitis needs early or delayed surgery individually. Previous studies using organ failure and infection as predictors of death obtained controversial results (Guo et al., 2013, 2014; Schepers et al., 2018). In our study, we assessed the impact of multiple clinical factors and comprehensive scores on surgical timing and postoperative mortality for the patients received the early or delayed surgery.

Early studies showed that the mortality of patients who received surgery within 2 weeks was much higher than that

TABLE 3 | Top important features for mortality after early surgery (<4 weeks).

LR		SVM		RF	
Significant variable	$\left\|\beta\right\|$	**Confirmed variable**	***meanImp***	**Variable**	***MeanDecreaseGini***
CRRT	744.77	Renal failure	10.32	Modified Marshall score on admission	5.76
Intra-abdominal bleeding	424.34	Onset of renal failure	10.19	Renal failure	4.87
Blood culture	373.11	Onset of fever	9.87	Onset of multiple organ failure	4.70
New-onset organ failure	297.63	Re-intervention	9.38	Onset of renal failure	3.97
Modified Marshall score on admission	147.89	Modified Marshall score on admission	9.02	Number of organ failure systems	3.23
POF pre-operation	134.08	Onset of multiple organ failure	7.54	Modified Marshall score pre-operation	2.74
Modified Marshall score pre-operation	97.74	Number of organ failure systems	7.51	Onset of fever	1.78
WBC	75.50	Multiple organ failure	7.45	Re-intervention	1.71
PCT	72.39	POF pre-operation	6.68	Duration of organ failure	1.62
Age	0.35	Modified Marshall score pre-operation	6.48	Age	1.62

LR, logistic regression; SVM, support vector machine; RF, random forest; POF, persistent organ failure; PCT, procalcitonin; WBC, white blood cell; CRRT, continuous renal replacement therapy.

TABLE 4 | Classification performance for mortality after early surgery (<4 weeks).

Model	Accuracy	Precision	Recall	F1-Measure	AUC
LR	0.90	0.82	0.88	0.83	0.94
SVM	0.94	0.94	0.90	0.91	0.93
RF	0.94	0.85	1.00	0.90	0.96
GAN-LR	0.90	0.89	0.92	0.90	0.97
GAN-SVM	0.99	0.99	0.99	0.99	0.99
GAN-RF	0.99	0.99	0.99	0.99	0.99

LR, logistic regression; SVM, support vector machine; RF, random forest; GAN, generative adversarial networks.

of surgery after 2 weeks (Besselink et al., 2007; Guo et al., 2013, 2014; Schepers et al., 2018), suggesting that early surgery should be conducted between 2 and 4 weeks. According to our analysis with multiple classifiers, IL-6, infected necrosis, onset of fever and CRP are important factors associated with the timing of surgery, which is consistent with the surgical indications used in clinic. The modified Marshall score is one of common factors used to assess patient mortality. Our analysis indicated that the mortality of the early surgery group was associated with the preoperative modified Marshall score and the modified Marshall score assessed at admission, suggesting that the modified Marshall score should be monitored in a real time for prediction. According to the Revised Atlanta Classification, organ failure is determined by modified Marshall score. The preoperative modified Marshall score was associated with the mortality after delayed surgery. Only two meaningful variables were obtained through stepwise regression of LR, including the preoperative modified Marshall score and circulatory failure. The preoperative modified Marshall score, the time of surgery, duration of organ failure and onset of renal failure were among of the top five important features selected by SVM and RF. A recent multicenter prospective cohort study reported that POF and multiple organ failure were the major determinants of AP severity, and the presence of infection was not associated with higher mortality (Sternby et al., 2017), consistent with our findings.

According to our knowledge, this is the first time to apply SVM and RF to predict the timing of surgery and postoperative mortality of patients with infected necrotizing pancreatitis. The classification performance of RF and SVM was better than LR. Especially when GAN was applied in the simulation, the accuracies were obviously improved. It is most likely because GAN can generate simulation samples with the same distribution as the actual samples, enhancing the sample size. In term of

TABLE 5 | Top important features for mortality after delayed surgery (≥4 weeks).

LR		SVM		RF	
Significant variable	$\lvert\beta\rvert$	**Confirmed variable**	***meanImp***	**Variable**	***MeanDecreaseGini***
Circulatory failure	60.16	Time of surgery	12.13	Time of surgery	167.62
Modified Marshall score pre-operation	14.17	Duration of organ failure	9.52	Duration of organ failure	122.11
		Modified Marshall score pre-operation	8.64	Onset of renal failure	111.15
		Onset of renal failure	8.60	Onset of fever	108.94
		Male	8.35	Modified Marshall score pre-operation	66.56
		Onset of multiple organ failure	7.34	Onset of multiple organ failure	52.52
		Onset of fever	7.30	Onset of single organ failure	45.50
		Modified Marshall score on admission	7.19	Onset of POF	45.18
		Number of organ failure systems	7.17	Male	43.82
		Blood culture	6.99	POF pre-operation	32.31

LR, logistic regression; SVM, support vector machine; RF, random forest; POF, persistent organ failure.

TABLE 6 | Classification performance for mortality after delayed surgery (≥4 weeks).

Model	Accuracy	Precision	Recall	F1-Measure	AUC
GAN-LR	0.97	0.69	0.70	0.63	0.92
GAN-SVM	0.99	0.80	1.00	0.88	0.99
GAN-RF	0.99	0.94	0.99	0.96	0.99

LR, logistic regression; SVM, support vector machine; RF, random forest; GAN, generative adversarial networks.

model classification performance, the classification accuracies of three models were high. Therefore, based on the patient's routine laboratory test and organ failure status, we can apply the classifiers to predict whether the patient should undergo early or delayed surgery individually to reduce patient mortality. Our classification results provide good references for clinicians to make personized surgical plans for patients with infected necrotizing pancreatitis.

However, there are some limitations of this study. Since the categorical variables cannot be applied to the traditional GAN, we changed the categorical variables into continuous variables and then put them into the GAN model. Although we have reached a conclusion consistent with Baowaly et al. by using our proposed GAN, we need to further verify with more samples.

In summary, we (1) applied a better machine learning model compared with a statistic model to predict the surgical timing (<4 or ≥ 4 weeks) in patients with infected necrotizing pancreatitis; (2) identified the key factors associated with surgical timing and postoperative survival for infected necrotizing pancreatitis and predicted the surgical timing by applying machine learning models; and (3) provided good references for clinicians in developing personalized surgical plans for patients with infected necrotizing pancreatitis.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

ETHICS STATEMENT

This study was approved by the ethics review board of West China Hospital of Sichuan University, and the need for informed consent was waived owing to the retrospective nature of the study.

AUTHOR CONTRIBUTIONS

Data collection: QG. Data analysis: LL and ZZha. Data interpretation: XY and HL. Writing of the manuscript: LL. Research conception: XZ and ZZho. Critical revision of the manuscript: WZ.

FUNDING

This work was supported by Sichuan Science and Technology Program (No. 2019YFS0147), 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University

(No. ZYJC18010), 1·3·5 project for disciplines of excellence-Clinical Research Incubation Project, West China Hospital, Sichuan University (No. 2019HXFH022), and Post-Doctor Research Project, West China Hospital, Sichuan University (No. 2019HXBH039).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00541/full#supplementary-material

REFERENCES

Abdelhafez, M., Andersson, R., Andren-Sandberg, A., Ashley, S., Baal, M. V., Baron, T., et al. (2015). Diagnostic strategy and timing of intervention in infected necrotizing pancreatitis: an international expert survey and case vignette study. *HPB* 18, 49–56. doi: 10.1111/hpb.12491

Association, O. U. G. S., Working, P. O. T. B., Association, O. S. O. G., and Pancreatic, S. O. G. B., and UK Working Party on Acute Pancreatitis (2005). UK guidelines for the management of acute pancreatitis. *Gut* 54(Suppl. 3), i1–i9. doi: 10.1136/gut.2004.057026

Banks, P. A. (1997). Practice guidelines in acute pancreatitis. *Am. J. Gastroenterol.* 92, 377–386.

Baowaly, M. K., Lin, C., Liu, C., and Chen, K. (2019). Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* 26, 228–241. doi: 10.1093/jamia/ocy142

Besselink, M. G. H., Verwer, T. J., Schoenmaeckers, E. J. P., Erik, B., Ridwan, B. U., Visser, M. R., et al. (2007). Timing of surgical intervention in necrotizing pancreatitis. *Arch. Surg.* 142, 1194–1201. doi: 10.1001/archsurg.142.12.1194

Creswell, A., White, T., Dumoulin, V., Kai, A., Sengupta, B., and Bharath, A. A. (2017). Generative adversarial networks: an overview. *IEEE Signal. Proc. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Deshpande, A. (2013). *Deep Learning Research Review Week 1: Generative Adversarial Nets*. Available online at: https://adeshpande3.github.io/Deep-Learning-Research-Review-Week-1-Generative-Adversarial-Nets (accessed July 17, 2018).

Douzas, G., and Bacao, F. (2017). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Syst. Appl.* 82, 40–52. doi: 10.1016/j.eswa.2017.03.073

Fiore, U., Santis, A. D., Perla, F., Zanetti, P., and Palmieri, F. (2017). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Inform. Sci.* 479, 448–455. doi: 10.1016/j.ins.2017.12.030

Gomatos, I. P., Halloran, C. M., Ghaneh, P., Raraty, M. G., Polydoros, F., Evans, J. C., et al. (2015). Outcomes from minimal access retroperitoneal and open pancreatic necrosectomy in 394 patients with necrotizing pancreatitis. *Ann. Surg.* 263, 992–1001. doi: 10.1097/SLA.0000000000001407

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets,"in *International Conference on Neural Information Processing Systems* (Montreal, QC), 2672–2680.

Guo, Q., Li, A., Xia, Q., Liu, X., Tian, B., Mai, G., et al. (2013). The role of organ failure and infection in necrotizing pancreatitis: a prospective study. *Ann. Surg.* 259, 1201–1207. doi: 10.1097/SLA.0000000000000264

Guo, Q., Li, A., Xia, Q., Lu, H., Ke, N., Du, X., et al. (2014). Timing of intervention in necrotizing pancreatitis. *J. Gastrointest. Surg.* 18, 1770–1776. doi: 10.1007/s11605-014-2606-1

Le, N. Q., Ho, Q. T., and Ou, Y. Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N. Q. K. (2019). iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol. Genet. Genomics* 294, 1173–1182. doi: 10.1007/s00438-019-01570-y

Le, N. Q. K., Huynh, T., Yapp, E. K. Y., and Yeh, H. (2019a). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. *Comput. Meth. Prog. Biol.* 177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Le, N. Q. K., Yapp, E. K. Y., Ho, Q., Nagasundaram, N., Ou, Y., and Yeh, H. (2019b). iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* 571, 53–61. doi: 10.1016/j.ab.2019.02.017

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXViv [preprint]*. arXiv:1701.06547.

Mowery, N. T., Bruns, B. R., MacNew, H. G., Agarwal, S., Enniss, T. M., Khan, M., et al. (2017). Surgical management of pancreatic necrosis: a practice management guideline from the Eastern Association for the Surgery of Trauma. *J. Trauma Acute Care Surg.* 83, 316–327. doi: 10.1097/TA.0000000000001510

Nature Medicine (2019). *Digital Medicine*. Available online at: https://www.nature.com/collections/egjifhdcih (accessed March 1, 2019).

Pascual, S., Bonafonte, A., and Serrà, J. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. *arXViv [preprint]*. arXiv:1703.09452.

Pezzilli, R., Zerbi, A., Campra, D., Capurso, G., Golfieri, R., Arcidiacono, P. G., et al. (2015). Consensus guidelines on severe acute pancreatitis. *Digest Liver Dis.* 47, 532–543. doi: 10.1016/j.dld.2015.03.022

Santana, E., and Hotz, G. (2016). Learning a Driving Simulator. *arXViv [preprint]*. arXiv:1608.01230.

Schepers, N. J., Bakker, O. J., Besselink, M. G., Ahmed Ali, U., Bollen, T. L., Gooszen, H. G., et al. (2018). Impact of characteristics of organ failure and infected necrosis on mortality in necrotising pancreatitis. *Gut* 68:314657. doi: 10.1136/gutjnl-2017-314657

Shenvi, S., Gupta, R., Kang, M., Khullar, M., Rana, S. S., Singh, R., et al. (2016). Timing of surgical intervention in patients of infected necrotizing pancreatitis not responding to percutaneous catheter drainage. *Pancreatology* 16, 778–787. doi: 10.1016/j.pan.2016.08.006

Sternby, H., Bolado, F., Canaval-Zuleta, H. J., Marra-López, C., Hernando-Alonso, A. I., Del-Val-Antoñana, A., et al. (2017). Determinants of severity in acute pancreatitis: a nation-wide multicenter prospective cohort study. *Pancreatology* 17:S67. doi: 10.1016/j.pan.2017.05.212

Tenner, S., Baillie, J., DeWitt, J., and Vege, S. S. (2013). American College of Gastroenterology guideline: management of acute pancreatitis. *Am. J. Gastroenterol.* 108, 1400–1415. doi: 10.1038/ajg.2013.218

Van Grinsven, J., van Santvoort, H. C., Boermeester, M. A., Dejong, C. H., van Eijck, C. H., Fockens, P., et al. (2016). Timing of catheter drainage in infected necrotizing pancreatitis. *Nat. Rev. Gastro Hepat,* 13:306. doi: 10.1038/nrgastro.2016.23

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., and Wang, F. Y. (2017). Generative adversarial networks: introduction and outlook. *IEEE/CAA J. Automat. Sin.* 4, 588–598. doi: 10.1109/JAS.2017.7510583

Working Group IAP/APA Acute Pancreatitis Guidelines, 2013∼Working Group IAP/APA Acute Pancreatitis Guidelines (2013). IAP/APA evidence-based guidelines for the management of acute pancreatitis. *Pancreatology* 13, e1–e15. doi: 10.1016/j.pan.2013.07.063

Yokoe, M., Takada, T., Mayumi, T., Yoshida, M., Isaji, S., Wada, K., et al. (2015). Japanese guidelines for the management of acute pancreatitis: Japanese Guidelines 2015. *J Hepato-Bil-Pan Sci.* 22, 405–432. doi: 10.1002/jhbp.259

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 11 June 2020
doi: 10.3389/fbioe.2020.00590

Diagnosis of Cervical Cancer With Parametrial Invasion on Whole-Tumor Dynamic Contrast-Enhanced Magnetic Resonance Imaging Combined With Whole-Lesion Texture Analysis Based on T2- Weighted Images

*Xin-xiang Li [1], Ting-ting Lin [2], Bin Liu [2] and Wei Wei [2]**

[1] Jiangsu Key Laboratory of Molecular and Functional Imaging, Department of Radiology, Zhongda Hospital, Medical School, Southeast University, Nanjing, China, [2] Department of Radiology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

Purpose: To evaluate the diagnostic value of the combination of whole-tumor dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) and whole-lesion texture features based on T2–weighted images for cervical cancer with parametrial invasion.

Materials and Methods: Sixty-two patients with cervical cancer (27 with parametrial invasion and 35 without invasion) preoperatively underwent routine MRI and DCE-MRI examinations. DCE-MRI parameters (K^{trans}, K_{ep}, and V_e) and texture features (mean, skewness, kurtosis, uniformity, energy, and entropy) based on T2-weighted images were acquired by two observers. All parameters of parametrial invasion and non-invasion were analyzed by one-way analysis of variance. The diagnostic efficiency of significant variables was assessed using receiver operating characteristic analysis.

Results: The invasion group of cervical cancer demonstrated significantly higher K^{trans} (0.335 ± 0.050 vs. 0.269 ± 0.079; $p < 0.001$), lower energy values (0.503 ± 0.093 vs. 0.602 ± 0.087; $p < 0.001$), and higher entropy values (1.391 ± 0.193 vs. 1.24 ± 0.129; $p < 0.001$) than those in the non-invasion group. Optimal diagnostic performance [area under curve [AUC], 0.925; sensitivity, 0.935; specificity, 0.829] could be obtained by the combination of K^{trans}, energy, and entropy values. The AUC values of K^{trans} (0.788), energy (0.761), entropy (0.749), the combination of K^{trans} and energy (0.814), the combination of K^{trans} and entropy (0.727), and the combination of energy and entropy (0.619) were lower than those of the combination of K^{trans}, energy, and entropy values.

Conclusion: The combination of DCE-MRI and texture analysis is a promising method for diagnosis cervical cancer with parametrial infiltration. Moreover, the combination of K^{trans}, energy, and entropy is more valuable than any one alone, especially in improving diagnostic sensitivity.

Keywords: cervical cancer, parametrial invasion, DCE-MRI, texture analysis, T2-weighted imaging

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Teng Li,
Anhui University, China
Zaiyi Liu,
Guangdong Provincial People's Hospital, China
Jihong Sun,
Sir Run Run Shaw Hospital, China

****Correspondence:***
Wei Wei
weiweill@126.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *26 February 2020*
Accepted: *14 May 2020*
Published: *11 June 2020*

Citation:
Li X, Lin T, Liu B and Wei W (2020) Diagnosis of Cervical Cancer With Parametrial Invasion on Whole-Tumor Dynamic Contrast-Enhanced Magnetic Resonance Imaging Combined With Whole-Lesion Texture Analysis Based on T2- Weighted Images. Front. Bioeng. Biotechnol. 8:590. doi: 10.3389/fbioe.2020.00590

INTRODUCTION

Cervical cancer is one of the most common malignant diseases of the female reproductive system, and it seriously threatens women's health and life. Accurate preoperative staging of cervical cancer plays an important role in clinical treatment decisions and prognosis. As a matter of principle, surgery is performed for cervical cancer without parametrial involvement, while tumors with parametrial invasion are treated with radio-chemotherapy. To the best of our knowledge, cervical cancer with parametrial invasion is closely related to recurrence and survival after treatment (Chung et al., 2010; Munagala et al., 2010; Noh et al., 2014; Kong et al., 2016; Xia et al., 2016; Dai et al., 2018). Therefore, accurate diagnosis of cervical cancer with parametrial invasion is of great clinical significance. Parametrial invasion is usually evaluated by conventional magnetic resonance (MR) imaging and gynecological examination. Several previous investigations showed that traditional imaging features, such as full-thickness disruption of the normal cervix stroma with nodular or spiculated lesions extending to the adjacent parametrium on T2-weighted images, were considered to be parametrial invasion (Freeman et al., 2012; Patel-Lippmann et al., 2017); however, image analysis is a subjective procedure with low interobserver agreement. An objective and quantitative method for evaluating parametrial infiltration in clinical practice is needed.

Currently, many new techniques have been applied at the molecular level, such as deep learning, proteomics, and protein interaction network (Wang et al., 2019, 2020; Deng et al., 2020; Hu et al., 2020). Besides, several imaging techniques (Park et al., 2014; Zhou et al., 2016) and radiomics (Mu et al., 2015; Meng et al., 2017) have been reported in the assessment of patients with parametrial invasion to determine the stage and treatment of cervical cancer. Chiappa et al. reported that 3D ultrasound volumes can be used to more precisely define the location and degree of cervical cancer invasion (Chiappa et al., 2015). Several studies show that the apparent diffusion coefficient (ADC) value of cervical cancer is significantly lower in cancer with parametrial invasion than in cancer without parametrial invasion (Park et al., 2014; Woo et al., 2018). Park et al. reported that merging high b-value diffusion-weighted MR imaging with background body signal suppression and T2-weighted high-spatial-resolution imaging could improve diagnostic efficiency of predicting cervical cancer with parametrial infiltration (Park et al., 2014). Recently, histogram analysis of ADC has shown potential to predict outcomes after concurrent chemo-radiotherapy in patients with cervical cancer (Meng et al., 2017). Moreover, the study showed that tracer uptake heterogeneity in tumors characterized by texture features based on fluorodeoxyglucose-positron emission tomography (18-FDG PET) is highly relevant to the stage of cervical cancer (Mu et al., 2015). To our knowledge, however, no reported studies demonstrated the role of dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) or texture analysis in evaluating cervical cancer with parametrial invasion.

The purpose of our study was to investigate the diagnostic value of the combination of whole-tumor volumetric DCE-MRI and texture features based on T2-weighted images for predicting parametrial invasion. These quantitative parameters might improve the diagnostic accuracy of cervical cancer with parametrial infiltration prior to treatment.

MATERIALS AND METHODS

Patients

This study was approved by our institutional review board and the patients provided written informed consent to participate. Seventy-five patients with histopathologically confirmed cervical cancer were admitted to our hospital from September 2017 to April 2019. Routine MRI sequences and DCE-MRI examinations were preoperatively performed. Thirteen patients were excluded according to the following criteria: (1) tumor diameter was <1 cm ($n = 7$); (2) image quality was not available for the next analysis ($n = 4$); (3) diameter of the necrotic lesions in the tumor was more than 5 mm ($n = 2$). Finally, 62 patients (25–56 years old, mean age 45 years) were eligible for the study (27 with parametrial invasion and 35 without invasion), including FIGO stage IA ($n = 14$), IB ($n = 9$), IIA ($n = 12$), IIB ($n = 15$), IIIA ($n = 7$), and IIIB ($n = 5$).

Imaging Acquisition

MRI was acquired by using a 3.0 T Signa HDxT MRI machine (GE Healthcare, USA) with an 8-channel phased array body coil. The scanning program for the pelvis was as follows: unenhanced axial T1-weighted imaging, axial and sagittal T2-weighted imaging, axial T2-weighted imaging with fat saturation, axial diffusion weighted imaging, axial DCE-MRI, and axial and sagittal contrast-enhanced T1-weighted imaging with fat saturation.

T2-weighted imaging was obtained using a fast spin echo sequence. The following protocol was used: repetition time (TR)/echo time (TE), 4600/30 ms; number of excitations, 2; section thickness, 6 mm; intersection gap, 2 mm; field of view (FOV), 240 × 240 mm; matrix size, 320 × 256; and total time, 2 min and 14 s.

DCE-MRI was obtained using T1-weighted fat-suppression images and a three dimensional (3D) liver acceleration volume acquisition (LAVA) sequence during the injection of 0.1 mmol/kg of gadodiamide (Omniscan, GE Healthcare, USA) at a rate of 2 ml/s and following a 20-mL saline flush at the same rate. The contrast medium was injected after the acquisition of three sets of pre-contrast T1 mapping using three flip angles: 5, 10, and 15 (total: 43 dynamics). The following protocol was used: TR/TE, 4.2/2.2 ms; 15° flip angle; FOV, 380 × 340 mm; matrix size, 320 × 224; section thickness, 4 mm; time resolution, 7.0 s; and total time, 4 min, 40 s.

Imaging Analysis and Parameter Acquisition

DCE-MRI data were processed by Omni-Kinetics (O.K.; GE Healthcare, China) software. Multi-flip angle T1 mapping transformed the signal intensity into contrast agent

TABLE 1 | Comparison of DCE-MRI and texture feature derived parameters.

Parameters	Invasion group (*n* = 27)	Non-invasion group (*n* = 35)	*P*-value	ICC	
				Inter (95% CI)	Intra (95% CI)
DCE-MRI					
K^{trans}, min^{-1}	0.335 ± 0.050	0.269 ± 0.079	<0.001	0.917 (0.834, 0.953)	0.841 (0.792, 0.878)
K_{ep}, min^{-1}	0.538 ± 0.103	0.526 ± 0.110	0.652	0.934 (0.869, 0.958)	0.893 (0.847, 0.961)
V_e	0.538 ± 0.095	0.511 ± 0.104	0.270	0.842 (0.793, 0.902)	0.927 (0.842, 0.972)
Texture					
Mean	679.53 ± 66.02	697.12 ± 59.70	0.260	0.924 (0.835, 0.959)	0.858 (0.786, 0.919)
Skewness	0.043 ± 0.242	0.085 ± 0.267	0.502	0.925 (0.834, 0.946)	0.915 (0.862, 0.968)
Kurtosis	3.252 ± 0.477	3.263 ± 0.653	0.940	0.826 (0.783, 0.879)	0.935 (0.883, 0.969)
Uniformity	0.912 ± 0.013	0.915 ± 0.034	0.719	0.904 (0.837, 0.961)	0.912 (0.852, 0.969)
Energy	0.503 ± 0.093	0.602 ± 0.087	<0.001	0.879 (0.802, 0.948)	0.892 (0.817, 0.948)
Entropy	1.391 ± 0.193	1.24 ± 0.129	<0.001	0.924 (0.846, 0.958)	0.921 (0.836, 0.971)

DCE-MRI, dynamic contrast-enhanced magnetic resonance imaging; K^{trans}, volume transfer constant; K_{ep}, flux rate constant between extravascular extracellular space and blood plasma; V_e, fraction of extravascular extracellular volume.

concentration. For the evaluation of the arterial input function (AIF), a region of interest (ROI) was manually placed on the iliac artery. Tumors were outlined on each slice from DCE-MRI images in order to show the volume of interest (VOI) of the whole tumor. The DCE-MRI parameters (K^{trans} [the volume transfer constant], K_{ep} [the flux rate constant], and V_e [fractional extravascular extracellular space volume]), were calculated using a modified Tofts model.

Texture features were acquired by axial T2-weighted images with ITK-SNAP software (version 3.6.0) and O.K. software. For each patient, the VOI was calculated using ITK-SNAP software by manually delineating the ROI along the edge of the tumor on each slice for the entire tumor by referencing the corresponding contrast-enhanced images. Texture features were extracted from the delineated VOI by the O.K. software.

DCE-MRI and texture features were independently measured by two radiologists using the O.K. software (XX.L. and TT.L. with 6 years and 12 years of clinical experience, respectively, in gynecologic oncology MR imaging). Reader 1 measured DCE-MRI and texture features twice in 1 week to estimate intraobserver reproducibility, and his first measurement was compared with the measurement obtained by reader 2 to assess interobserver agreement. The mean of the two measured data of reader 1 was statistically analyzed. The intraclass correlation coefficient (ICC) of more than 0.75 edindicated good agreement.

Statistical Analysis

Quantitative metrics were indicated as mean ± standard error, and the normality test was assessed using the Kolmogorov-Smirnov method. The normal distribution of the DCE-MRI and texture feature parameters data was compared between the invasion group and the non-invasion group was using one-way analysis of variance; comparisons among the metrics of each group were processed using least significant difference (LSD) test. Non-normal distribution of data were compared using the Mann–Whitney test. Receiver operating characteristic (ROC) curve analysis was used to assess the diagnostic ability of DCE-MRI and texture feature parameters in diagnosing parametrial invasion. The AUC was compared using the Delong Clarke-Pearson method (DeLong et al., 1988). Cut-off values were obtained by maximizing Youden's index (sensitivity+specificity-1). Statistical analysis was performed using SPSS 23.0 (IBM Corp., Armonk, NY) and GraphPad Prism 8.0 (GraphPad software, San Diego, CA). P<0.05 was considered as the threshold for statistical significance.

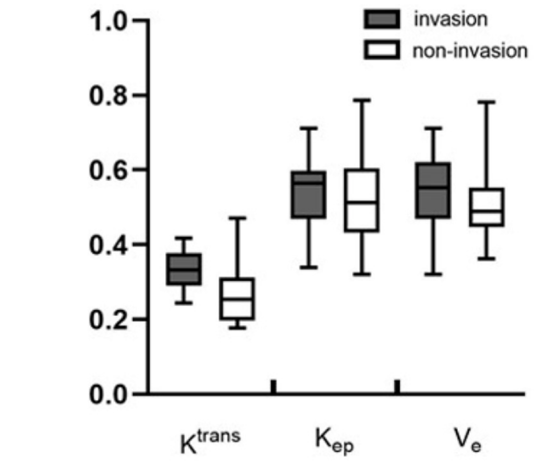


FIGURE 1 | Boxplots of the DCE-MRI parameters of K^{trans}, K_{ep}, and V_e between the invasion and non-invasion groups.

RESULTS

Excellent intra- and interobserver agreements were found in the measurements of DCE-MRI and texture features metrics (**Table 1**). The intra- and interobserver ICCs of K^{trans}, K_{ep}, and

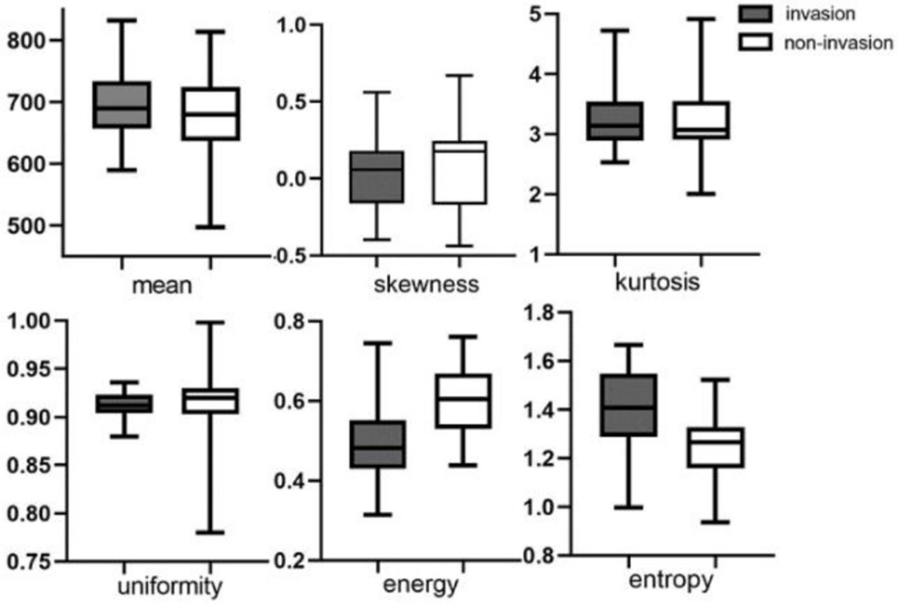


FIGURE 2 | Boxplots of texture features of mean value, skewness, kurtosis, uniformity, energy, and entropy between invasion and non-invasion groups.

V_e were 0.917 and 0.841, 0.934 and 0.893, and 0.842 and 0.927, respectively. Meanwhile, the intra- and interobserver ICCs of the mean value were 0.924 and 0.858, of skewness were 0.925 and 0.915, of kurtosis were 0.826 and 0.935, of uniformity were 0.904 and 0.912, of energy were 0.879 and 0.892, and of entropy were 0.924 and 0.921.

Metrics derived from DCE-MRI and texture features were compared between the invasion group and the non-invasion group and are summarized in **Table 1** and **Figures 1**, **2**. The invasion group indicated a significantly higher K^{trans} (0.335 $\pm$ 0.050 vs. 0.269 $\pm$ 0.079; $p < 0.001$), lower energy values (0.503 $\pm$ 0.093 vs. 0.602 $\pm$ 0.087; $p < 0.001$), and higher entropy values (1.391 $\pm$ 0.193 vs. 1.24 $\pm$ 0.129; $p < 0.001$) than the non-invasion group, while there was no significant difference for K_{ep}, V_e, mean, skewness, kurtosis, or uniformity.

ROC analysis showed that setting the K^{trans} cut-off value to ≥ 0.286 min^{-1} produced the best diagnostic performance for diagnosing parametrial infiltration (AUC, 0.788; sensitivity, 0.839; specificity, 0,657). The best diagnostic ability could be obtained by setting the threshold value of energy at ≤ 0.488 (AUC, 0.761; sensitivity, 0.710; specificity, 0.714). Setting the critical value of entropy at ≥ 1.387 obtained the best diagnostic index (AUC, 0.749; sensitivity, 0.581; specificity, 0.943). The combination of K^{trans} and energy had a significantly better diagnostic index than an independent diagnosis of K^{trans}, energy, and entropy ($p = 0.036$, $p = 0.029$, $p = 0.047$). However, using a combination of K^{trans} and entropy (AUC, 0.727; sensitivity, 0.806; specificity, 0.657) and a combination of energy and entropy (AUC, 0.619; sensitivity, 0.548; specificity, 0.771) as the diagnostic marker achieved a significantly worse performance than other single and combination parameters. The combination of K^{trans}, energy, and entropy (AUC, 0.925; sensitivity, 0.935; specificity, 0.829) resulted in a significantly better diagnostic performance than K^{trans}, energy, entropy, a combination of K^{trans} and energy, a combination of K^{trans} and entropy, or a combination of K^{trans} and energy ($p = 0.042$, $p = 0.037$,

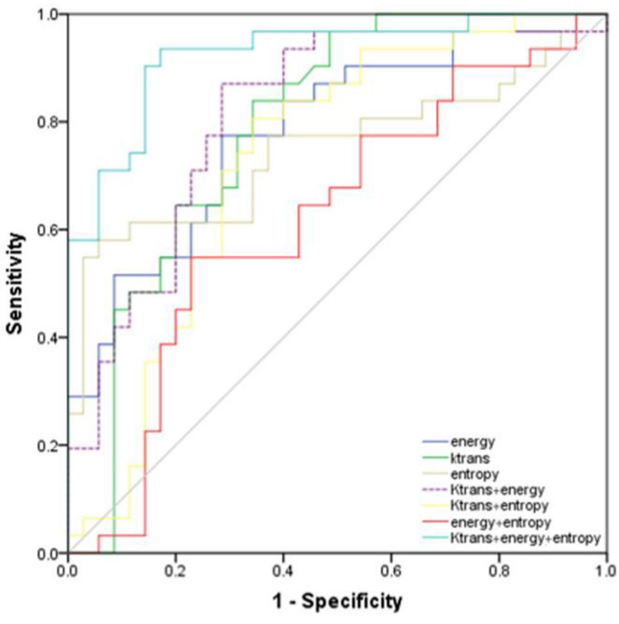


FIGURE 3 | Receiver operating characteristic curves of the energy, K^{trans}, entropy, combination of K^{trans} and energy, combination of K^{trans} and entropy, combination of energy and entropy, combination of K^{trans}, energy, and entropy for diagnosis of invasion and non-invasion of cervical cancer.

TABLE 2 | Diagnostic efficiency of each parameter and their combined metrics.

Parameters	Cut-off value	AUC	Sensitivity	Specificity
K^{trans}, min^{-1}	0.286	0.788 (0.690–0.849)	0.839	0.657
Energy	0.488	0.785 (0.717–0.857)	0.774	0.714
Entropy	1.387	0.749 (0.657–0.828)	0.581	0.943
K^{trans} + energy		0.813 (0.748–0.859)	0.871	0.714
K^{trans} + entropy		0.728 (0.604–0.830)	0.806	0.657
Energy + entropy		0.619 (0.481–0.757)	0.548	0.771
K^{trans} + energy + entropy		0.925 (0.853–0.976)	0.935	0.829

Data in parentheses indicate 95% confidence intervals. AUC, area under curve; K^{trans}, volume transfer constant.

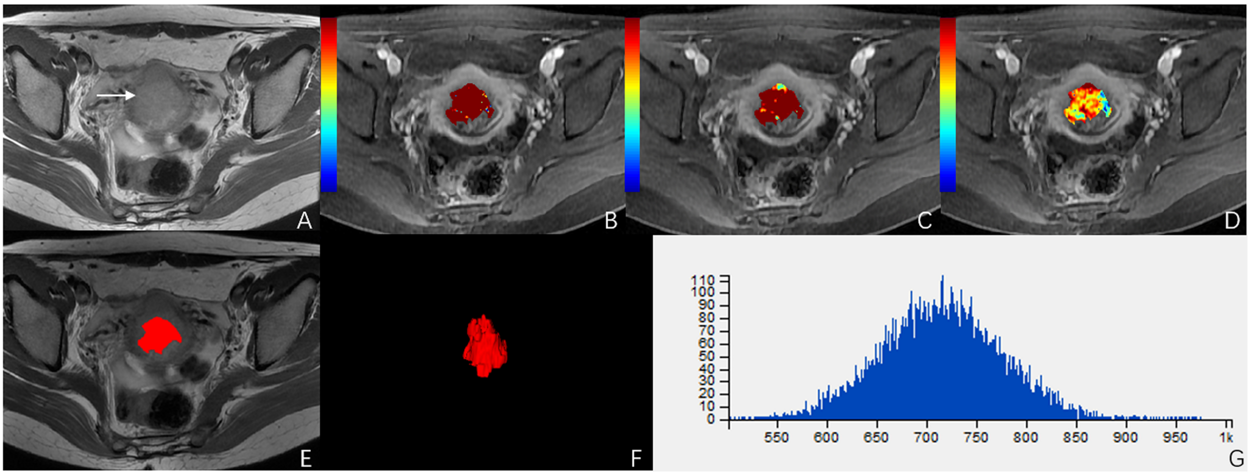


FIGURE 4 | A 51-year-old woman with stage IIB cervical cancer. **(A)** T2-weighted image (T2WI) shows a slightly hyperintense cervical mass. **(B–D)** K^{trans}, K_{ep}, and V_e parametric maps are derived from DCE-MRI. The corresponding values are 0.577/min, 0.639/min, 0.694. **(E,F)** The volume of the tumor is drawn from T2WI. **(G)** Histogram map of the entire tumor.

$p = 0.018$, $p = 0.048$, $p = 0.007$, $p = 0.029$, respectively) (**Figure 3**). **Table 2** shows the detailed diagnostic performances. The representative images of DCE-MRI and texture features of cervical cancer with and without invasion are summarized in **Figures 4**, **5**.

DISCUSSION

The signal intensity kinetics acquired by DCE-MRI suggest the underlying microvessel density, perfusion, permeability, and the extracellular-extravascular space composition of tumors (Zahra et al., 2007; Bonekamp et al., 2016). DCE-MRI can predict the response to and outcomes of radiotherapy in patients with cervical cancer (Tao et al., 2019). Tao et al. reported (2019) that the K^{trans} of high-grade ductal carcinoma *in situ* of the breast is higher than that of low-grade ductal carcinoma. Li et al. (2015) reported that the K^{trans} of high-grade glioma is higher than that of low-grade glioma. stoLikewise, the present study infound that the K^{trans} value of the invasive group was higher than that of the non-invasive group. The parameter K^{trans} reflected tumor angiogenesis, which is proportional to the density of the tumor vessels. This indicated that the angiogenesis of the invasive cervical cancer group was greater than that of the non-invasive group, and the malignant degree of the invasive group was higher than that of the non-invasive group. The more malignant the tumors are, the more angiogenesis they have. Early cervical cancer consists mainly of neovascularization, but the blood vessels are few in number and have low permeability. The growth rate of advanced cervical cancer is faster than that of early cervical cancer, and the tumor's demand for blood oxygen is increased, so that a large number of new blood vessels are formed. The angiogenesis and the permeability of the blood vessels is increased. The tortuous course of the blood vessels increases their permeable area. Moreover, the endothelial cells of blood vessels are irregular. Therefore, the contrast agent permeates through the gaps in the blood vessels more easily than it does in early cervical cancer. We can conclude that the parameters of K^{trans} in the infiltration group were higher than those in the non-infiltration group.

Texture analysis is a new image post-processing computer technology that quantitatively analyzes the distribution rules and characteristics of image pixels and reflects lesion heterogeneity

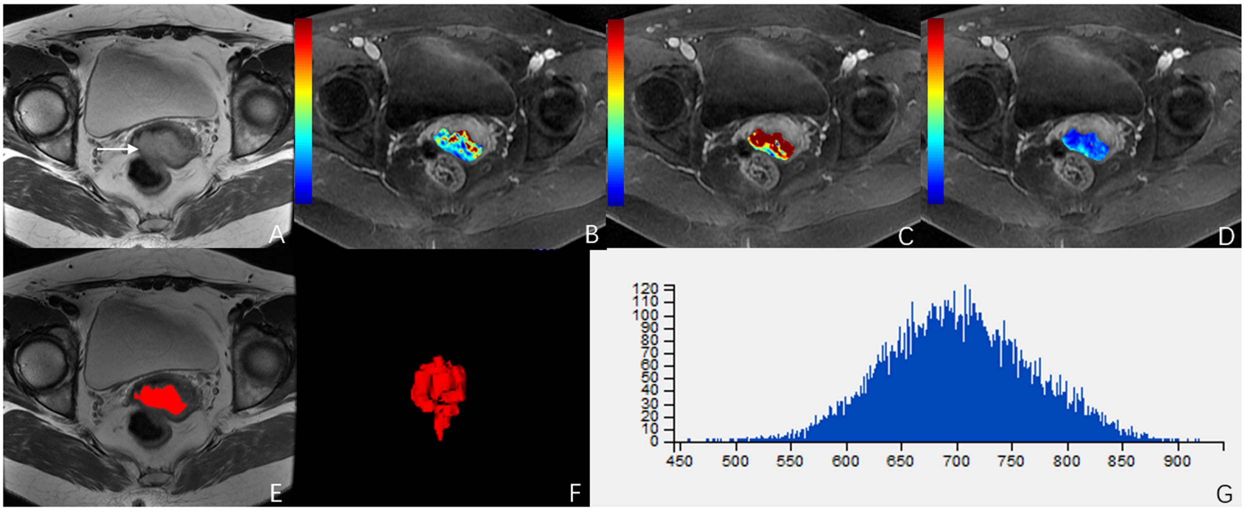


FIGURE 5 | A 49-year-old woman with stage IIA cervical cancer. **(A)** T2-weighted image (T2WI) shows a slightly hyperintense cervical mass. **(B–D)** K^{trans}, K_{ep}, and V_e parametric maps are derived from DCE-MRI. The corresponding values are 0.196/min, 0.430/min, 0.396. **(E,F)** The volume of the tumor is drawn from T2WI. **(G)** Histogram map of the entire tumor.

and the fine differences of tumors. Energy reflects uniformity and texture the coarseness of the images. The better distributed the gray of the image is, the greater its value. In this study, the energy value of the non-invasive group was larger than that of the invasive group, which indicated that the images of the invasive group were less uniform than those of the non-invasive group. This may be owing to the cystoid degeneration and necrosis in the invasive cervical cancer group. Entropy, reflecting the basic degree of chaos in the gray levels, is a measure of the image information. It is mainly used to evaluate the uniformity of image texture. The entropy value of the parametrial infiltration group was higher than that of the non-infiltration group, which indicated that the distribution of image pixels in the infiltration group was more discrete and disordered than that in the non-infiltration group. The reason for this difference may be related to the degree of malignancy. Several studies showed that tumors with a high degree of malignancy have high heterogeneity (Ng et al., 2013; Zhang et al., 2017a,b), and a high entropy value represents high tumor heterogeneity (Guan et al., 2017), and this is in accordance with our study showing that advanced cervical cancer representing high heterogeneity has high entropy. Guan et al. (2017) showed that cervical cancers with higher (IIB-IVA) rather than lower (IB-IIA) FIGO stages had lower energy and higher entropy of texture features based on ADC images. This result is consistent with our study showing that advanced cervical cancer has lower energy and higher entropy than early stage cervical cancer. The mean value, skewness, kurtosis, and homogeneity had no statistical significance in diagnosing cervical cancer with parametrial infiltration. These parameters may have significance if we performed multiple sequences of MRI for texture analysis.

In the present study, K^{trans}, energy, entropy, and combinations of them had the optimal diagnostic performance for diagnosing cervical cancer with parametrial infiltration; particularly, the combination of K^{trans}, energy and entropy had the highest AUC (0.925) and sensitivity (93.5%). This indicated that the combination of K^{trans}, energy, and entropy was more significant than the other parameters for the diagnosis of parametrial infiltration. In other words, DCE-MRI, representing quantitative perfusion information at the molecular level, and texture features, representing a mathematical model of the gray distribution of quantitative image pixels, are the most valuable for the diagnosis of cervical cancer with parametrial infiltration. Thus, we can use more accurate quantitative parameters at the microscopic level instead of making a subjective diagnosis with a larger margin of error to evaluate parametrial infiltration. Quantitative parameters can be used as an important ancillary diagnostic tool for routine MR examination and can provide a reference for the establishment of an artificial intelligence prediction model of cervical cancer with parametrial infiltration.

Several limitations of the present study are as follows. First, the sample size of this study was relatively small. Second, the ROIs of the tumors were manually performed, which might increase the variability of the data measurement. Third, our study did not distinguish the pathological types of cervical cancer, such as squamous cell carcinoma, adenocarcinoma, and small cell carcinoma.

In conclusion, this study showed that the invasion group of cervical cancer demonstrated significantly higher K^{trans}, lower energy values, and higher entropy values than those in the non-invasion group. Both DCE-MRI and texture analysis were valuable in the diagnosis. A combination of DCE-MRI and texture analysis may be a promising method to improve accuracy in diagnosing cervical cancer with parametrial infiltration prior to treatment and has great significance in the medical field.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical research ethics board of the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

WW, TL, and BL contributed conception and design of the study. XL organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. WW, XL, TL, and BL wrote sections of the manuscript. All authors contributed to manuscript revision, read and approve the submitted version.

FUNDING

This work was supported by the National Natural Science Youth Science Fund Project (No. 81501468), by the Fundamental Research Funds for the Central Universities (No. WK9110000002).

REFERENCES

Bonekamp, D., Wolf, M. B., Edler, C., Katayama, S., Schlemmer, H. P., Herfarth, K., et al. (2016). Dynamic contrast enhanced MRI monitoring of primary proton and carbon ion irradiation of prostate cancer using a novel hypofractionated raster scan technique. *Radiother. Oncol.* 120, 313–319. doi: 10.1016/j.radonc.2016.05.012

Chiappa, V., Di Legge, A., Valentini, A. L., Gui, B., Micco, M., Ludovisi, M., et al. (2015). Agreement of two-dimensional and three-dimensional transvaginal ultrasound with magnetic resonance imaging in assessment of parametrial infiltration in cervical cancer. *Ultrasound Obstet. Gynecol.* 45, 459–469. doi: 10.1002/uog.14637

Chung, H. H., Nam, B. H., Kim, J. W., Kang, K. W., Park, N. H., Song, Y. S., et al. (2010). Preoperative [18F]FDG PET/CT maximum standardized uptake value predicts recurrence of uterine cervical cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 37, 1467–1473. doi: 10.1007/s00259-010-1413-5

Dai, Y. F., Xu, M., Zhong, L. Y., Xie, X. Y., Liu, Z. D., Yan, M. X., et al. (2018). Prognostic significance of solitary lymph node metastasis in patients with stages IA2 to IIA cervical carcinoma. *J. Int. Med. Res.* 46, 4082–4091. doi: 10.1177/0300060518785827

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845.

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., et al. (2020). Developing Computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms21072274

Freeman, S. J., Aly, A. M., Kataoka, M. Y., Addley, H. C., Reinhold, C., and Sala, E. (2012). The revised FIGO staging system for uterine malignancies: implications for MR imaging. *Radiographics* 32, 1805–1827. doi: 10.1148/rg.326125519

Guan, Y., Li, W., Jiang, Z., Zhang, B., Chen, Y., Huang, X., et al. (2017). Value of whole-lesion apparent diffusion coefficient (ADC) first-order statistics and texture features in clinical staging of cervical cancers. *Clin. Radiol.* 72, 951–958. doi: 10.1016/j.crad.2017.06.115

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for QSAR prediction. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009. [Epub ahead of print].

Kong, T.-W., Chang, S.-J., Piao, X., Paek, J., Lee, Y., Lee, E. J., et al. (2016). Patterns of recurrence and survival after abdominal versus laparoscopic/robotic radical hysterectomy in patients with early cervical cancer. *J. Obstetr. Gynaecol. Res.* 42, 77–86. doi: 10.1111/jog.12840

Li, X., Zhu, Y., Kang, H., Zhang, Y., Liang, H., Wang, S., et al. (2015). Glioma grading by microvascular permeability parameters derived from dynamic contrast-enhanced MRI and intratumoral susceptibility signal on susceptibility weighted imaging. *Cancer Imag.* 15:4. doi: 10.1186/s40644-015-0039-z

Meng, J., Zhu, L., Zhu, L., Ge, Y., He, J., Zhou, Z., et al. (2017). Histogram analysis of apparent diffusion coefficient for monitoring early response in patients with advanced cervical cancers undergoing concurrent chemo-radiotherapy. *Acta Radiol.* 58, 1400–1408. doi: 10.1177/0284185117694509

Mu, W., Chen, Z., Liang, Y., Shen, W., Yang, F., Dai, R., et al. (2015). Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. *Phys. Med. Biol.* 60, 5123–5139. doi: 10.1088/0031-9155/60/13/5123

Munagala, R., Rai, S. N., Ganesharajah, S., Bala, N., and Gupta, R. C. (2010). Clinicopathological, but not socio-demographic factors affect the prognosis in cervical carcinoma. *Oncol. Rep.* 24, 511–520. doi: 10.3892/or_00000887

Ng, F., Ganeshan, B., Kozarski, R., Miles, K. A., and Goh, V. (2013). Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. *Radiology* 266, 177–184. doi: 10.1148/radiol.12120254

Noh, J. M., Park, W., Kim, Y. S., Kim, J. Y., Kim, H. J., Kim, J., et al. (2014). Comparison of clinical outcomes of adenocarcinoma and adenosquamous carcinoma in uterine cervical cancer patients receiving surgical resection followed by radiotherapy: a multicenter retrospective study (KROG 13-10). *Gynecol. Oncol.* 132, 618–623. doi: 10.1016/j.ygyno.2014.01.043

Park, J. J., Kim, C. K., Park, S. Y., Park, B. K., and Kim, B. (2014). Value of diffusion-weighted imaging in predicting parametrial invasion in stage IA2-IIA cervical cancer. *Eur. Radiol.* 24, 1081–1088. doi: 10.1007/s00330-014-3109-x

Patel-Lippmann, K., Robbins, J. B., Barroilhet, L., Anderson, B., Sadowski, E. A., and Boyum, J. (2017). MR imaging of cervical Cancer. *Magn. Reson. Imaging Clin. N. Am.* 25, 635–649. doi: 10.1016/j.mric.2017.03.007

Tao, W. J., Zhang, H. X., Zhang, L. M., Gao, F., Huang, W., Liu, Y., et al. (2019). Combined application of pharamcokinetic DCE-MRI and IVIM-DWI could improve detection efficiency in early diagnosis of ductal carcinoma *in situ*. *J. Appl. Clin. Med. Phys.* 20, 142–150. doi: 10.1002/acm2.12624

Wang, B., Wang, L., Zheng, C. H., and Xiong, Y. (2019). Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/tcbb.2019.2953908. [Epub ahead of print].

Wang, W., Zhou, Y., Cheng, M. T., Wang, Y., Zheng, C. H., Xiong, Y., et al. (2020). Potential pathogenic genes prioritization based on protein domain interaction network analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/tcbb.2020.2983894. [Epub ahead of print].

Woo, S., Kim, S. Y., Cho, J. Y., and Kim, S. H. (2018). Apparent diffusion coefficient for prediction of parametrial invasion in cervical cancer: a critical evaluation based on stratification to a Likert scale using T2-weighted imaging. *Radiol. Med.* 123, 209–216. doi: 10.1007/s11547-017-0823-x

Xia, X., Xu, H., Wang, Z., Liu, R., Hu, T., and Li, S. (2016). Analysis of prognostic factors affecting the outcome of stage IB-IIB cervical cancer treated by radical hysterectomy and pelvic lymphadenectomy. *Am. J. Clin. Oncol.* 39, 604–608. doi: 10.1097/COC.0000000000000100

Zahra, M. A., Hollingsworth, K. G., Sala, E., Lomas, D. J., and Tan, L. T. (2007). Dynamic contrast-enhanced MRI as a predictor of tumour response to radiotherapy. *Lancet Oncol.* 8, 63–74. doi: 10.1016/s1470-2045(06)71012-9

Zhang, G. M., Shi, B., Sun, H., Jin, Z. Y., and Xue, H. D. (2017a). Differentiating pheochromocytoma from lipid-poor adrenocortical adenoma by CT texture analysis: feasibility study. *Abdom. Radiol. (NY)* 42, 2305–2313. doi: 10.1007/s00261-017-1118-3

Zhang, G. M., Sun, H., Shi, B., Jin, Z. Y., and Xue, H. D. (2017b). Quantitative CT texture analysis for evaluating histologic grade of urothelial carcinoma. *Abdom. Radiol.* 42, 561–568. doi: 10.1007/s00261-016-0897-2

Zhou, G., Chen, X., Tang, F., Zhou, J., Wang, Y., and Wang, Z. (2016). The value of diffusion-weighted imaging in predicting the prognosis of stage IB-IIA cervical squamous cell carcinoma after radical hysterectomy. *Int. J. Gynecol. Cancer* 26, 361–366. doi: 10.1097/IGC.0000000000000613

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 26 June 2020
doi: 10.3389/fbioe.2020.00588

Integrated Analysis of DEAD-Box Helicase 56: A Potential Oncogene in Osteosarcoma

Chen Zhu[1†], *Xianzuo Zhang*[1*†], *Nikolaos Kourkoumelis*[2], *Yong Shen*[3,4*] *and Wei Huang*[1]

[1] Division of Life Sciences and Medicine, Department of Orthopedics, The First Affiliated Hospital of USTC, University of Science and Technology of China, Hefei, China, [2] Department of Medical Physics, School of Health Sciences, University of Ioannina, Ioannina, Greece, [3] Institute on Aging and Brain Disorders, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, [4] Division of Life Sciences and Medicine, Neurodegenerative Disorder Research Center, University of Science and Technology of China, Hefei, China

Background: Osteosarcoma is a solid tumor common in the musculoskeletal system. The DEAD-box helicase (DDX) families play an important role in tumor genesis and proliferation.

Objective: To screen potential molecular targets in osteosarcoma and elucidate its relationship with DDX56.

Methods: We employed the Gene Expression Omnibus and The Cancer Genome Atlas datasets for preliminary screening. DDX56 expression was measured by RT-qPCR in three osteosarcoma cell lines. Biological roles of DDX56 were explored by Gene ontology, Kyoto Encyclopedia of Genes and Genomes and Ingenuity Pathway Analysis. Cell proliferation, cycle, and apoptosis assays were performed using Lentivirus^{TM} knockdown technique.

Results: It was found that DDX56 expression was regularly upregulated in osteosarcoma tissue and cell lines, while DDX56 knockdown inhibited cell proliferation and promoted cell apoptosis.

Conclusions: The findings suggest DDX56 as a potential therapeutic target for the treatment of osteosarcoma.

Keywords: biomarker, DEAD-box RNA helicases 56 (DDX56), osteosarcoma, oncogene, proliferation

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Chun Hoi Yan,
The University of Hong Kong,
Hong Kong
Tao Cheng,
Shanghai Jiao Tong University, China

****Correspondence:***
Xianzuo Zhang
zhangxianzuo@foxmail.com
Yong Shen
shenyong@ustc.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *01 March 2020*
Accepted: *14 May 2020*
Published: *26 June 2020*

Citation:
Zhu C, Zhang X, Kourkoumelis N, Shen Y and Huang W (2020) Integrated Analysis of DEAD-Box Helicase 56: A Potential Oncogene in Osteosarcoma. Front. Bioeng. Biotechnol. 8:588. doi: 10.3389/fbioe.2020.00588

INTRODUCTION

Osteosarcoma derives from primitive bone-forming mesenchymal cells. It is a primary bone neoplasm characterized by the production of osteoid or immature bone by the malignant cells (Anderson, 2016). Being the most common primary malignancy of bone in children and adolescents, the incidence of osteosarcoma are 4–5 per year per million for all races and both sexes (Ottaviani and Jaffe, 2009). The survival rate of patients with osteosarcoma has improved mostly due to marked advances in diagnosis and chemotherapy (Anderson, 2016). However, the high rate of relapse and distant metastasis of osteosarcoma result in poor long-term survival (Kumar et al., 2017). Thus, there is an urgent need to develop new treatment strategies for osteosarcoma.

High-throughput microarrays are promising tools for identifying candidate molecular targets in medical oncology. During the last decade, numerous gene expression profiling studies on osteosarcoma oncogenesis and proliferation were performed using microarray technology and showed hundreds of differentially expressed genes (DEGs) involved in different pathways, biological processes, or molecular functions.

The emerging roles of Asp–Glu–Ala–Asp (DEAD)-box RNA helicases have recently been acknowledged in disparate cellular functions. DEAD-box (DDX) RNA helicases play a crucial role not only in unwinding double-stranded RNA molecules but also in transcription, splicing, RNA transport, ribosome biogenesis, RNA editing, RNA decay, and translation (Sugiura et al., 2007; Xu and Hobman, 2012). Although DDX56 is reported to be required in virus infection (Reid and Hobman, 2017), affecting the response to abiotic stress and host–pathogen interaction (Pragya et al., 2007; Umate et al., 2010), the relationship of DDX RNA helicases with malignancies remains unclear. The relationship between DDX family and cancer is worth further investigating since a number of DEAD-box RNA helicases were recently reported to be implicated in solid tumor progression and chemotherapy resistance (Kuramitsu et al., 2013; He et al., 2018).

MATERIALS AND METHODS

Patients and Samples

The gene expression microarray profile of GSE126209 was downloaded from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The mRNA profiles of osteosarcoma tumors and adjacent normal tissues were generated by high-throughput sequencing based on the GPL20301 platform (Illumina HiSeq 4000, *Homo sapiens*). Eleven samples were obtained from six Chinese Uyghur patients, in which 10 samples were paired tumor/normal specimens from the same patients. Specifically, gene expression profiles of mesenchymal stem cell osteosarcoma patients were compared with those in non-neoplastic patient to identify the DEGs. The DEAD-box family was given special attention among the DEGs. qPCR was used to examine if DDX56 is highly expressed in osteosarcoma cell lines. Subsequently, we employed the Lentivirus^{TM} technique to examine the effect of DDX56 silencing on human osteosarcoma cell growth *in vitro*.

Identification of Differentially Expressed mRNAs

Paired *t*-test was used to filter differentially expressed mRNAs between tumor and adjacent normal tissues. We selected differentially expressed genes according to the *p*-value threshold and absolute value of fold change (FC). A value of $p < 0.05$ with $|FC| > 2$ was considered to represent a significant difference. The Ensembl Gene ID of the mRNAs was transferred into gene symbol using the Biomart module in Ensembl (http://www.ensembl.org/biomart/martview/).

Hierarchical Clustering

The differentially expressed profiles of mRNAs in DDX 56 family were clustered using a hierarchical cluster algorithm with average linkage and Spearman's rank correlation distance, as provided by the software EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/). The clustering was performed using the methods outlined in a previous publication (Misha et al., 2004). Results were visualized with the help of heatmaps and dendrograms.

Protein–Protein Interaction Network Analysis

The protein–protein interaction (PPI) pairs between differentially expressed mRNAs were identified using the IID (Integrated Interactions Database, version 2018-11) database (Kotlyar et al., 2016), tissue-specific protein–protein interactions (PPIs) with larger information (a total of 1,566,043 PPIs among 68,831 proteins). The PPI interaction in this study was specified in musculoskeletal tissues. Furthermore, Cytoscape (version 3.5.0) was used to establish the PPI network and calculate the parameters of nodes and edges. Top nodes in the DDX56 family net were chosen according to the network topology property indicators, and were analyzed by CytoNCA in Cytoscape for factors including degree, betweenness centrality, and closeness centrality. In general, a high indicator score in network topology denotes an important role in the network. Top nodes with the highest degree were selected for further study.

mRNA Profile Data and Survival Analysis

The mRNA profile and its corresponding survival data were retrieved from The Cancer Genome Atlas (TCGA) database (https://tcga-data.nci.nih.gov/tcga/). These data were analyzed using the UALCAN (http://ualcan.path.uab.edu/) portal tools (Chandrashekar et al., 2017). The UALCAN tools enable graphs and plots depicting gene expression and patient survival information based on gene expression. Additional information about the selected genes was provided by GTEx (https://gtexportal.org/). Genes positively and negatively correlated with DDX56 in sarcoma (SARC) patients were screened out according to GTEx Profiles (Lonsdale et al., 2013). Extremely low-expressed genes (median TPM < 0.5) were filtered out.

Osteosarcoma Cell Lines

Human osteosarcoma cell lines (HOS, Sao-2, and U-2 OS) were purchased from the Shanghai Cell Bank (Shanghai, China). Cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM; HyClone, Tauranga, New Zealand) supplemented with 10% fetal bovine serum (FBS; Gibco, Rockville, MD, USA), 100 µg/ml of streptomycin (Sigma-Aldrich, St. Louis, MO, USA), and 100 U/ml of penicillin (Sigma-Aldrich), followed by incubation in a humidified atmosphere with 5% CO_2 at room temperature.

Functional Enrichment Analysis

Gene ontology (GO) analysis, which organizes genes into hierarchical categories and uncovers gene regulatory networks on the basis of biological process and molecular function, was used to analyze the main function of differentially expressed genes (Gene Ontology, 2006). The KEGG pathway analysis was then used to identify the significant pathways for these genes (Kanehisa et al., 2004). The Database for Annotation, Visualization and Integrated Discovery (DAVID; https:/david.ncifcrf.gov/) provides a comprehensive set of functional annotation tools to analyze high-throughput gene function. GO and KEGG pathway enrichment analysis were performed using DAVID. We were only interested in biological processes, cell components, molecule functions, and KEGG pathways at the significant level ($p < 0.05$, FDR < 0.05, and an enrichment score of >1.5).

Ingenuity Pathway Analysis (IPA)

The core pathway analysis was performed with Ingenuity Pathway Analysis (Andreas et al., 2014). IPA contains a curated database of networks and biological relationships based on original peer-reviewed articles. A geneset including DDX56 and closely related genes were uploaded and analyzed separately using the IPA software (Qiagen) (https://apps.ingenuity.com/).

Western Blot Analysis

Cells were harvested in RIPA buffer. Protein concentration was measured using the BCA protein assay (HyClone-Pierce, Rockford, IL, USA). Equal amounts of total protein of each treatment were separated using 12.5% SDS-PAGE and further transferred onto PVDF membranes. Membranes were incubated with mouse anti-FLAG or anti-GAPDH antibodies (Santa Cruz Biotechnology, Santa Cruz, CA, USA). Secondary antibodies conjugated to horseradish peroxidase and ECL Western blotting reagents were used for detection.

Quantitative Real–Time Polymerase Chain Reaction

Total RNA was extracted using the Trizol reagent (Invitrogen, Shanghai, China) and reverse transcribed to single-stranded cDNA. The cDNA was then used as a template for the following polymerase chain reaction (PCR). The primers used were as follows: for DDX56 forward, 5′-CCG CTT ATG CTA TTC CGA TGC-3′ and reverse, 5′-TGC GAG ATG GGG TCC CTA CTA TAG-3′; and for GAPDH forward, 5′-TGA CTT CAA CAG CGA CAC CCA-3′ and 5′-CAC CCT GTT GCT GTA GCC AAA-3′. GAPDH was used as an internal control. The PCR products of DDX56 and GAPDH were 258 and 121 bp, respectively. All samples were examined in triplicates.

Recombinant Lentiviral Vector Production and Cell Infection

The interfering target sequence of DDX56 (ACTCAAGGAGCTGATATTA) was designed from the full-length DDX56 sequence (NM_019082) by GeneChem Co. Ltd. (Shanghai, China). After testing knockdown efficiencies, the stem-loop oligonucleotides were synthesized and inserted into the lentivirus-based pGCSIL-GFP (GeneChem Co. Ltd.) with AgeI/EcoRI sites. For lentivirus infection, U-2 OS cells were cultured into six-well plates, and then, the DDX56–shRNA–lentivirus or negative control (NC) lentivirus was added according to a multiplicity of infection (MOI). After 72 h of infection, the cells were observed under a fluorescence microscope (MicroPublisher 3.3RTV; Olympus, Tokyo, Japan). After 120 h of infection, the cells were harvested to determine knockdown efficiency by quantitative RT-PCR.

Cell Growth Assay

Cell growth was measured using multiparametric high-content screening (HCS). Briefly, U-2 OS cells at the logarithmic phase after being infected with either the NC lentivirus or DDX56–shRNA lentivirus were seeded at 2,000 cells/well into 96-well plates; the cells were then incubated at 37°C with 5% CO_2 for 5 days. The cells in the plates were counted using the Celigo® Image HCS Cytometer (Nexcelom Bioscience LLC, Lawrence, MA, USA) for each day's analysis. In each well, at least 800 cells were analyzed. Each experiment was performed in triplicate.

In vitro Proliferation Assay

MTT assays were performed to measure the rate of cell proliferation *in vitro*. Briefly, the cells transfected with shDDX56 or shCtrl were planted into 96-well plates at a density of 1×10^5 cells/well and then cultured for 24, 48, or 72 h, respectively. The transfected cells were incubated with 25 μl of MTT (Sigma-Aldrich) for 4 h at 37°C, followed by removing of supernatants and adding of 150 μl of DMSO (Sigma-Aldrich). The absorbance value was measured at 450 nm with a microplate reader (BioTek Instruments, Winooski, VT, USA).

Cell Apoptosis Analysis

Flow cytometry (FCM) analysis was used to determine the cell cycle distribution or detect apoptosis. Briefly, U-2 OS cells were infected with DDX56–shRNA or NC plasmids and incubated at 37°C for 1, 2, 3, 4, or 5 days. At the indicated time point, adherent cells were collected. The suspension was filtered through a 300 mesh, and the DNA content of the stained nuclei was analyzed for the cell cycle phase by BD FACS Calibur flow cytometer (BD Biosciences, San Diego, CA, USA). Each experiment was performed in triplicate. Cell apoptosis was assayed by staining with Annexin V-APC (eBioscience, San Diego, CA, USA) and detected by FCM. For the analysis of apoptosis, U-2 OS cells were cultured into six-well plates. After 48 h of transfection with DDX56–shRNA or NC plasmids, the cells were collected and washed twice with ice-cold PBS. The cell concentrations were adjusted to 1 × 106/ml with 1× staining buffer. One hundred microliters of cell suspension was stained with 5 μl of Annexin V-APC at room temperature in the dark for 15 min. Cells were analyzed using FCM within 1 h. All experiments were performed in triplicate.

Statistical Analysis

Statistical analysis was performed using SPSS for Windows version 23.0 (SPSS, Inc., Chicago, IL, USA). The Student's t-test was used for raw data analysis. The random variance model t-test was performed using BRB-ArrayTools (v4.6, http://linus.nci.nih.gov/BRB-ArrayTools.html) (Wright and Simon, 2003). The statistical data for each group were presented as the mean ± SD. Because the sample size was limited, the adjusted p-values were too large after multiple testing control. We used raw value of $p < 0.05$ as threshold for nominally significant differential expression. Notably, multiple testing adjustment with FDR < 0.05 was used to filtrate enriched GO and KEGG pathways.

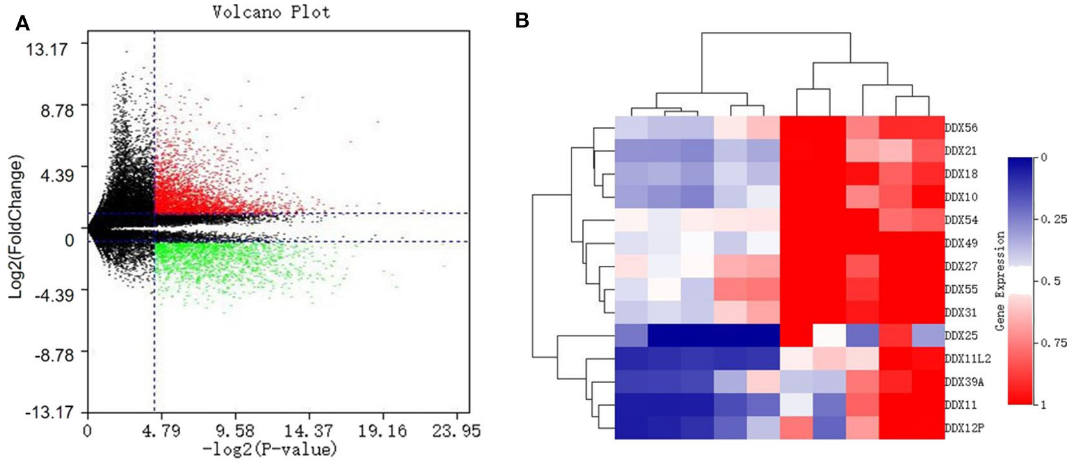


FIGURE 1 | Identification of differentially expressed genes (DEGS) in the DEAD-box helicase (DDX) family between osteosarcoma tissue and adjacent normal tissue. **(A)** Volcano plot of all 60,492 expression genes included in the GSE126209 dataset. Red spots: up-regulated DEGs; green spots: down-regulated DEGs. **(B)** The heatmap with hierarchical clustering for differentially expressed genes in DDX family.

RESULTS

Identification and Preliminary Screening of DEGs

With a fold change (FC) cut-off value >2 and a value of $p < 0.05$, a total of 4,939 mRNAs (3,301 up-regulated and 1,637 down-regulated) were identified as differentially expressed between tumor and adjacent normal tissue from the gene expression microarray profile (**Figure 1A**). Genes (4,424) succeeded in the Ensembl Gene ID—gene symbol transferring and 515 genes failed. The 4,424 genes were then uploaded to the IID website for bone tissue-specific protein–protein interactions. A total of 32,292 PPIs were identified among those genes. We selected genes in DEAD-box (DDX) RNA helicases family for further screening (**Table 1** and **Figure 1B**). The top rank gene DDX56 with highest node degree was chosen as the potential focal target.

mRNA Profile Data and Survival Analysis

The TCGA data retrieved from UALCAN portal were used to analyze the gene DDX56 expression in SARC patients. These data revealed that when compared with normal tissues, DDX56 exhibited a significant higher expression level in tumor ($p < 0.05$) (**Figure 2A**). This difference is independent of gender and race (**Figures S2A,B**). These findings were consistent with the previous DDX56 expression analysis in GSE126209 dataset and *in vitro* validation in three different osteosarcoma cell lines. The gene expression of DDX56 is upregulated in HOS, Saos-2, and U-2 OS cell (**Figure 3**). Survival analysis was also performed to evaluate whether DDX56 expression levels could predict overall prognosis. However, using all of the TCGA data obtained, the Kaplan–Meier plot demonstrated no significant differences ($p = 0.81$) (**Figure 2B**). The patients with higher DDX56 expression had a low survival rate before 3 years of onset, while it was the opposite after 3 years. The crosspoint of two plots is near 1,926 days. Stratification was made, but no demographic bias was found (**Figures S2C,D**).

TABLE 1 | Protein–protein interaction network statistics of differentially expressed genes in DEAD-box helicase (DDX) family between osteosarcoma tissue and adjacent normal tissue.

Rank	Gene symbol	Degree	Betweenness centrality	Closeness centrality
1	DDX56	206	2.24E−04	0.371734
2	DDX54	193	3.29E−04	0.385697
3	DDX55	167	3.69E−04	0.363594
4	DDX5	141	2.07E−04	0.412069
5	DDX31	141	1.11E−03	0.362973
6	DDX6	140	4.36E−04	0.392576
7	DDX27	131	1.11E−04	0.373173
8	DDX11	124	1.70E−04	0.3568
9	DDX10	115	1.56E−04	0.369817
10	DDX3X	87	9.14E−05	0.402981
11	DDX1	80	5.91E−05	0.398811
12	DDX17	79	7.82E−05	0.398782
13	DDX20	69	3.69E−05	0.369504
14	DDX12P	66	4.81E−06	0.338923
15	DDX49	64	2.62E−05	0.35585

Functional Enrichment Analysis

The TCGA data were also used to predict potential genes relevant to DDX56 function through the UALCAN portal. This prediction uses additional gene information base on GTEx profiles. This online *in silico* analysis yields a total of 480 potential genes relevant to DDX56 function. We found 204 genes co-expressing and interacting with DDX56 by intersecting these genes with the above DEGs (**Figure 4A**). More specifically, when the relationship with DDX56 was divided into positive and negative and the differential expression was divided into upregulated and downregulated, there were 199 genes in this

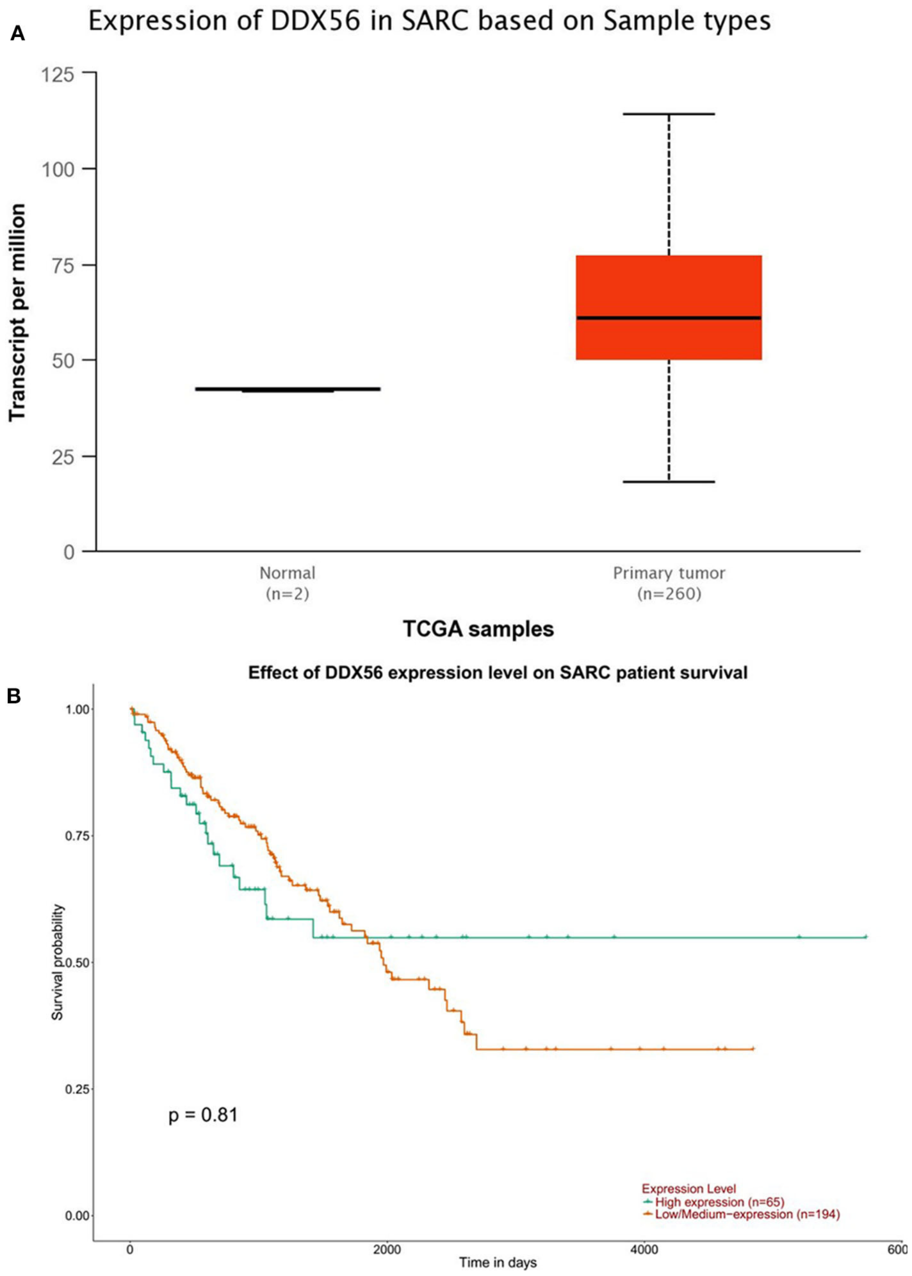


FIGURE 2 | The Cancer Genome Atlas (TCGA) gene expression and survival plot of DDX56 in sarcoma patients. **(A)** The expression was significantly higher in sarcoma patients than in negative control (NC; $p < 0.05$). **(B)** Survival plot of DDX56 expression level and race on sarcoma (SARC) patient survival ($p = 0.81$).

intersection (**Figure 4B**). The GO analysis and KEGG analysis were performed using DAVID. The top enriched biological processes were ribonucleoprotein complex biogenesis, ncRNA metabolic process and RNA processing. The top enriched cell components were nucleolus, nucleoplasm, and intracellular ribonucleoprotein complex. The top enriched molecule functions were poly(A) RNA binding, protein binding, and ATP-dependent RNA helicase activity, respectively (**Figure 4C**). The most enriched pathways include spliceosome, ribosome biogenesis in eukaryotes, and homologous recombination (**Figure 4D**).

Ingenuity Pathway Analysis

The core pathway analysis was performed using the IPA software. The role of DDX56 in cell function includes pluripotency, replication, and growth. It has several mutations found in liver neoplasm, melanoma, and pancreatic ductal adenocarcinoma. Physical interactions, including RNA–RNA, protein–protein, protein–nucleic acid, and protein–cell or tissue were also found (**Supplementary Material**). A physical interaction network was built based on these findings (**Figure 5**). This manually curated database returned a network consisting of three diseases, eight transcription regulators, one growth factor, two cytokines, one ion channel, four enzymes, and several other integrities. The tp53 gene is a joint node in this IPA network, which has direct or indirect contacts with many other nodes, and participates in the communication and regulation of the entire access network. The nodes referencing relevant diseases in this network include proliferations, apoptosis, and cell division process of tumor cells.

Lentivirus-Mediated Knockdown of DDX56

In order to validate the gene function of DDX56 in osteosarcoma, knockdown of the expression of DDX56 was performed by introducing a lentivirus cell infection model specifically designed. The knockdown efficiency was determined by external Western blot analysis using human embryonic kidney 293T cells. As is shown in **Figure 6A**, the target protein expression was detected by Western blotting in the cells, but was greatly reduced in the DDX56–shRNA-infected cultures, indicating effective knockdown of the target sequence. To further explore the role of DDX56, we knocked down DDX56 in the U-2 OS cell lines. As shown in **Figure 6B**, the proportion of infected cells was >80% for both the DDX56–shRNA and NC lentivirus by day 3 post-infection. DDX56 mRNA levels were assessed by real-time PCR at day 5 post-infection with either the DDX56–shRNA or NC lentivirus. The DDX56–shRNA lentivirus-infected cultures had significantly lower levels of DDX56 mRNA compared to levels in the cultures infected with the NC lentivirus (**Figure 6C**), which indicates the success of DDX56 knockdown in targeted cells.

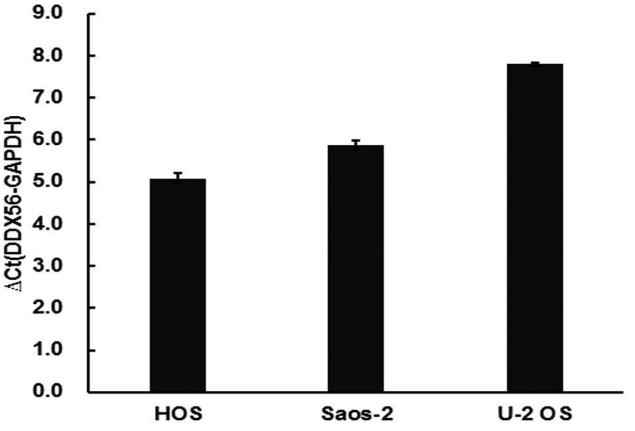


FIGURE 3 | Validation of DDX56 expression levels in three osteosarcoma cell lines. Expression of DDX56 mRNA was measured by real-time qPCR in the indicated cell lines. A constitutively expressed GAPDH gene was used as an internal control (Ct, cycle threshold; ΔCt = Ct target gene – Ct internal control).

Knockdown of DDX56 in U-2 OS Cells Reduces Cell Proliferation

To examine the effect of DDX56 on cell growth, U-2 OS cells cocultured with DDX56–shRNA or NC lentivirus were seeded into 96-well plates and were monitored by high-content screening (HCS) every day for 5 days. As illustrated in **Figure 7A** and confirmed by quantification in **Figure 7B**, control-transfected cells greatly expanded over the 5 days of the experiment, while the number of DDX56–shRNA-transfected cells did not change. The cell growth rate was defined as: Cell count at n days/cell count at first day, where $n = 2, 3, 4,$ and 5 (**Figures 7B,C**). The results of the present study showed that DDX56 knockdown significantly inhibited cell growth rate of the U-2 OS cells.

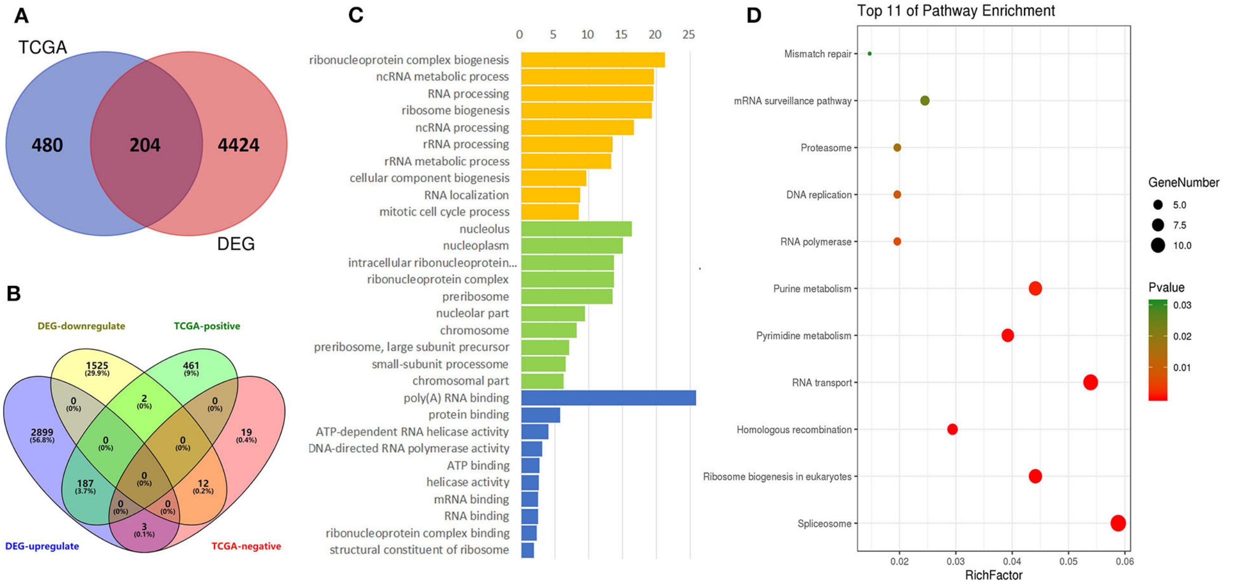


FIGURE 4 | Identification of DEGs, gene ontologies (GOs), and KEGG pathways relevant to DDX56. **(A)** Venn diagram for selected identical DEGs and TCGA-relevant genes. **(B)** Four-dimensional Venn diagram for up/down-regulated DEGs and positive/negative-related TCGA genes. **(C)** Functional GO enrichment of DDX56-relevant DEGs. **(D)** KEGG pathway enrichment.

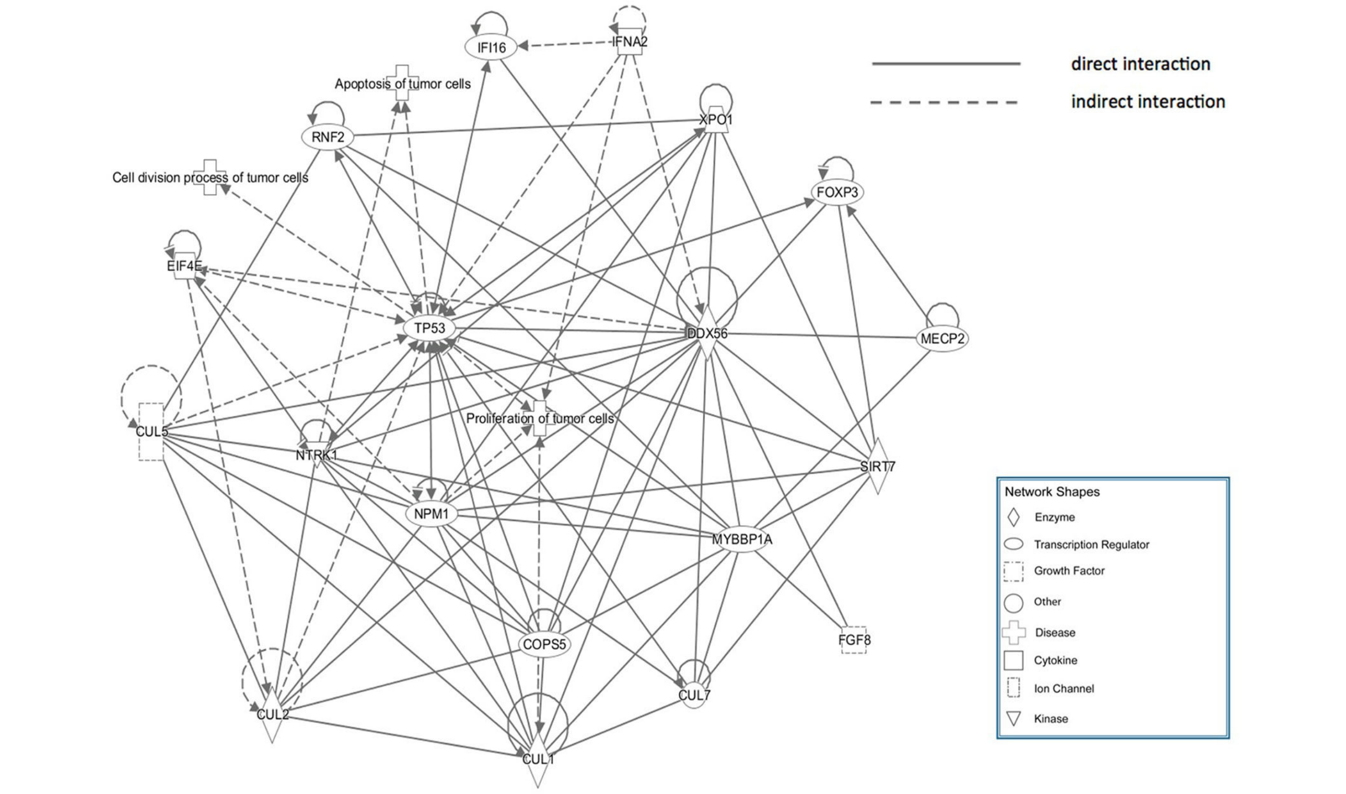


FIGURE 5 | Ingenuity Pathway Analysis (IPA) network of genes that directly interact with DDX56, which were enriched in cell proliferation, apoptosis, and cell cycling.

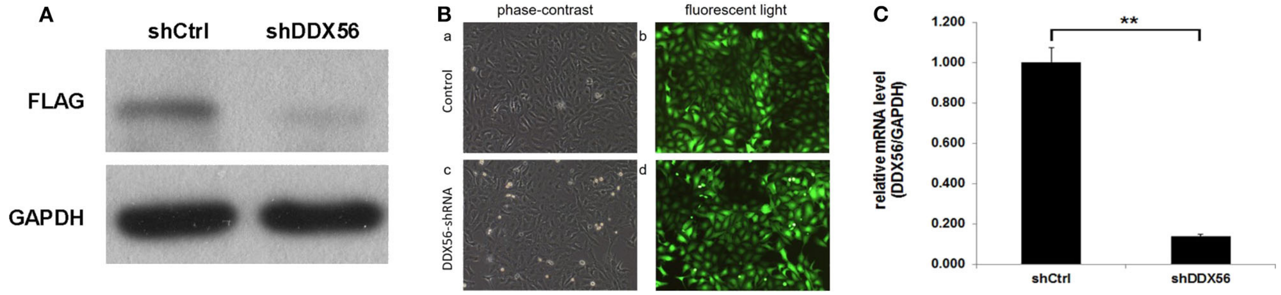


FIGURE 6 | Knockdown of DDX56 protein expression in 293T cells. **(A)** External validation of DDX56 knockdown efficiency in 293T cells. DDX56 protein expression was analyzed by Western blotting in control-transfected (NC) and DDX56–shRNA-transfected 293T cells. GAPDH was used as a loading control. **(B)** Fluorescent microscopic images of U-2 OS cell lines infected with DDX56–shRNA and NC lentivirus vectors. Note that most of the cells express GFP. Magnification, ×100. **(C)** DDX56 mRNA expression was analyzed by real-time qPCR. Compared with shCtrl, DDX56 mRNA expression was markedly decreased after silencing by RNAi (shCtrl, sham shRNA interfered control cells; shDDX56, DDX56 targeted shRNA interfered U-2 OS cells; NC, normal control; $^{**}p < 0.01$).

Knockdown of DDX56 in U-2 OS Cells Inhibit Clone Formation

In order to validate the cell clonogenic capacity change after DDX56 knockdown, MTT and clonogenic assays were used. Three days after shRNA lentivirus infection, the cells were plated in six-well plates; the number of plated cells was 400, and the number of clones was observed after 9 days. As is shown in **Figure 7D**, the results showed that the number of colonies in the experimental group decreased, suggesting that the DDX56 expression is closely related to the clonogenic capacity in U-2 OS cells.

Knockdown of DDX56 in U-2 OS Cells Can Trigger Cell Apoptosis

The Annexin V–APC staining and flow cytometry analysis was carried out to test the relationship between cell apoptosis and DDX56 expression in U-2 OS cells. As is shown in **Figure 7E**, 4 days after shRNA lentivirus infection, U-2 OS cells in the

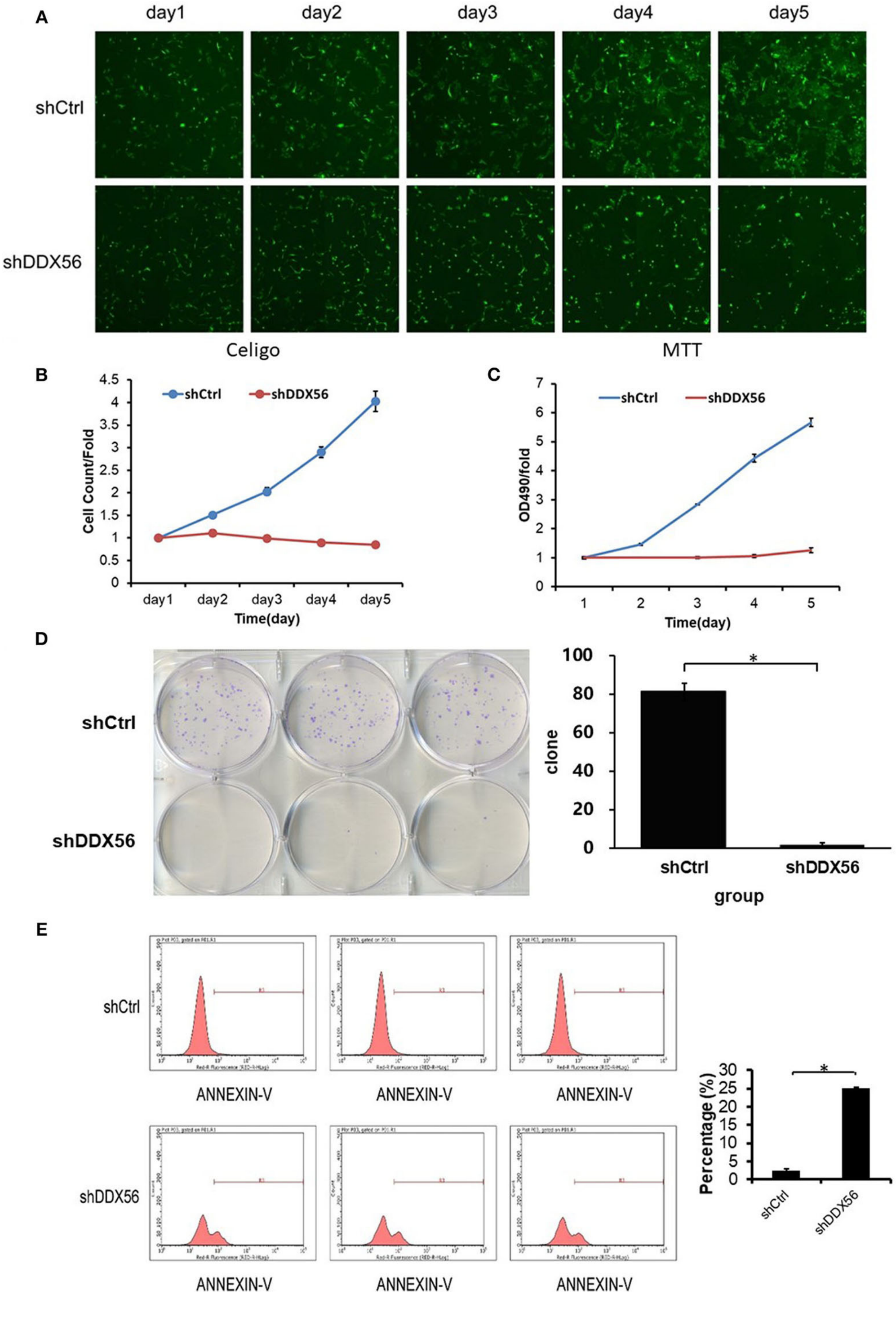


FIGURE 7 | Effect of DDX56 knockdown on U-2 OS cell growth. **(A)** Cells were infected with the control or DDX56–shRNA lentivirus, and high-content cell imaging was applied every day as indicated to acquire raw images (unprocessed by software algorithm) of cell growth. **(B)** Cells were seeded into 96-well plates and infected with the control or DDX56–shRNA lentivirus, and cell growth was assayed every day for 5 days (NC vs. DDX56–shRNA, $p < 0.05$). **(C)** MTT measurement of cell proliferation was performed in cells infected with the control or DDX56–shRNA lentivirus. The number of viable cells was 2,000 per well. The optical density (OD) at 490 nm was recorded in 5 days (NC vs. DDX56–shRNA, $p < 0.05$). **(D)** Effect of DDX56 knockdown on U-2 OS cell clonogenic ability. Colony formation assay was performed. Cells were seeded into a six-well plate 3 days after lentivirus. The number of viable cells was 400. The picture (left) was captured 9 days after seeding using a digital camera. The statistics showed significant difference in clonogenic potential between groups (NC vs. DDX56–shRNA, *$p < 0.05$). **(E)** Effect of DDX56 knockdown on U-2 OS cell apoptosis. Cell death was determined by Annexin V staining and flow cytometry. Cell cultures showed a significant increase in apoptosis compared with NC (*$p < 0.05$).

experimental group showed a significant increase in apoptosis portion (shCtrl 2.53 ± 0.33% vs. shDDX56 25.05 ± 0.24%, $p < 0.001$). These results indicate that DDX56 expression is a determinant of cell apoptosis in U-2 OS cells.

DISCUSSION

Osteosarcoma is the most prevalent primary bone tumor in children, adolescents, and elderly adults (Mckenna et al., 1987). The current overall treatment efficiency and recurrence remain unsatisfactory as the molecular mechanisms underlying the pathogenesis have not been fully determined. Recent studies revealed the potential effects of RNA splicing, assembly, and adjustment on tumor genesis (Inoue et al., 2019; Shuai et al., 2019; Suzuki et al., 2019). The DDX RNA helicases family represents the largest family of RNA helicases to be involved in cellular metabolism (Cordin et al., 2006; Patrick and Paul, 2006; Patrick, 2008). In this study, we found 15 encoding genes in the DEAD-box helicase family among 4,939 differentially expressed mRNAs between osteosarcoma tumor and adjacent normal tissues. Since protein–protein interactions play essential roles in various biological progresses (Wang et al., 2019), we then mapped the PPI network using the above coding genes. DDX56, the node gene with the highest degree in the PPI network was selected for further study.

Interaction network analysis has been proven effective in assisting to understand the pathogenesis of complex diseases (Wang et al., 2020). In this study, the ingenuity pathway analysis was used to find out potential mechanism and core pathways that might relate DDX56 to tumor cell proliferations, cell division, and apoptosis. The IPA database manually screened existing knowledge from over 20 years' literature and provides sensitive and accurate predictions on molecular interactions. Based on the DDX56-related osteosarcoma IPA network, the TP53 gene exhibited a crucial role in the connection and regulation of these nodes. The tumor protein p53 (TP53), also known as p53, is the most frequently mutated human gene that regulates the tumor suppression processes (Wang et al., 2014; Wang and Sun, 2016). It controls cell cycle arrest and apoptosis induced by chemotherapeutic agents including doxorubicin, by activating Bax, p21, PUMA (p53 Upregulated Modulator of Apoptosis), and Noxa (Levine et al., 1991). Previous studies found that p53 suppresses osteosarcoma cell proliferation, metastasis, and angiogenesis through inhibition of the PI3K/AKT/mTOR pathway (Song et al., 2015). Activation of p53-dependent signaling pathway promotes apoptosis in osteosarcoma cells and enhances sensitivity of osteosarcoma to the chemotherapy (Yuan et al., 2007; Yang et al., 2012). In the present study, we performed corresponding experiments to validate the effects of lentivirus-mediated DDX56 knockdown on these processes. First, we detected the mRNA and protein expression levels of DDX56 in osteosarcoma using public data and RT-PCR assay. We showed that DDX56 was upregulated in the GSE126209 dataset, TCGA SARC patients, and validated using osteosarcoma cell lines. Then, we downregulated DDX56 in U-2 OS cell lines via transfection of shRNA plasmids. We evaluated the role of DDX56 knockdown in proliferation, cell division, and apoptosis of osteosarcoma cells. As expected, we observed that DDX56 knockdown exerted a significant inhibitory effect on proliferation and clonogenic capacity, while significantly promoting cell apoptosis in U-2 OS cells. These findings suggest that the DDX56-modulated oncogenesis and p53 signaling-related osteosarcoma neoplasia may share a common molecular pathway. Further studies regarding the underlining mechanisms are worth exploring.

Moreover, we found 204 genes co-expressing and interacting with DDX56 in the studied gene profile. These genes are primarily involved in RNA processing-related pathways, including spliceosome, ribosome biogenesis in eukaryotes, and homologous recombination. Notably, alternative RNA splicing is an essential process to yield proteomic diversity in human malignancies (Inoue et al., 2019; Shuai et al., 2019), especially including osteosarcoma (Ajiro et al., 2016). Several DDX family members were reported to play roles in alternative splicing (Linder and Jankowsky, 2011; Bourgeois et al., 2016). DDX5 and DDX17 contribute to tumor cell invasiveness by regulating alternative splicing of several DNA- and chromatin-binding factors (Peters and Doets, 2009). As DDX56 shares common structures with the DDX family members (Linder and Jankowsky, 2011), DDX56 may also change splicing by spliceosome assembly alteration. A recent study has verified that DDX56 cell promotes proliferation in colorectal cancer through alternative splicing tumor suppressor WEE1 (Voss et al., 2015). We found in our experiment that DDX56 knockdown exerted a significant inhibitory effect on proliferation and clonogenic capacity, while significantly promoting cell apoptosis in U-2 OS cells. However, the underlining mechanism remains to be further verified.

Survival analysis considering DDX56 expression on the prognosis of osteosarcoma was performed using data from the TCGA database. Oddly, these data do not support significant difference in clinical outcomes between patients with high- and low-expressed DDX56 probably due to the significant heterogeneity between samples. In practice, it was also recognized that the overall prognosis was poor before chemotherapy (Anderson, 2016). Moreover, the number of samples with osteosarcoma in the TCGA database is limited. Even though it is the largest samples volume currently available with clinical and expression data, there are only 65 primary sarcoma patients included with high DDX56 expression, given that osteosarcomas represent fewer than 1% of cancers overall. In the TCGA database, DDX56 is overexpressed in osteosarcoma among other cancers. External validation was performed *in vitro* using several osteosarcoma cell lines to suggest that DDX56 might be a novel oncogene in osteosarcoma. More rigid designed prospective clinical survival observation as well as mechanical studies should be performed in order to further validate this hypothesis.

There are some limitations that should be acknowledged. First of all, the sample size is relatively small due to the low overall incidence of the rare musculoskeletal malignancy. Second, the selected subjects all come from the Chinese Uyghur population. Given the potential ethnic specificity, the genetic background is possibly different between the Uyghur cells and purchased cells, which might affect the extension of the conclusion. Third, the IPA program builds its models by querying the known literature. Unknown interactions could not be discovered through this analysis, and thus, it is likely that there are highly relevant interactions that do not emerge in IPA. The next limitation is the fact that, since IPA queries only known associations and interactions, genes about which little or nothing is known about the function of their products cannot be identified as hubs using this method. In addition, the PPI network was constructed using a previous published database (IID) (Kotlyar et al., 2016); those unidentified but existing protein interaction relationships may have been missed. Given these limitations, the models generated here must be considered preliminary and incomplete. Other predictive tools, such as predictions of protein interactions based on molecular structure, specific groups, may be good cross-validations (Deng et al., 2020; Hu et al., 2020). Further studies should also be done to illustrate the underlying mechanisms.

In conclusion, we have identified DDX56 as a novel oncogene using bioinformatics tools and demonstrated that DDX56 was overexpressed in osteosarcoma tissues and cell lines. Furthermore, DDX56 knockdown inhibited cell proliferation and promoted cell apoptosis in osteosarcoma. These findings propose that DDX56 may be considered as a potential therapeutic target for the treatment of osteosarcoma.

DATA AVAILABILITY STATEMENT

The datasets analyzed in the present study are available in the Gene Expression Omnibus repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126209.

ETHICS STATEMENT

All experiments were conducted on commercially supplied cell lines. Therefore, ethics approval and written informed consent was not required for this study.

AUTHOR CONTRIBUTIONS

XZ conceived the idea and designed the project. WH performed the data analysis. CZ and XZ wrote the paper. XZ, CZ, and NK revised the manuscript. YS provided the administrative support. CZ obtained the funding support. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 81871788), the project for Science and Technology leader of Anhui Province (Grant No. 2018H177), the Scientific Research Fund of Anhui Education (Grant No. 2017jyxm1097), and the Anhui Provincial Post-doctoral Science Foundation (Grant No. 2019B302).

ACKNOWLEDGMENTS

We thank Mr. Donglin Cheng for his technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00588/full#supplementary-material

Figure S1 | The heatmap with hierarchical clustering for all 4,940 differentially expressed genes (DEGs) in GSE126209 dataset.

Figure S2 | (A) Gender difference on expression of DEAD-box helicase (DDX)56 was not significant ($p > 0.05$). **(B)** The expressions of DDX56 are indifferent among sarcoma patients from different races ($p > 0.05$). **(C)** Survival plot-effect of DDX56 expression level and gender on sarcoma (SARC) patient survival. **(D)** Survival plot-effect of DDX56 expression level and race on SARC patient survival.

REFERENCES

Ajiro, M., Jia, R., Yang, Y., Zhu, J., and Zheng, Z. M. (2016). A genome landscape of SRSF3-regulated splicing events and gene expression in human osteosarcoma U2OS cells. *Nucleic Acids Res.* 44, 1854–1870. doi: 10.1093/nar/gkv1500

Anderson, M. E. (2016). Update on survival in osteosarcoma. *Orthoped. Clin. North Am.* 47:283. doi: 10.1016/j.ocl.2015.08.022

Andreas, K., Jeff, G., Jack, P., and Stuart, T. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30:523. doi: 10.1093/bioinformatics/btt703

Bourgeois, C. F., Mortreux, F., and Auboeuf, D. (2016). The multiple functions of RNA helicases as drivers and regulators of gene expression. *Nat. Rev. Mol. Cell Biol.* 17, 426–438. doi: 10.1038/nrm.2016.50

Chandrashekar, D. S., Bashel, B., Sah, B., Creighton, C. J., Ponce-Rodriguez, I., Bvsk, C., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Cordin, O., Banroques, J., Tanner, N. K., and Linder, P. (2006). The DEAD-box protein family of RNA helicases. *Gene* 367, 17–37. doi: 10.1016/j.gene.2005.10.019

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., et al. (2020). Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms21072274

Gene Ontology, C. (2006). The gene ontology (GO) project in 2006. *Nucleic Acids Res.* 34, D322–D326. doi: 10.1093/nar/gkj021

He, Y., Zhang, D., Yang, Y., Wang, X., Zhao, X., Zhang, P., et al. (2018). A double-edged function of DDX3, as an oncogene or tumor suppressor, in cancer progression. *Oncol. Rep.* 39, 883–892. doi: 10.3892/or.2018.6203

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for QSAR prediction. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009. [Epub ahead of print].

Inoue, D., Chew, G. L., Liu, B., Michel, B. C., Pangallo, J., D'Avino, A. R., et al. (2019). Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature* 574, 432–436. doi: 10.1038/s41586-019-1646-9

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* 44, D536–D541. doi: 10.1093/nar/gkv1115

Kumar, S., Alazraki, A., George, B., Martin, M., Katzenstein, H. M., and Cash, T. (2017). Pulmonary relapse of osteosarcoma following presentation with a pneumomediastinum and localized disease at diagnosis. *J. Pediatr. Hematol. Oncol.* 39, e446–e449. doi: 10.1097/MPH.0000000000000821

Kuramitsu, Y., Tominaga, W., Baron, B., Tokuda, K., Wang, Y., Kitagawa, T., et al. (2013). Up-regulation of DDX39 in human malignant pleural mesothelioma cell lines compared to normal pleural mesothelial cells. *Anticancer Res.* 33, 2557–2560.

Levine, A. J., Momand, J., and Finlay, C. A. (1991). The p53 tumour suppressor gene. *Nature* 351, 453–456. doi: 10.1038/351453a0

Linder, P., and Jankowsky, E. (2011). From unwinding to clamping–the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.* 12:505. doi: 10.1038/nrm3154

Lonsdale, J. T., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Mckenna, W. G., Barnes, M. M., Kinsella, T. J., Rosenberg, S. A., Lack, E. E., Glatstein, E., et al. (1987). Combined modality treatment of adult soft tissue sarcomas of the head and neck. *Int. J. Radiat. Oncol. Biol. Phys.* 13, 1127–1133. doi: 10.1016/0360-3016(87)90184-2

Misha, K., Patrick, K., Culhane, A. C., Steffen, D., Jan, I., Christine, K. R., et al. (2004). Expression profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Res.* 32, 465–470. doi: 10.1093/nar/gkh470

Ottaviani, G., and Jaffe, N. (2009). The epidemiology of osteosarcoma. *Cancer Treat. Res.* 152:3. doi: 10.1007/978-1-4419-0284-9_1

Patrick, L. (2008). mRNA export: RNP remodeling by DEAD-box proteins. *Curr. Biol.* 18, R297–R299. doi: 10.1016/j.cub.2008.02.027

Patrick, L., and Paul, L. (2006). Bent out of shape: RNA unwinding by the DEAD-box helicase vasa. *Cell* 125, 219–221. doi: 10.1016/j.cell.2006.03.030

Peters, K. S., and Doets, H. C. (2009). Midterm results of cementless total hip replacement in rapidly destructive arthropathy and a review of the literature. *Hip Int.* 19, 352–358. doi: 10.1177/112070000901900409

Pragya, K., Surya, K., Michal, G., Ruth, S., and Simon, B. (2007). *STRESS RESPONSE SUPPRESSOR1* and *STRESS RESPONSE SUPPRESSOR2*, two DEAD-box RNA helicases that attenuate Arabidopsis responses to multiple abiotic stresses. *Plant Physiol.* 145, 814–830. doi: 10.1104/pp.107.099895

Reid, C. R., and Hobman, T. C. (2017). The nucleolar helicase DDX56 redistributes to West Nile virus assembly sites. *Virology* 500, 169–177. doi: 10.1016/j.virol.2016.10.025

Shuai, S., Suzuki, H., Diaz-Navarro, A., Nadeu, F., Kumar, S. A., Gutierrez-Fernandez, A., et al. (2019). The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* 574, 712–716. doi: 10.1038/s41586-019-1651-z

Song, R., Tian, K., Wang, W., and Wang, L. (2015). P53 suppresses cell proliferation, metastasis, and angiogenesis of osteosarcoma through inhibition of the PI3K/AKT/mTOR pathway. *Int. J. Surg.* 20, 80–87. doi: 10.1016/j.ijsu.2015.04.050

Sugiura, T., Sakurai, K., and Nagano, Y. (2007). Intracellular characterization of DDX39, a novel growth-associated RNA helicase. *Exp. Cell Res.* 313, 782–790. doi: 10.1016/j.yexcr.2006.11.014

Suzuki, H., Kumar, S. A., Shuai, S., Diaz-Navarro, A., Gutierrez-Fernandez, A., De Antonellis, P., et al. (2019). Recurrent non-coding U1-snRNA mutations drive cryptic splicing in Shh medulloblastoma. *Nature* 574, 707–711. doi: 10.1038/s41586-019-1650-0

Umate, P., Tuteja, R., and Tuteja, N. (2010). Genome-wide analysis of helicase gene family from rice and Arabidopsis: a comparison with yeast and human. *Plant Mol. Biol.* 73:449. doi: 10.1007/s11103-010-9632-5

Voss, M. R. H. V., Vesuna, F., Trumpi, K., Brilliant, J., Berlinicke, C., Leng, W. D., et al. (2015). Identification of the DEAD box RNA helicase DDX3 as a therapeutic target in colorectal cancer. *Oncotarget* 6, 28312–28326. doi: 10.18632/oncotarget.4873

Wang, B., Wang, L., Zheng, C. H., and Xiong, Y. (2019). Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2953908. [Epub ahead of print].

Wang, W., Zhou, Y., Cheng, M. T., Wang, Y., Zheng, C. H., Xiong, Y., et al. (2020). Potential pathogenic genes prioritization based on protein domain interaction network analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.2983894. [Epub ahead of print].

Wang, X., Chen, J. X., Liu, J. P., You, C., Liu, Y. H., and Mao, Q. (2014). Gain of function of mutant TP53 in glioblastoma: prognosis and response to temozolomide. *Ann. Surg. Oncol.* 21:1337. doi: 10.1245/s10434-013-3380-0

Wang, X., and Sun, Q. (2016). TP53 mutations, expression and interaction networks in human cancers. *Oncotarget* 8:624. doi: 10.18632/oncotarget.13483

Wright, G. W., and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19, 2448–2455. doi: 10.1093/bioinformatics/btg345

Xu, Z., and Hobman, T. C. (2012). The helicase activity of DDX56 is required for its role in assembly of infectious West Nile virus particles. *Virology* 433, 226–235. doi: 10.1016/j.virol.2012.08.011

Yang, W., Wei, Y., Shi, Y., Li, Y., and Li, R. (2012). Arsenic trioxide induces apoptosis of p53 null osteosarcoma MG63 cells through the inhibition of catalase. *Med. Oncol.* 29, 1328–1334. doi: 10.1007/s12032-011-9848-5

Yuan, X., Zhu, X., Huang, X., Sheng, P., He, A., and Yang, Z. (2007). Interferon-a enhances sensitivity of human osteosarcoma U2OS cells to doxorubicin by p53-dependent apoptosis. *Acta Pharmacol. Sin.* 28, 1835–1841. doi: 10.1111/j.1745-7254.2007.00662.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fbioe.2020.00719

Predictive Value of the Texture Analysis of Enhanced Computed Tomographic Images for Preoperative Pancreatic Carcinoma Differentiation

Zhang Longlong[1], *Li Xinxiang*[2], *Ge Yaqiong*[3] *and Wei Wei*[2]*

[1] Department of Radiology, Anhui Provincial Hospital Affiliated to Anhui Medical University, Hefei, China, [2] Department of Radiology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China, [3] GE Healthcare China, Shanghai, China

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Yonglong Luo,
Anhui Normal University, China
Junfeng Xia,
Anhui University, China

****Correspondence:***
Wei Wei
weiweill@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *24 March 2020*
Accepted: *08 June 2020*
Published: *30 June 2020*

Citation:
Longlong Z, Xinxiang L, Yaqiong G and Wei W (2020) Predictive Value of the Texture Analysis of Enhanced Computed Tomographic Images for Preoperative Pancreatic Carcinoma Differentiation. Front. Bioeng. Biotechnol. 8:719. doi: 10.3389/fbioe.2020.00719

Purpose: To assess the utility of texture analysis for predicting the pathological degree of differentiation of pancreatic carcinoma (PC).

Methods: Eighty-three patients with PC who went through postoperative pathology diagnose and CT examination were selected at Anhui Provincial Hospital. Among them, 34 cases were moderately differentiated, 13 cases were poorly differentiated, and 36 cases were moderately poorly differentiated. The images in the arterial and venous phase (VP) with the lesions at their largest cross section were selected to manually outline the region of interest (ROI) to delineate lesions using open-source software. A total of 396 features were extracted from the ROI using AK software. Spearman correlation analysis and random forest selection by filter (rfSBF) in the caret package of R studio were used to select the discriminating features. The receiver operating characteristic ROC analysis was used to evaluate their discriminative performance.

Results: Twelve and six features were selected in the arterial and VPs, respectively. The areas under the ROC curve (AUC) in the arterial phase (AP) for diagnosing poorly differentiated, moderately differentiated and moderate-poorly differentiated cases were 0.80, 1, and 0.80 in the training group and 0.77, 1, and 0.77 in the test group; in the VP, the values were 0.81, 1, and 0.82 in the training group and 0.74, 1, and 0.74 in the test group.

Conclusion: Texture analysis based on contrast-enhanced CT images can be used as an adjunct for the preoperative assessment of the pathological degrees of differentiation of PC.

Keywords: pancreatic carcinoma, texture analysis, contrast-enhanced CT, pathological grading, machine learning

INTRODUCTION

Although all the efforts have been made to develop better pancreatic carcinoma (PC) treatment strategies, the prognosis still remains poor. PC, which is a tumor with a very high degree of malignancy and it is usually found at an advanced stage (Luu et al., 2019). Early diagnosis remains challenging if the tumor is not located close to the common bile duct, causing obstructive jaundice (Siegel et al., 2016). Some patients have already missed the best treatment opportunity when PC is discovered. PC cells are highly invasive and prone to metastasis and the 5-year survival rate is only 5–7% (Schima et al., 2007; Siegel et al., 2016). Wang et al. (2003, 2007) reported that the degree of enhancement and differentiation of the tumor were inversely proportional to the degree of malignancy. The lower the degree of enhancement, the higher the degree of malignancy, and the lower the degree of differentiation, the higher the degree of malignancy. Surgery is the only effective method to cure PC and is still considered for most lesions (Cloyd and Poultsides, 2015; Wong et al., 2018). Different pathological grades of PC have different prognoses (Kurihara et al., 2015; Matsumoto et al., 2015). Therefore, the preoperative prediction of the pathological grade and differentiation of lesions is very important in the treatment and prognostic evaluation of patients with PC (Kurihara et al., 2015; Matsumoto et al., 2015). PC can be detected with computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US), but CT is still the most commonly used method for the diagnosis of PC (Chun-Ye et al., 2012). Recently, texture analysis technology has become a topic of growing interest. This technology is widely used in the diagnosis, differential diagnosis and pathological grading of diseases by extracting potential information from the image and analyzing the extracted texture features. When the radiologists diagnose the disease, it is mainly based on the observation of the image of the lesion and the clinical manifestation of the patient. The application of texture analysis in imaging can quantify the imaging features of the disease, thus assisting the imaging physician in the diagnosis (Davnall et al., 2012; Ng et al., 2013; Rao et al., 2014; Hanania et al., 2016; Choi et al., 2018).

Some researches had shown CT texture analysis was helpful to the prediction of the resectability and prognosis in patients after neoadjuvant therapy for pancreatic ductal adenocarcinoma and the prediction of pancreatic neuroendocrine tumor grade (Canellas et al., 2018; Choi et al., 2018; Wong et al., 2018). Sandrasegaran et al. (2019) has reported that CT texture analysis was easy to perform on contrast-enhanced CT and it could determine prognosis in patients with unresectable PC. Huang et al.'s (2019) study showed that two-dimensional texture analysis was a feasible quantitative technique for the differential diagnosis of pancreatic lymphoma from pancreatic adenocarcinoma, and the diagnostic performance was similar to CT characteristics. To the best of our knowledge, texture analysis has not been used as a method for predicting PC differentiation. The purpose of this study is to explore the value of texture analysis on CT images for predicting PC differentiation.

MATERIALS AND METHODS

Ethical Approval

The studies involving human participants were reviewed and approved by Medical Research Ethics Committee of The First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Hospital).

Patients

A retrospective analysis was performed on patients at the Department of Imaging of the Anhui Provincial Hospital from 2013 to 2019 who met the following inclusion criteria: (1) contrast-enhanced multiphase abdominal CT scan before surgery; (2) lesion size ≥10 mm; (3) a single mass. The exclusion criteria were as follows: (1) received chemotherapy or other treatment before CT examination; (2) poor-quality images; (3) the tumor was transferred, so the patient could not undergo surgery. All included patients underwent surgical treatment within 2 weeks after enhanced CT scan. The chi-square test was used to compare the differences in the sex distribution between the groups. $P < 0.05$ was considered statistically significant. A total of 83 patients with PC diagnosed by pathology from the Anhui Provincial Hospital were collected, including 54 males and 29 females, aged from 41 to 76 years, with a mean age of 60.69 ± 9.12 years, as shown in **Figure 1**. The main symptoms of the patients when they came to the hospital for treatment were jaundice (74 cases, 89.1%), yellow urine (44 cases, 53.0%), abdominal discomfort (including abdominal distension or abdominal pain, 50 cases, 60.2%), itchy skin (5 cases, 6.0%), fever (2 cases, 2.4%), and diarrhea (1 case, 1.2%); in addition, one patient was found to have no obvious clinical symptoms on medical examination.

Scanning Method

All patients fasted for 6 h before the CT scan. The examination was carried out according to the following protocol: intramuscular injection of anisodamine (654-II) 10–15 mg and 800–1000 mL of clear water orally 15 min before the scan. All examinations were performed using a multidetector CT scanner (DiscoveryHD750, Gemstone Spectral Imaging, GE Healthcare, Milwaukee, WI, United States). All patients underwent routine unenhanced/three-phase enhanced CT scans to determine the extent of the lesion. The parameters for abdominal CT were as follows: tube voltage 120 kVp, tube current 250–350 mA, slice thickness 5 mm, slice interval 5 mm, field of view 35–50 cm, matrix 512 × 512, rotation time 0.7 s and pitch 1.375. After unenhanced CT scans were obtained, each patient received 1.5 mL per kilogram of body weight of non-ionic iodinated contrast material (Iohexol, Omnipaque 300, GE Healthcare, Shanghai, China), which was administered at a rate of 3.0 mL/s using a power injector (Stellant; Medrad, Warrendale, PA, United States). The scanning delay for arterial phase (AP) imaging was determined using automated scan triggering software (SmartPrep; GE Healthcare, Milwaukee, WI, United States). AP scanning automatically began 10 s after the density in the descending aorta reached 100 HU on the

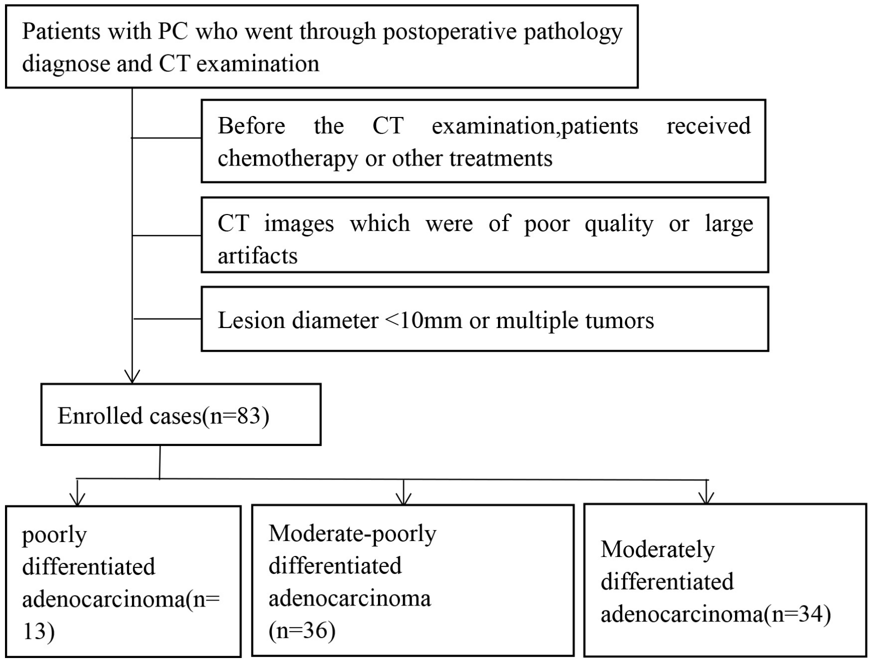

Patients with PC who went through postoperative pathology diagnose and CT examination
Before the CT examination,patients received chemotherapy or other treatments
CT images which were of poor quality or large artifacts
Lesion diameter <10mm or multiple tumors
Enrolled cases(n=83)
poorly differentiated adenocarcinoma(n= 13)
Moderate-poorly differentiated adenocarcinoma (n=36)
Moderately differentiated adenocarcinoma(n=34)

FIGURE 1 | Screening and grouping flow chart of enrolled cases in this study.

monitoring scan. At delays of 30 s and 3 min after AP scanning, venous phase (VP) and delayed phase (DP) acquisitions commenced, respectively.

Region-of-Interest (ROI) Segmentation and Radiomics Feature Extraction

The images of the arterial and VPs of all patients were collected from the CT ICPACS workstation of the Department of Radiology, Anhui Provincial Hospital, and exported in DICOM format. Two imaging physicians with 15 years of diagnostic experience in the CT diagnosis of abdominal disease used open-source software[1] to delineate Region-of-Interests (ROIs) containing the target lesions in the arterial and venous images. Open-source ITK-SNAP software was used for ROI sketching. During the process of sketching, the operators selected the largest area of the enhancement on the AP and VP of the tumor, paying attention to avoiding the pancreatic duct, blood vessels, calcification, and necrotic cystic areas to minimize errors.

Statistical Analysis and Clinical Predictive Model-Building

To ensure the intra- and inter-observer reproducibility, 30 patients were randomly selected and delineated by the radiologists1 for twice to calculate the intra-observer ICCs, and delineated by radiologist2 for once to calculate the inter-observer ICCs. An ICC>0.75 indicated good reproducibility. Radiologist1 finished the rest delineation.

The original images were normalized by transforming them into standard intensity ranges with a mean value of 0 and a standard deviation of 1 (z-score transformation) before the image features were extracted. AK software (GE Healthcare, Analysis Kit, Version: 3.2.0. R) was used to extract a total of 396 feature parameters, of which 42 were histogram features, nine were form factors, 154 were gray level co-occurrence matrix (GLCM)

[1] www.itksnap.org

TABLE 1 | Comparison of the clinical data of the cases grouped according to the degree of differentiation.

Degree of pathological differentiation	Number of cases (patients)	Sex		Age (year)	CT value (Hu)	
		Male	Female		Arterial phase	Venous phase
Poor differentiation	13	9	4	42~76	52.1 ± 9.2	67.7 ± 8.4
Moderate-poor differentiation	36	24	12	43~76	56.2 ± 9.4	72.5 ± 13.2
Moderate-poor differentiation	34	21	13	41~75	65.7 ± 9.9	84.5 ± 11.7
T value		χ^2 = 0.303		*F* = 1.213	*F* = 12.660	*F* = 13.498
P value		>0.5		>0.05	<0.05	<0.05

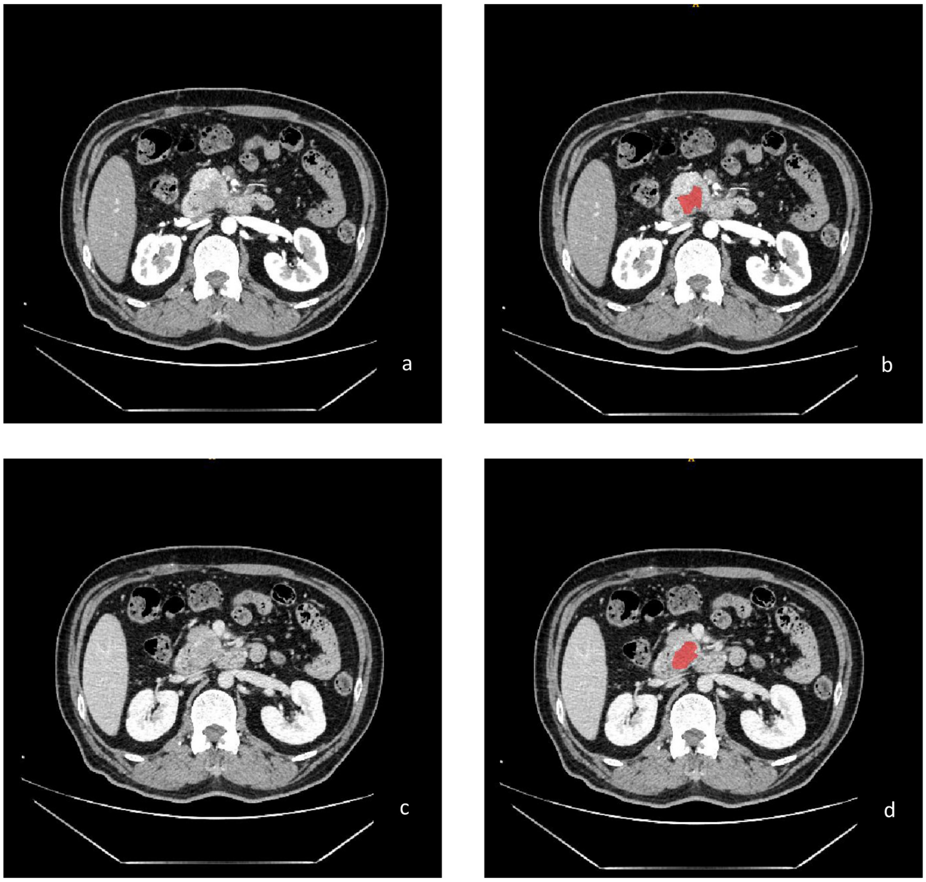


FIGURE 2 | Enhanced CT images of moderately differentiated PC in the arterial phase **(a)** and venous phase **(c)**; the treatment ROI using ITK-SNAP software **(b,d)**.

features, 11 were gray-level size zone matrix (GLSZM) features, and 180 were run length matrix (RLMs) features.

All statistical analyses were performed in R (3.5.1[2]). Patients were randomly assigned to the training and test group at a ratio of 7:3. In training group, first, Spearman correlation analysis was used to eliminate features with correlation coefficients >0.9. Next, random forest selection by filter (rfSBF) in the "random forest" package with ten-fold cross validation tests was used to select the best feature subsets in each phase. Then, the conditional inference tree "ctree" of the "train" function in the "caret" package was applied in the best sub-feature groups to train the predictive model. Afterward, the "pROC" package was applied to evaluate the discriminative performance of the model and validated in the test group. Other parameters, including accuracy, sensitivity and specificity, were calculated by the "confusion matrix."

RESULTS

Clinical Characteristics

A total of 83 cases of PC were included in this study. The postoperative pathological grade included 13 cases of poorly differentiated adenocarcinoma, 36 cases of moderately poorly differentiated adenocarcinoma, and 34 cases of moderately differentiated adenocarcinoma. The CT values of the AP and VP of the three groups were statistically significant ($P < 0.05$) (**Table 1**). Part of the images and ROIs are shown in **Figure 2**.

Feature Selection and Radiomics Signature Building

The range of the inter-observer ICCs in the Artery phase were from 0.0096 to 1, and the intra-observer ICCs were from 0.14 to 1; In VP, inter-observer ICCs were from 0.06 to 0.94, and intra-observer ICCs were from 0.26 to 1.0, 282 features for Artery phase and 274 for VP with an ICC value bigger than 0.75 were selected for the further analysis. Spearman correlation analysis and the random forest method were used to select features. The final retained features of the arterial and VPs are shown in **Figure 3**. Twelve features were retained in the AP (a), six features were retained in the VP (b), and 12 features were retained in the combined group (c); the x-axis represents the weight of the features, the y-axis represents the last retained features, and the larger the weight is, the more predictive the feature is. In the AP, the three most predictive features in the poorly differentiated group, moderately differentiated group

[2] http://www.Rproject.org

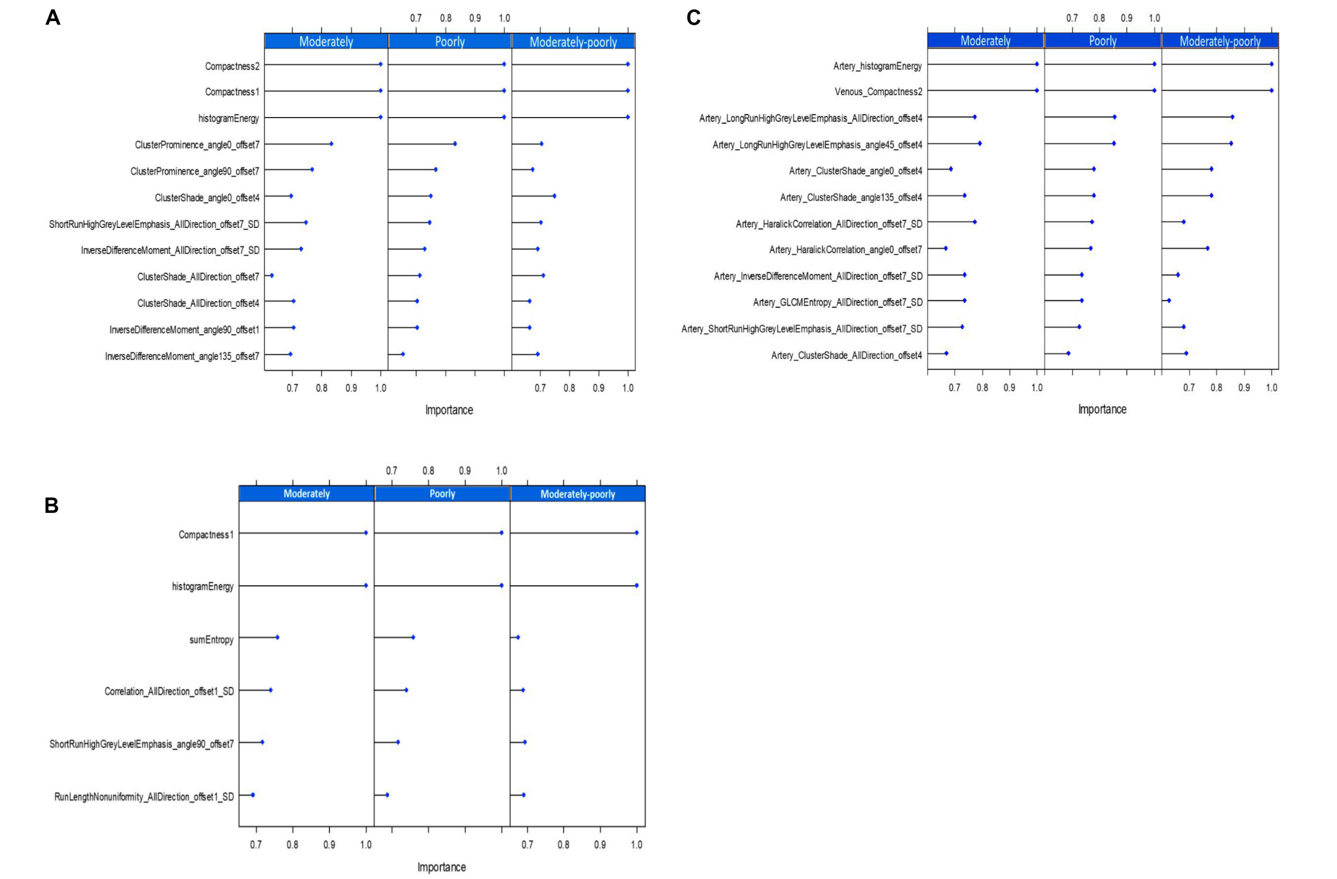


FIGURE 3 | The best texture features obtained by the random forest, The arterial phase **(A)**, the venous phase **(B)**, and the combined group **(C)**.

and moderate-poorly differentiated group were compactness2, compactness1 and histogramEnergy; in the VP, the two most predictive features were compactness1 and histogramEnergy; in the combined group, the two most predictive features were artery-histogramenergy and venous-compactness2. We can see that in both the AP and VP, compactness1 from "Form factor" and histogramEnergy from "histogram" show the best predictive performance. The explanation of some texture features is shown in **Table 2**.

The AUC(95%CI) value of the three models to differentiate the degree of differentiation in the training group and test group of PC patients were shown in **Figures 4A–F**. Other parameters, including sensitivity, specificity, and model accuracy were shown in **Table 3**, The overall accuracies in the AP and VP were 0.77(95%CI: 0.64–0.87) and 0.77(95%CI: 0.62–0.86) in the training group and 0.74(95%CI: 0.52–0.89) and 0.70(95%CI: 0.47–0.86) in the test group, respectively. We can see that neither the AUC nor sensitivity and specificity in both the AP and VP have a one hundred percent value in differentiating moderate differentiation. Additionally, the model has good performance in poor differentiation and moderate-poor differentiation, and the accuracy of the model was higher than 0.74 in both the training and test groups. Thus, it can be considered that texture analysis can be used for enhanced CT scanning and has obvious texture features. The comparison also has a certain predictive effect, and the malignancy of pancreatic cancer can be evaluated to a certain extent.

DISCUSSION

Pancreatic carcinoma is a very malignant tumor with a 5-year survival rate of <6%. And it is possible that PC will become the second leading cause of death from malignancy in the next two decades (Rahib et al., 2014; Ferlay et al., 2015). The disease is usually detected at a late clinical stage that is difficult to treat. Therefore, the preoperative evaluation of the degree of malignancy and resectability of PC is very important for the operation, postoperative treatment and prognosis of patients (Kurihara et al., 2015; Matsumoto et al., 2015). This study provides a radiomics features based machine learning model for predicting the degree of differentiation of PC before surgery. And the results show that the model has high feasibility and credibility.

Texture analysis has been shown great value in medical image preprocessing This technology is not affected by photon noise and can quantitatively measure tumor heterogeneity.

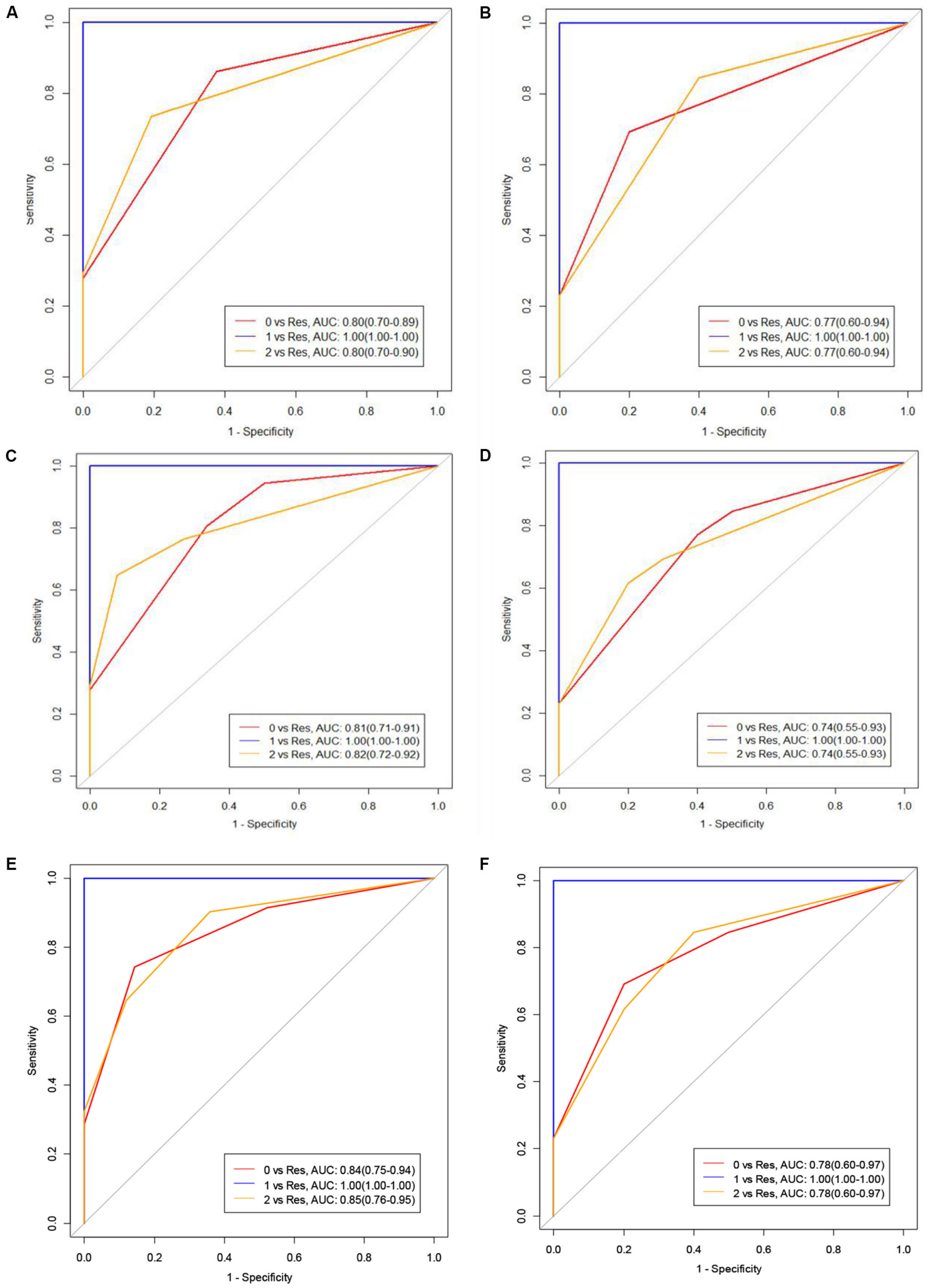


FIGURE 4 | (A,B) represent the training and test groups with moderately differentiated, poorly differentiated, and moderate-poorly differentiated pathologies of PC. **(C,D)** represent the AUC values of the venous phase of the training and test groups with moderately differentiated, poorly differentiated and moderate-poorly differentiated PC. **(E,F)** represent the combined groups (Class 0 means moderately differentiated, Class 1 means poorly differentiated, Class 2 means moderate-poorly differentiated, Res means the rest of the cases).

TABLE 2 | Features measured with different texture analysis methods by AK software.

Texture feature groups	Parameters
Histogram Features (Gray intensity information and its distribution of the lesion, for example HistogramEnergy describes the severity of the change in image brightness information, the smaller the change, the greater the Energy.)	Mean, Variance, Uniformity, Skewness, Kurtosis, Energy, Entropy
Form Factor Features (The shape of the lesion, For example, Compactness, describing the degree of roundness or sphericity of the lesion; if the lesion is more spherical, the Compactness value is greater.)	Volume CC, Surface, Surface Volume Ratio, Compactness, Maximum 3D Diameter
GLCM Features (Obtained by counting the probability of pixel pairs in different directions and step sizes)	Entropy, Inertia, Inverse Difference Moment;
RLM Features (obtained by counting the probability of multiple occurrences of pixels in different directions and steps)	Short Run Emphasis, Low Gray Level Run Emphasis, Short Run Low Gray Level Emphasis;
GLSZM Features (obtained by counting the number of pixels with the same adjacent gray value, so as to obtain the gray connected area matrix)	Small Zone Emphasis, Low Gray Level Zone Emphasis, Short Run Low Gray Level Emphasis

GLCM, gray level co-occurrence matrix; RLM, Run Length Matrix; GLSZM, Gray Level Size Zone Matrix. All of them are used to describe the complexity of the lesion site, level changes, and the thickness of the texture and other information.

It is widely used in the diagnosis, differential diagnosis and therapeutic evaluation of tumors (Davnall et al., 2012; Ng et al., 2013). Texture analysis can parameterize the potential information in the inspected image to obtain more abundant quantitative data, which will facilitate structured analysis and the processing of data. Kim et al. (2019) and Sandrasegaran et al. (2019) determined that texture analysis is useful for predicting a patient's prognosis and resectability of the tumor, after neoadjuvant therapy for PDAC. Canellas et al. (2018) showed that texture analysis has certain feasibility in the classification of pancreatic neuroendocrine tumors. We can boldly hypothesize that this technology can also be used for pathological grading of pancreatic cancer.

Watanabe et al. (2014) showed that when more tumor-associated fibrosis is present, the stronger the tumor invasiveness, and the higher the degree of malignancy and that the overexpression of fibroblast activation protein in PC tissue causes an increase in the lesion interstitial fiber component. Blocking the contrast agent into the lesion weakens the degree of enhancement of the lesion. The results of this study are consistent with the findings of Wang et al. (2003, 2007), that the higher the malignancy of the tumor, the lower the degree of differentiation and enhancement. In addition, Eilaghi et al. (2017) showed that CT image-based texture analysis can effectively distinguish between tumor and normal pancreatic tissue, and CT texture features may become an imaging biomarker for the postoperative overall survival rate. Our study showed that the feasibility of CT image-based texture analysis for the preoperative prediction of PC differentiation.

The limitations of this study are as follows: (1) the sample size was limited, and the measured values of the parameters may be biased; (2) there was no uniform standard when selecting the ROI in this study, and manual image sketching is time-consuming and laborious; and (3) in this study, feature extraction was based on a single enhanced image for analysis. The better way is to use 3D stereo modeling to extract texture features for analysis. In actual work, imaging physicians use CT plain scans and three-phase enhancement methods to analyze and diagnose the lesions or perform multiple imaging examinations at the same time to achieve a more comprehensive diagnosis.

In summary, CT texture analysis in PC has a clear predictive value for identifying differences in tumor grading. It provides a new method for assessing the malignant degree of tumor grading. It has the advantage of being a non-traumatic examination method, it is not dependent the opinion or experience of radiologists, that still permits the accurate diagnosis of patients with cancer.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Research Ethics Committee of The First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Hospital). The

TABLE 3 | Sensitivity and specificity of the arterial and venous phases training and test groups of PC with different degrees of differentiation.

Phase	Arterial			Phase	Venous		
	Group	Sensitivity	Specificity		Group	Sensitivity	Specificity
Training group: Accuracy 0.77(95%CI:0.64–0.87)	Class 0	0.63	0.86	Training group: Accuracy 0.77(95%CI:0.62–0.86)	Class 0	0.50	0.94
	Class 1	1.00	1.00		Class 1	1.00	1.00
	Class 2	0.81	0.74		Class 2	0.92	0.65
Test group: Accuracy 0.74(95%CI:0.52–0.89)	Class 0	0.80	0.69	Test group: Accuracy 0.70(95%CI:0.47–0.86)	Class 0	0.50	0.85
	Class 1	1.00	1.00		Class 1	1.00	1.00
	Class 2	0.60	0.84		Class 2	0.80	0.62

patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WW and LX contributed to conception and design of the study. ZL organized the database. GY performed the statistical analysis. ZL and GY wrote the first draft of the manuscript. ZL, WW, LX, and GY wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Youth Science Fund Project (No. 81501468). The Fundamental Research Funds for the Central Universities (No. WK9110000002).

REFERENCES

Canellas, R., Burk, K. S., Parakh, A., and Sahani, D. V. (2018). Prediction of pancreatic neuroendocrine tumor grade based on CT features and texture analysis. *AJR Am. J. Roentgenol.* 210, 341–346. doi: 10.2214/ajr.17.18417

Choi, T. W., Kim, J. H., Yu, M. H., Park, S. J., and Han, J. K. (2018). Pancreatic neuroendocrine tumor: prediction of the tumor grade using CT findings and computerized texture analysis. *Acta Radiol.* 59, 383–392. doi: 10.1177/0284185117725367

Chun-Ye, Q., Xun, S., and Ming, G. (2012). Correlation between spiral CT preoperative staging of pancreatic cancer and PTEN and COX-2 expression. *Hepato Gastroenterol.* 59, 2000–2002. doi: 10.5754/hge12031

Cloyd, J. M., and Poultsides, G. A. (2015). Non-functional neuroendocrine tumors of the pancreas: advances in diagnosis and management. *World J. Gastroenterol.* 21, 9512–9525. doi: 10.3748/wjg.v21.i32.9512

Davnall, F., Yip, C. S. P., Ljungqvist, G., Selmi, M., Ng, F., Sanghera, B., et al. (2012). Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging* 3, 573–589. doi: 10.1007/s13244-012-0196-6

Eilaghi, A., Baig, S., Zhang, J., Karanicolas, P., Gallinger, S., Khalvati, F., et al. (2017). CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma - a quantitative analysis. *BMC Med. Imaging* 17:38. doi: 10.1186/s12880-017-0209-5

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Hanania, A. N., Bantis, L. E., Feng, Z., Wang, H., Tamm, E. P., Katz, M. H., et al. (2016). Quantitative imaging to evaluate malignant potential of IPMNs. *Oncotarget* 7, 85776–85784. doi: 10.18632/oncotarget.11769

Huang, Z.-X., Li, M., He, D., Wei, Y., Yu, H.-P., Wang, Y., et al. (2019). Two-dimensional texture analysis based on CT images to differentiate pancreatic lymphoma and pancreatic adenocarcinoma: a preliminary study. *Acad Radiol.* 26, e189–e195. doi: 10.1016/j.acra.2018.07.021

Kim, B. R., Kim, J. H., Ahn, S. J., Joo, I., Choi, S.-Y., Park, S. J., et al. (2019). CT prediction of resectability and prognosis in patients with pancreatic ductal adenocarcinoma after neoadjuvant treatment using image findings and texture analysis. *Eur. Radiol.* 29, 362–372. doi: 10.1007/s00330-018-5574-0

Kurihara, T., Kogo, M., Ishii, M., Yoneyama, K., Kitamura, K., Shimada, K., et al. (2015). Practical prognostic index for survival in patients with unresectable pancreatic cancer treated with gemcitabine or S-1. *Hepato Gastroenterol.* 62, 478–484.

Luu, A. M., Hoehn, P., Vogel, S. R., Reinacher-Schick, A., Munding, J., Uhl, W., et al. (2019). Pathologic complete response of pancreatic cancer following neoadjuvant FOLFIRINOX treatment in hepatic metastasized pancreatic cancer. *Visceral Med.* 35, 387–391. doi: 10.1159/000497827

Matsumoto, I., Murakami, Y., Shinzeki, M., Asari, S., Goto, T., Tani, M., et al. (2015). Proposed preoperative risk factors for early recurrence in patients with resectable pancreatic ductal adenocarcinoma after surgical resection: a multi-center retrospective study. *Pancreatology* 15, 674–680. doi: 10.1016/j.pan.2015.09.008

Ng, F., Ganeshan, B., Kozarski, R., Miles, K. A., and Goh, V. (2013). Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. *Radiology* 266, 177–184. doi: 10.1148/radiol.12120254

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74, 2913–2921. doi: 10.1158/0008-5472.can-14-0155

Rao, S.-X., Lambregts, D. M. J., Schnerr, R. S., van Ommen, W., van Nijnatten, T. J. A., Martens, M. H., et al. (2014). Whole-liver CT texture analysis in colorectal cancer: does the presence of liver metastases affect the texture of the remaining liver? *United Eur. Gastroenterol. J.* 2, 530–538. doi: 10.1177/2050640614552463

Sandrasegaran, K., Lin, Y., Asare-Sawiri, M., Taiyini, T., and Tann, M. (2019). CT texture analysis of pancreatic cancer. *Eur. Radiol.* 29, 1067–1073. doi: 10.1007/s00330-018-5662-1

Schima, W., Ba-Ssalamah, A., Kölblinger, C., Kulinna-Cosentini, C., Puespoek, A., and Götzinger, P. (2007). Pancreatic adenocarcinoma. *Eur. Radiol.* 17, 638–649. doi: 10.1007/s00330-006-0435-7

Siegel, R. L., Miller, K. D., and Jemal, A. (2016). Cancer statistics, 2016. *CA Cancer J. Clin.* 66, 7–30. doi: 10.3322/caac.21332

Wang, Z.-Q., Li, J.-S., Lu, G.-M., Zhang, X.-H., Chen, Z.-Q., and Meng, K. (2003). Correlation of CT enhancement, tumor angiogenesis and pathologic grading of pancreatic carcinoma. *World J. Gastroenterol.* 9, 2100–2104. doi: 10.3748/wjg.v9.i9.2100

Wang, Z.-Q., Lu, G.-M., Chen, Y.-X., Wu, J., Quan, Z.-F., and Li, J.-S. (2007). Value of CT enhancement degree in differential diagnosis between pancreatic carcinoma and inflammatory pancreatic mass. *Zhonghua yi xue za zhi* 87, 1120–1122.

Watanabe, H., Kanematsu, M., Tanaka, K., Osada, S., Tomita, H., Hara, A., et al. (2014). Fibrosis and postoperative fistula of the pancreas: correlation with MR imaging findings–preliminary results. *Radiology* 270, 791–799. doi: 10.1148/radiol.13131194

Wong, K. P., Tsang, J. S., and Lang, B.-H. (2018). Role of surgery in pancreatic neuroendocrine tumor. *Gland Surg.* 7, 36–41. doi: 10.21037/gs.2017.12.05

Conflict of Interest: GY was employed by the company GE Healthcare, Shanghai.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

METHODS
published: 30 June 2020
doi: 10.3389/fbioe.2020.00669

A New Method Based on CEEMD Combined With Iterative Feature Reduction for Aided Diagnosis of Epileptic EEG

Mengran Zhou[1,2], *Kai Bian*[1]*, *Feng Hu*[1] *and Wenhao Lai*[1]

[1] School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, China, [2] State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology, Huainan, China

In the clinical diagnosis of epileptic diseases, the intelligent diagnosis of epileptic electroencephalogram (EEG) signals has become a research focus in the field of brain diseases. In order to solve the problem of time-consuming and easily influenced by human subjective factors, artificial intelligence pattern recognition algorithm has been applied to EEG signals recognition. However, at present, the common empirical mode decomposition (EMD) signal decomposition algorithm does not consider the problem of mode aliasing. The EEG features obtained by feature extraction may be mixed with some unimportant features that affect the classification accuracy. In this paper, we proposed a new method based on complementary ensemble empirical mode decomposition (CEEMD) combined with iterative feature reduction for aided diagnosis of epileptic EEG. First of all, the evaluation indexes of decomposing and reconstructing signals by several methods were compared. The CEEMD was selected as the decomposition method of the signals. Then, the support vector machine recursive elimination (SVM-RFE) was used to reduce 9 features extracted from EEG data. The support vector classification of the gray wolf optimizer (GWO-SVC) recognition model was established for different feature subsets. By comparing the classification accuracy of training set and test set of different feature subsets, and considering the complexity of the model reflected by the number of features selected by SVM-RFE, the analysis showed that the 6 feature subsets with fewer features and higher classification accuracy could reflect the key information of epileptic EEG. The accuracy of the training set classification was 99.38% and the test set was as high as 100%. The recognition time was only 1.6551 s. Finally, in order to verify the reliability of the algorithm proposed in this paper, the proposed algorithm compared with the classification model established by the raw EEG signals and the optimization model established by other intelligent optimization algorithms. It is found that the algorithm used in this paper has higher classification accuracy and faster recognition time than other processing methods. The experimental results show that CEEMD combined with SVM-RFE is feasible for rapid and accurate recognition of EEG signals, which provides a theoretical basis for the aided diagnosis of epilepsy.

Keywords: intelligent diagnosis, EEG signals, complementary ensemble empirical mode decomposition, feature reduction, gray wolf optimizer

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Haibao Wang,
First Affiliated Hospital of Anhui Medical University, China
Huaqing Zhu,
Anhui Medical University, China

****Correspondence:***
Kai Bian
kbian92@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *17 April 2020*
Accepted: *28 May 2020*
Published: *30 June 2020*

Citation:
Zhou M, Bian K, Hu F and Lai W (2020) A New Method Based on CEEMD Combined With Iterative Feature Reduction for Aided Diagnosis of Epileptic EEG. Front. Bioeng. Biotechnol. 8:669. doi: 10.3389/fbioe.2020.00669

INTRODUCTION

Epilepsy is a chronic disease of nervous system disorder caused by abnormal discharge of brain neurons (Sheng et al., 2018). Worldwide, the number of epileptics has exceeded 50 million (Yang et al., 2011). The symptoms of epilepsy patients are usually sudden loss of consciousness, muscle convulsions, etc., which make epilepsy patients have a high mortality rate (Kobow et al., 2012), so their daily life has been greatly troubled. If the epilepsy of seizure type can be accurately identified and classified so that doctors can take reasonable treatment plans, it can help epileptics avoid the risk of disease in advance (Chen E. et al., 2018). Therefore, it is of great significance to strengthen the early diagnosis and late treatment of epilepsy.

The analysis of electroencephalogram (EEG) signals has the characteristics of high efficiency, small damage, and low cost. It has become the main clinical diagnosis method of epilepsy. This method needs experienced doctors to observe the high amplitude synchronous rhythms such as sharp wave and spike-wave in EEG during the epileptic seizure for a long time with the naked eye (Lévesque et al., 2017), which will not only consume a lot of energy but also may get wrong diagnosis results due to various uncertain factors. Therefore, it is necessary to develop a method of automatic recognition of epileptic EEG. In recent years, machine learning and deep learning algorithms have been widely applied in the biomedical and health field (Wang et al., 2019, 2020; Deng et al., 2020; Hu et al., 2020). Artificial intelligence combined with EEG has achieved good results in the diagnosis and prediction of epilepsy and other diseases. For example, Bajaj and Pachori (2013) used empirical mode decomposition (EMD) to decompose EEG signals and improved the classification accuracy of epilepsy detection by analyzing the first three natural mode function components. Puspita et al. (2017) extracted the mean, standard deviation and median statistical features of EEG data, and then used the back-propagation neural network (BNN) to establish the classification and recognition model of EEG data of epilepsy patients and achieved the best classification results. Cao et al. (2017) combined the short-time Fourier transform (STFT) with a convolutional neural network (CNN) and used the deep learning algorithm to avoid the process of manual feature selection in EEG recognition. The analysis steps of EEG mainly include preprocessing of raw signals, feature extraction, recognition, and classification. EMD is often used as the decomposition method of EEG signals. However, only one or some IMF components selected by subjective experience are taken as the research object, which cannot completely contain the useful information of the original signals, so the accuracy of EEG obtained is low, and it cannot effectively identify different types of EEG. Several typical EEG feature indexes are extracted directly for classification and recognition. This method cannot judge whether the extracted EEG features are all effective EEG feature indexes, which not only increases the recognition time but also affects the accuracy of classification.

Complementary ensemble empirical mode decomposition (CEEMD) is a signal decomposition method developed on the basis of empirical mode decomposition (EMD) (Muñoz-Gutiérrez et al., 2018), it has obvious advantages in dealing with non-linear and non-stationary signals. Satija et al. (2017) used a modified CEEMD algorithm to achieve automatic detection and classification of ECG noise. Chen and Hsiao (2018) used the CEEMD method to extract hidden signals from the respiratory inductance plethysmography (RIP) signals based on the frequency bands of different respiratory muscles. Amezquita-Sanchez et al. (2016) combined CEEMD with magnetoencephalography (MEG) to distinguish patients with mild cognitive impairment (MCI). Support vector machine recursive feature elimination (SVM-RFE) is a feature selection method, it can eliminate the feature information of low importance, and effectively remove the interference of redundant information (Tapia et al., 2012), which is conducive to the establishment of the classification model. SVM-RFE has been widely used in biomedical research. Ding et al. (2015) proposed a method of SVM-RFE combined with voxel-based morphometry (VBM) to analyze MRI data and realized the automatic classification of smokers and non-smokers. Anaissi et al. (2016) used the ensemble SVM-RFE algorithm to select the characteristic genes in the genomic data. Bisdas et al. (2018) adopted the SVM-RFE method to select the most discriminative diagnostic biomarkers. Gray wolf optimizer (GWO) is a new swarm intelligent optimization algorithm (Yamany et al., 2015). It can improve the performance of the SVM training model and has the advantages of simplicity and efficiency. Ramakrishnan and Sankaragomathi (2017) used the modified region growing (MRG) and GWO to achieve the accurate segmentation of CT brain tumor images. Shankar et al. (2018) proposed an improved GWO to optimize the performance of multi-kernel SVM for thyroid disease classification.

In this paper, CEEMD was used to decompose the raw epileptic EEG signals into natural mode functions (IMF) of different frequencies, then these component signals were reconstructed and their linear and non-linear features were extracted. SVM-RFE was used to eliminate non-key features and reduce the influence of redundant features on recognition accuracy. Finally, the GWO-SVC classification model based on GWO optimized support vector classification (SVC) algorithm was applied to classify the EEG signals, which provided a theoretical basis for the aided diagnosis of epilepsy.

MATERIALS AND METHODS

Selection of Experimental Data

The experimental data in this paper were from the EEG database of the epilepsy research center of the University of Bonn, Germany (Andrzejak et al., 2001). The sampling frequency of EEG signal acquisition system was 173.61 Hz, and the range of filtering bandwidth was 0.53–40 Hz. EEG data have been preprocessed to remove the artifacts and the data were widely used in public, so the experimental results have high reliability and contrast. The data set consists of five data subsets (denoted A–E), each of which contains 100 single-channel signals with a time of 23.6 s, and each single-channel signal contains 4,097 sampling points, and the bit of A/D conversion is 12 bits. The band-pass filter with a bandwidth of 0.53–40 Hz was used for filtering. Subsets A and B were EEG signals from the scalp surface of 5 healthy volunteers when they opened and closed their eyes, respectively. Subset C was the EEG signals of the hippocampal formations in five epileptic patients. Subset D was the EEG signal of the epileptogenic area with interictal epilepsy. Subset E was the EEG signal of the epileptogenic area during the ictal epilepsy.

The hardware condition of the computer used in the experiment was the Intel Core i7 processor, 4GB memory, win7 system. Under the environment of MATLAB r2016b (MathWorks, USA), the algorithm was used to simulate and test the data. The support vector machine chose the libsvm-mat-3.1 toolkit (Chang and Lin, 2011) to run.

Complementary Ensemble Empirical Mode Decomposition

CEEMD is an improved signal decomposition method for EEMD proposed by Yeh et al. (2010). This method not only solves the problems of residual white noise and complex processing in EEMD (Wu and Huang, 2009) decomposition but also effectively suppresses the modal aliasing in the EMD decomposition method (Wu and Huang, 2010). The decomposition process of the CEEMD algorithm is based on EMD, adding a pair of auxiliary white noise with the same amplitude and opposite sign to the raw signals. These raw signals are decomposed into several intrinsic mode functions (IMFs) and residuals with clearer physical meaning. As the number of added noise increases, the residual amount of noise in reconstruction data will decrease, and the final residual amount can be almost ignored (Chen D. et al., 2018).

The decomposition steps of CEEMD are as follows:

Step 1: A pair of random Gaussian white noises with the same amplitude and opposite signs are added to the signal to form two new decomposition signals.

$$\begin{cases} S_{+i}(t) = S(t) + N_i^{+}(t) \\ S_{-i}(t) = S(t) + N_i^{-}(t) \end{cases} \tag{1}$$

Where $S(t)$ is the raw signal, $N_i(t)$ is the white noise added for the i time, $S_{+i}(t)$ is the signal obtained by adding the positive white noise for the i time, and $S_{-i}(t)$ is the signal obtained by adding the negative white noise for the i time. Generally, the value is 0.01–0.5 times of the standard deviation of the original signal.

Step 2: EMD algorithm is used to decompose $S_{+i}(t)$ and $S_{-i}(t)$ to get their IMF components and residual terms.

$$\begin{cases} S_{+i}(t) = \sum\limits_{j=1}^{m} I_{+ij}(t) + R_{+i}(t) \\ S_{-i}(t) = \sum\limits_{j=1}^{m} I_{-ij}(t) + R_{-i}(t) \end{cases} \tag{2}$$

Where $I_{+ij}(t)$ denotes the j IMF component from $S_{+i}(t)$ decomposition, $I_{-ij}(t)$ denotes the j IMF component from $S_{-i}(t)$ decomposition, $R_{+i}(t)$ and $R_{-i}(t)$ denote the corresponding residual terms, respectively.

Step 3: Step 1 and *step 2* are repeated for m times, and random white noise is added each time until the residual terms can no longer be decomposed.

Step 4: Calculate the mean value of IMF components obtained by decomposition, and take the mean value as the result of IMF component.

$$C_j(t) = \frac{1}{2m} \sum_{i=1}^{m} \left(I_{+ij}(t) + I_{-ij}(t)\right) \tag{3}$$

where $C_j(t)$ denotes the first IMF component obtained by CEEMD.

Support Vector Machine Recursive Feature Elimination

Support vector machine recursive feature elimination (SVM-RFE) is a feature selection method based on feature sorting technology proposed by Guyon et al. (2002). The function of RFE is to rank features by greedy strategy. Starting from the complete set, the least relevant features are eliminated one by one to complete the backward feature reduction, and finally, the optimal feature subset is obtained. SVM-RFE is a combination of SVM and RFE. In the process of SVM training, the weight of features can reflect their contribution to classification decision-making. Therefore, the weight of a classifier can be used as the basis of feature ranking, and then the relatively unimportant features are deleted one by one according to the weight of classifier until a certain number of features with higher importance are left. The combination of the SVM classification algorithm and feature selection process can improve the effectiveness of feature selection.

The steps of iterative reduction feature of SVM-RFE method are as follows:

Step 1: Input training sample data $D = \{d_1, d_2, ..., d_3\}^T$ and category label $L = \{l_1, l_2, ..., l_n\}^T$

Step 2: Initialize feature set $\alpha = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ and rearrange feature set $\beta = \{\}$

Step 3: The SVM classifier is used to train the input data, and the parameter information of the support vector is $\delta = SVMtrain(D, L)$

Step 4: Calculate the cost function of features

$$f(x) = \frac{1}{2} D^T U(x) - \frac{1}{2} D^T U(-x) \tag{4}$$

Where $U(x)$ is a matrix with element $a_i a_j K(x_i, x_j)$, $U(-x)$ is the matrix after eliminating x features, and K denotes the kernel function of correlation between x_i and x_j

Step 5: The weight coefficient w is used as the ranking criterion of feature importance to reorder new features. Get a new feature order set $\beta = \{\beta_1, \beta_2, ..., \beta_n\}$, and remove the feature with the smallest weight coefficient from the current order set, repeat Step 3–Step 5, until enough features are deleted

Step 6: A set of nested feature subsets $Z_1 \subset Z_2 \cdots Z_n$ is defined, $Z_i (i = 1, 2, \cdots, n)$ represents a subset of the top most important features selected from the feature set, and uses the recognition rate of the classifier as the evaluation index to select the best subset.

Gray Wolf Optimizer Combine With Support Vector Classification

Gray wolf optimizer (GWO) is an advanced heuristic group intelligent optimization algorithm proposed by Mirjalili et al. (2014). This algorithm is mainly an optimized search method which simulates the social hierarchy of gray wolf and the way of preying on its prey. It has strong convergence performance, few parameters and easy to realize, and so on. SVM is originally a two classification model and can be used to solve multi-classification problems. It is a linear classifier with the largest interval defined in the feature space, which makes it different from the perceptron (Utkin et al., 2016). The learning strategy of SVM is to maximize the interval. SVM is a non-linear classifier in essence. SVM algorithm can be used for pattern classification or nonlinear regression, and SVC is the algorithm used by SVM to solve classification problems (Chen et al., 2010). The classification performance of the SVC model is affected by the penalty coefficient c and kernel function parameter g. Through the GWO algorithm, the SVC parameters are optimized to find the best classification parameters c and g, so as to obtain the GWO-SVC model with good performance.

The specific parameter optimization steps are as follows:

Step 1: α, β, and γ are three different classes of primitive wolves with the same scale generated from feasible region $W = \{w_1, w_2, \cdots, w_n\}$
Step 2: Initialize the position of the original wolves, obtain the fitness μ of gray wolf individuals in the population, and define the optimal and suboptimal fitness as c and g, respectively
Step 3: Select the fitness of the top three, and set the corresponding gray wolf to α, β, and γ in order
Step 4: Constantly move the position of gray wolf when it preys on prey and updates the subordinate wolves. The updating formula is as follows:

$$\begin{cases} Q_\alpha = \left|W(t) - H_1 W_\alpha\right| \\ Q_\beta = \left|W(t) - H_2 W_\beta\right| \\ Q_\gamma = \left|W(t) - H_3 W_\gamma\right| \end{cases} \tag{5}$$

$$\begin{cases} W_1 = W_\alpha - K_1 Q_\alpha \\ W_2 = W_\beta - K_2 Q_\beta \\ W_3 = W_\gamma - K_3 Q_\gamma \end{cases} \tag{6}$$

$$W(t+1) = \tfrac{1}{3}(W_1 + W_2 + W_3) \tag{7}$$

Where W_α, W_β, and W_γ denote the location of the gray wolf, and H_1, H_2, H_3, K_1, K_2, and K_3 are scale factors
Step 5: Update the values of α, H, and K. If the constraints are not met, go to *step 2*
Step 6: Use output parameters c and g to build SVC model for classification and recognition.

Evaluation Index

The effect of a signal processing method is determined by the comparison of some digital evaluation indexes, such as pearson correlation coefficient (Pr), signal to noise ratio (SNR), and mean absolute error (MAE) (Ou-Yang et al., 2012). Generally, the larger the Pr value is, the greater the linear correlation between signals is. The larger the SNR value is, the more useful the restored signal is and the less the distortion is. The smaller the MAE value is, the better the effect of signal filtering is.

The expression of the Pr:

$$Pr = \frac{\sum_{i=1}^{m} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{m} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{m} (Y_i - \overline{Y})^2}} \tag{8}$$

The expression of the SNR:

$$SNR = 10\log_{10} \frac{\sum_{i=1}^{m} X_i^2}{\sqrt{\sum_{i=1}^{m} (X_i - Y_i)^2}} \tag{9}$$

The expression of the MAE:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |X_i - Y_i| \tag{10}$$

Where X_i is the original signal, Y_i is the processed signal, $\overline{X}$ is the average value of the signal, and $\overline{Y}$ is the standard deviation of the signal.

Feature Extraction

Because the information contained in EEG is usually recessive, it is difficult to find all the rules through observation, so it is necessary to extract the features of EEG data. Because of its unique characteristics, EEG is different from other physiological signals, and the characteristics of different EEG are also different. The purpose of EEG feature extraction is to extract relatively effective feature indexes from many EEG features. At present, there are many EEG characteristics studied in the literature, such as mean, variance, standard deviation, range, fluctuation coefficient (Yuan et al., 2012), variation coefficient (Vinton et al., 2004), sample entropy (Arunkumar et al., 2018), kurtosis (Javidi et al., 2011) and skewness (Gandhi et al., 2012). In this paper, we extracted the above nine features from EEG signals for analysis.

RESULTS AND ANALYSIS

Analysis of EEG Signal of Primary Epilepsy

One single-channel signal is selected from subset D with interictal epilepsy and subset E with ictal epilepsy for waveform analysis. The raw epileptic EEG signal is shown in **Figure 1**, and the single-channel signal contains 4,097 sampling points. **Figure 1A** shows the EEG signal during the interictal epilepsy. The waveform of the signal is relatively stable with little fluctuation. The amplitude range is −252∼123 μV. **Figure 1B** is the EEG signal during the ictal epilepsy, which fluctuates violently and has regularity. The amplitude range is −890∼1,367 μV. The amplitude of EEG in the ictal period is obviously larger than that in the interval

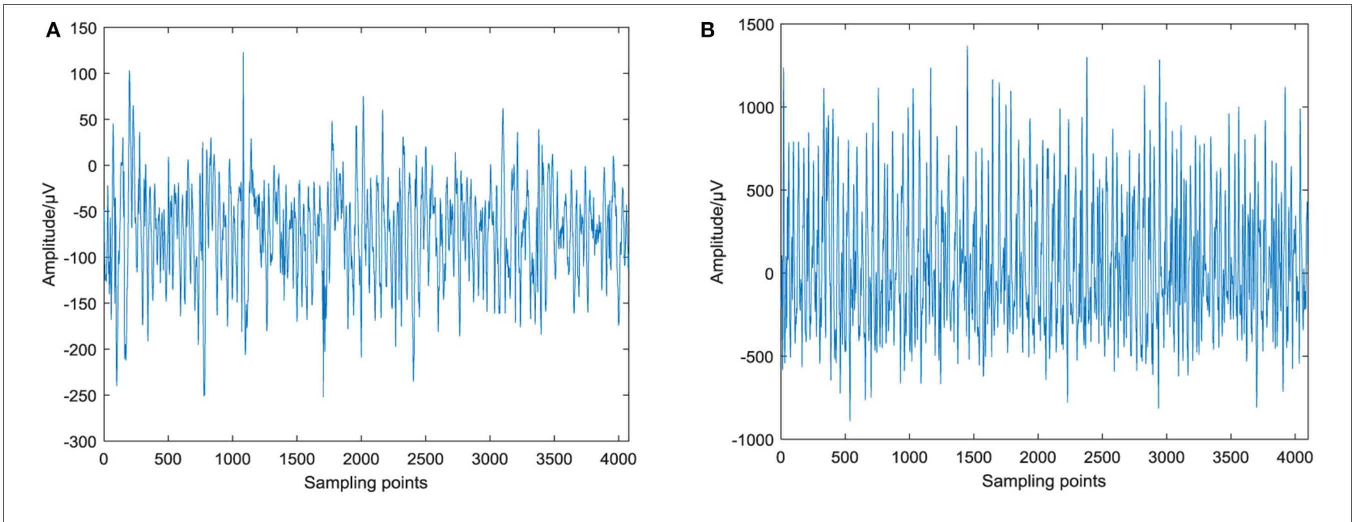


FIGURE 1 | Raw EEG signal of epilepsy. **(A)** EEG signal of interictal epilepsy and **(B)** EEG signal of ictal epilepsy.

period, and the fluctuation gap is obvious, which indicates that the signal is excited and fluctuates violently in the ictal period. This phenomenon is consistent with the state of EEG activity with ictal epilepsy.

CEEMD Based on Signal Evaluation Index

EMD and CEEMD are used to decompose epileptic EEG signals, and the Intrinsic Mode Function (IMF) components of each order are obtained. Based on the MATLAB platform, the standard deviation of the added white noise is set to 0.2 times of the raw signal of the standard deviation. The number of iterations is set to 100, and the number of IMF is set to 9 (not including the trend). The signal decomposition of EEG between interictal and ictal period are shown in **Figure 2**. The raw EEG signal is decomposed into nine IMF and one residual term. The decomposed IMF components are arranged in the order of frequency from high to low, and each component has its own amplitude and frequency. With the increase of IMF component orders, the more stable the signal changes, the smaller the corresponding energy. The signal changes during the ictal period are more intense than during the interictal period. The amplitudes of the first four orders are larger than those of other orders. It can be seen from **Figures 2A,C** that the amplitude of IMF in each stage of ictal EEG signal processed by EMD is larger than that of the interictal EMD, and the difference is obvious. High-frequency signals with small amplitude appear in some sampling points of the first three IMF components, that is to say, there are different degrees of mode aliasing, which is more obvious in the ictal period. However, it can be seen from **Figures 2B,D** that there is no small-amplitude and high-frequency signals in the first three stages of EEG signals and seizure signals processed by CEEMD, which indicates that CEEMD can solve the problem of mode aliasing caused by EMD decomposition. There are great differences in amplitude and frequency between interictal EEG and ictal EEG.

EMD, EEMD, and CEEMD are used to decompose the IMF component of the ictal period signals and conduct correlation analysis with the original signals, as shown in **Figure 3**. It can be concluded from the correlation property that the *Pr* of IMF2 and IMF3 decomposed by EMD is >0.5, which shows a strong correlation. The *Pr* of IMF2 reaches a maximum value of 0.6932, followed by a decreasing trend of IMF. The results of EEMD and CEEMD show that the *Pr* of IMF2, IMF3, and IMF4 are more than 0.5, which shows a strong correlation. The *Pr* of the two decomposition methods reach the maximum at IMF3, and their values are 0.8316 and 0.8300, respectively. The *Pr* of the latter IMF shows a decreasing trend. In addition to the first two IMF components, the *Pr* of the remaining eight IMF components decomposed by EMD are smaller than the *Pr* of the corresponding components decomposed by EEMD and CEEMD. The evaluation indexes of 10 IMF decomposed by different decomposition methods are shown in **Table 1**. The difference between the average *Pr* of the IMF decomposed by CEEMD and EEMD is very small and larger than that of EMD. The average *Pr* of EMD and EEMD is close, and both are smaller than the CEEMD decomposition method. CEEMD's average *MAE* is also smaller than the other two signal decomposition methods. In general, the CEEMD has relatively good signal evaluation indexes. However, from the signal evaluation index, it can be seen that the average *Pr* of different IMF decomposed by three methods is between 0.1 and 0.3, which shows weak correlation, indicating that a single IMF cannot represent all the information of the raw EEG signals. We need to select some useful IMF components for signal reconstruction in order to avoid the influence of distorted signals on the subsequent EEG recognition.

Generally, it is considered that *Pr* has no correlation in the range of 0–0.09. The threshold value is set to 0.1, IMF components below the threshold value are deleted, and the components above the threshold value are reserved for signal reconstruction. As can be seen from **Figure 3**, the IMF1–IMF4

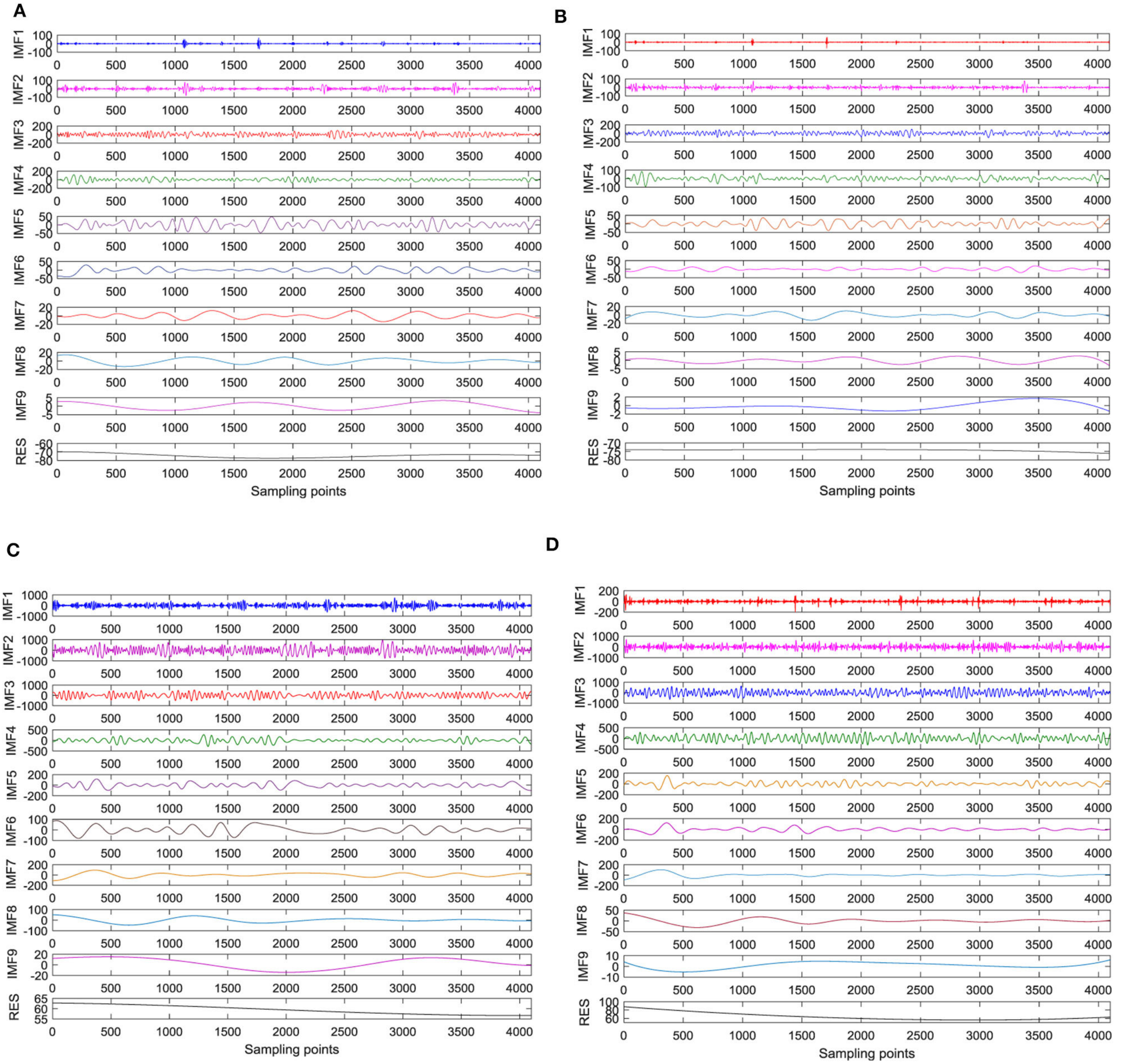


FIGURE 2 | The signal decomposition of EEG between interictal and ictal period. **(A)** EMD decomposition during the interictal period, **(B)** CEEMD decomposition during the interictal period, **(C)** EMD decomposition during the ictal period, and **(D)** CEEMD decomposition during the ictal period.

components decomposed by EMD, the *Pr* of the IMF1–IMF6 components decomposed by EEMD and CEEMD are all higher than 0.1. We select these IMF components to reconstruct the EEG signals. The evaluation indexes reconstructed by different decomposition methods are shown in **Table 2**. After reconstruction, the evaluation indexes of EEG signals are better than those of a single IMF component signal. The *Pr* of reconstructed signals and raw signals are all >0.9, showing a strong correlation, which shows that signal reconstruction is a necessary job. In conclusion, CEEMD is better than the other two methods in decomposing and reconstructing the signals, and CEEMD is chosen as the preprocessing method of the raw EEG signals.

The above simulation experiment is to analyze the correlation of one channel of epileptic EEG during the ictal period, and the next is to analyze the correlation of two kinds of EEG signals during the interictal and ictal period. Each type of signal has 100 channels. Here, one channel is selected from the two types of signals for further correlation analysis. As shown in **Figure 4**, the maximum correlation component of EEG signals during the interictal period is IMF4, the maximum correlation IMF component of EEG signals during the ictal period is IMF2, and

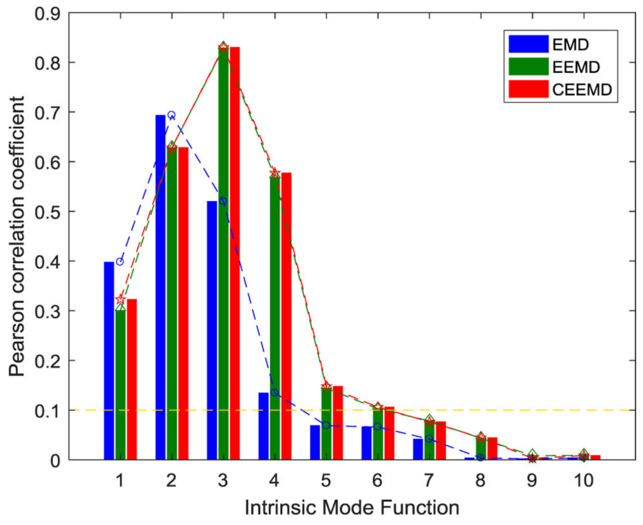


FIGURE 3 | Correlation between IMF and raw signals in different stages of ictal period signals.

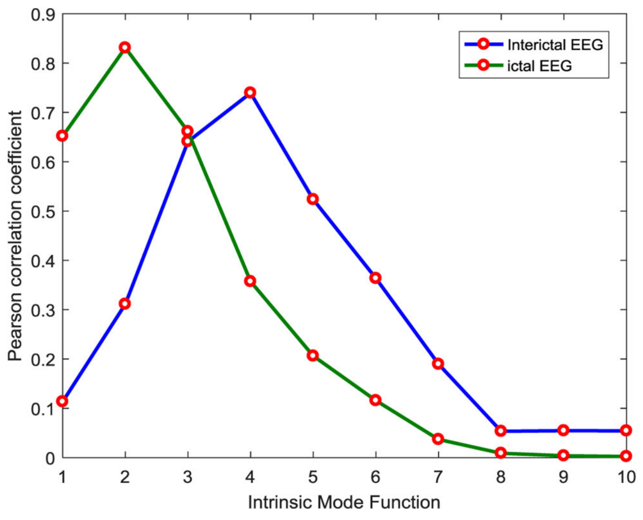


FIGURE 4 | Correlation between IMF components of different channels and raw signals.

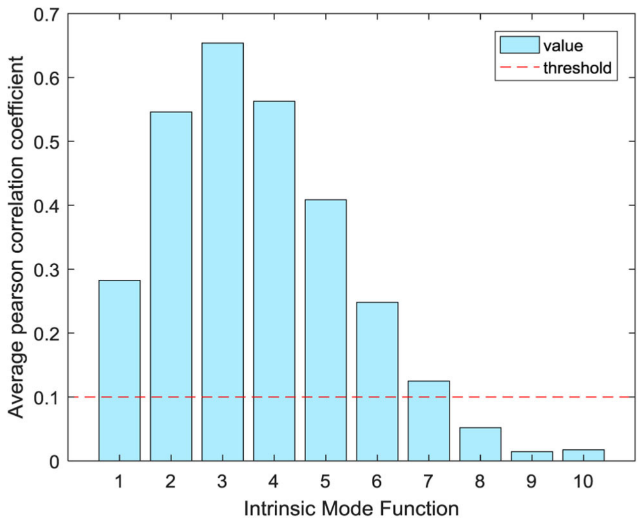


FIGURE 5 | Correlation between IMF components of all channels and raw signals.

TABLE 1 | Evaluation indexes of different decomposition methods.

Method	Average *Pr*	Average *SNR*	Average *MAE*
EMD	0.1964	0.4946	296.4282
EEMD	0.2701	0.4233	296.5334
CEEMD	**0.2745**	**0.7692**	**287.3643**

Bold values indicate the best experimental results more intuitively.

TABLE 2 | Evaluation indexes reconstructed by different decomposition methods.

Method	*Pr*	*SNR*	*MAE*
EMD	0.9856	12.4403	72.8892
EEMD	0.9954	5.5925	158.0966
CEEMD	**0.9959**	**14.3365**	**67.1919**

Bold values indicate the best experimental results more intuitively.

the maximum correlation IMF components of different types of signals are different. The EEG samples of 200 channels are decomposed by the CEEMD method, and the average *Pr* of IMF components is used as the division basis of useful signals. The threshold value is set to 0.1. As shown in **Figure 5**, the average *Pr* of IMF1–IMF7 components is higher than 0.1. Finally, we select these seven IMF components to reconstruct all EEG signals.

Feature Extraction of EEG Signals

The feature extraction of 200 single-channel signals reconstructed from the interictal and ictal period is carried out. The extracted 9 features, namely mean, variance, standard deviation, range, fluctuation coefficient, variation coefficient, sample entropy, kurtosis, and skewness will be used in the next iterative feature reduction analysis. In the extraction of sample entropy, $m = 2$, $r = 0.2$ std. Because 9 features will produce many combinations of different feature subsets, it will lead to low training efficiency and model performance degradation. Therefore, the SVM-RFE algorithm should be used to rank the epileptic EEG data according to the weight of feature importance and select the combination of the optimal features.

Reduction of Secondary Features and Establishment of Classification Models

When SVM-RFE is used to reduce the secondary features of data, it is necessary to normalize the data to [0,1] interval first to avoid the adverse effect of a too large difference between different features of data on the experimental results. Gaussian radial basis function (RBF) is used as a kernel function of SVM. The weight values of different features are shown in **Figure 6**. The sequence numbers 1–9 correspond to the nine features extracted from

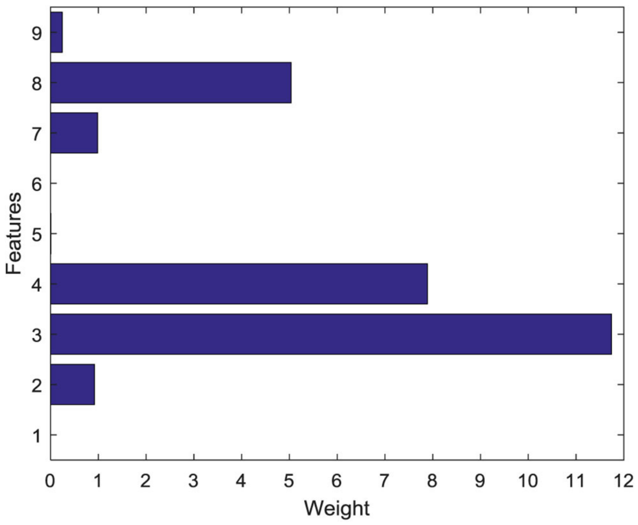


FIGURE 6 | Weight values of different features.

TABLE 3 | Classification accuracy of different feature subsets.

Feature subset	Feature numbers	Best c	Best g	Training set/%	Test set/%
{3,4,8,2,7,9,5,1,6}	9	60.1100	6.9761	100 (160/160)	97.5 (39/40)
{3,4,8,2,7,9,5,1}	8	33.9487	8.6581	99.38 (159/160)	100 (40/40)
{3,4,8,2,7,9,5}	7	70.2355	5.2786	99.38 (159/160)	100 (40/40)
{3,4,8,2,7,9}	**6**	79.1905	7.7580	**99.38 (159/160)**	**100 (40/40)**
{3,4,8,2,7}	5	72.8569	9.2752	98.75 (158/160)	97.5 (39/40)
{3,4,8,2}	4	1.1068	70.1738	98.13 (157/160)	97.5 (39/40)
{3,4,8}	3	15.0708	12.2651	97.5 (156/160)	97.5 (39/40)
{3,4}	2	9.8490	88.9158	96.25 (154/160)	95 (38/40)
{3}	1	24.4377	39.4796	93.13 (149/160)	95 (38/40)

Bold values indicate the best experimental results more intuitively.

the EEG signals, respectively. This figure fully reflects that there are obvious differences in the importance of each feature of the EEG signals. It can be seen that the weight value of the standard deviation feature is the largest, indicating that the feature covers a lot of useful information on the EEG data. The weight of mean value, fluctuation coefficient, and variation coefficient is very small, which shows that the importance of these three characteristics is relatively low. According to the weight values of different features, the new features are sorted as {3,4,8,2,7,9,5,1,6}.

Because the first feature is the last one to be eliminated, it is also the most important feature. Therefore, based on all feature combinations in the new feature sorting, the features with the lowest importance in the current feature set are eliminated one by one feature at a time, and the number of features is reduced iteratively until it is reduced to the most important standard deviation feature. There are nine different feature sets. 80% (160) of 200 epileptic EEG signals are divided into training sets and the remaining 20% (40) into test sets. The data of different feature combinations in EEG signals are input to the GWO-SVC classification models in turn. The accuracy of the training set and test set obtained by a training classifier is used as the evaluation index of secondary feature reduction to select the optimal subset. In order to ensure the accuracy of classification results and the efficiency of recognition process at the same time, the initial number of the gray wolf is set to 20, the maximum number of iterations is set to 50, and the search interval of penalty coefficient and kernel function parameter is [0,100].

The classification accuracy of different feature subsets is shown in **Table 3**. The accuracy of the training set is on the decline, while the accuracy of the test set is on the rise and then on the decline. When the number of features in the feature subset is reduced from 9 to 8, the accuracy of the test set reaches the maximum of 100% for the first time, and only one channel EEG signal in the training set is misclassified. When the number of features is reduced to 6, the accuracy of the training set and the test set begins to decline. There are eight iterations until there is only one feature left. The purpose of secondary feature reduction is to improve the classification accuracy by filtering features or to reduce the dimension of feature set without reducing the classification accuracy. Although the accuracy of the training set of the full feature set is 100%, there are EEG signals in the test set which are misclassified, and the number of features is the most, which results in the low efficiency of model training. Finally, the subset {3,4,8,2,7,9} of six features with fewer features and higher classification accuracy is selected as the result of the SVM-RFE algorithm.

Based on the 9 features extracted from the raw EEG data, the SVC model without parameter optimization is established, and RBF is chosen as the kernel function. In libsvm-mat-3.1 toolkit, the default value of a penalty coefficient c is 1, and the default value of a kernel function parameter g is the reciprocal of feature number (1/features). In order to clearly express the difference between the test category and the actual category, the blue "○" in the figure is the actual category of the input sample, and the red "∗" is the predicted result of the classification model. If "○" and "∗" coincide, the sample is correctly classified. The classification results of the raw EEG signals by SVC are shown in **Figure 7**. In the training set, 23 EEG signals were identified incorrectly, including three EEG signals in the interictal period and 20 EEG signals in the ictal period. A total of four EEG signals in the test set were identified incorrectly, and they were all EEG signals during the ictal period. The EEG signals processed by CEEMD are classified by GWO-SVC as shown in **Figure 8**. Only one EEG signal in the training set is identified incorrectly, which was the 73rd EEG signal in the interictal period. All the EEG signals in the test set are correctly identified. It can be seen that the training and test set of the GWO-SVC model established by the EEG signal after CEEMD processing has significantly better recognition results than the SVC classification model established by the unprocessed raw EEG signals. It shows that the method in this paper is applicable to the aided diagnosis of epileptic EEG, and it realizes the precise identification of EEG signals.

Comparison With Other Methods

In order to verify the classification effect and superiority of the proposed method for epilepsy EEG recognition, the algorithm

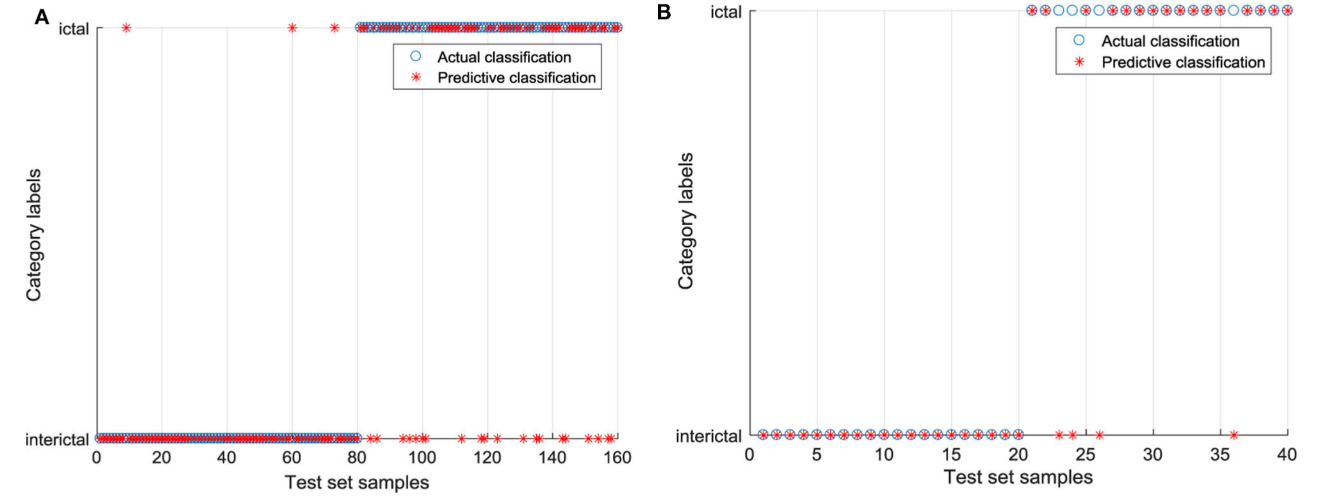


FIGURE 7 | The classification results of the raw EEG signals by SVC. **(A)** Classification diagram of training set and **(B)** Classification diagram of test set.

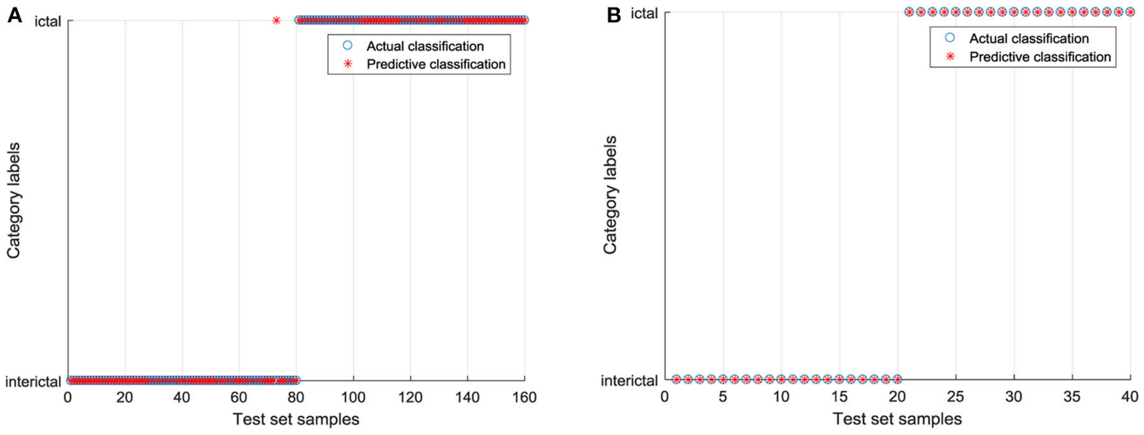


FIGURE 8 | The EEG signals processed by CEEMD are classified by GWO-SVC. **(A)** Classification diagram of training set and **(B)** Classification diagram of test set.

in this paper not only performs longitudinal comparative analysis and research with the SVC classification results of the unoptimized parameters of the raw EEG data but also compares with the classification results of grid search (GS), genetic algorithm (GA), particle swarm optimization (PSO), artificial bee colony (ABC), cuckoo search (CS), and firefly algorithm (FA) intelligent optimization algorithms. Other classifiers are similar to the GWO algorithm. The number of initial population is set to 20, the maximum number of iterations is set to 50, the search interval of penalty coefficient and kernel function parameters is [0,100], and the EEG data are normalized to [0,1] interval. Through such work, the unity of initial conditions can be ensured. **Table 4** shows the classification results of different processing methods. It can be seen that the number of features selected by the model without parameter optimization and parameter optimization is different. The modeling time of SVC without parameter optimization is short, but the accuracy of the training set is low. It takes less time to establish the SVC model without parameter optimization, but the accuracy of the training set is low. CEEMD has little effect on the accuracy of the SVC model without parameter optimization. The accuracy of the training set and test set of GWO-SVC model is significantly higher than that of SVC. Compared with the raw EEG signals, the training set and test set accuracy of the model is improved after the signal is processed by CEEMD and SVM-RFE. Compared with the SVC model based on the raw EEG signals, the accuracy of training set classification and test set classification of the optimization model based on the algorithm in this paper is improved by 13.755 and 10%, respectively.

The classification results of different optimization algorithms are shown in **Table 5**. The training set classification accuracy of the GS algorithm optimization model is the lowest, and the recognition time is long. Although the FA algorithm can make the classification accuracy of the training set reach 100%, the

TABLE 4 | Classification results of different treatment methods.

Processing method	Feature numbers	Best c	Best g	Training set/%	Test set/%	Time (s)
Raw+SVC	9	1	0.1111	85.625 (137/160)	90 (36/40)	0.0739
Raw+GWO-SVC	9	47.9615	34.1838	100 (160/160)	95 (38/40)	2.6507
Raw+SVM-RFE+SVC	5	1	0.2	86.875 (139/160)	95 (38/40)	0.0731
CEEMD+SVC	9	1	0.1111	85.625 (137/160)	90 (36/40)	0.0503
CEEMD+GWO-SVC	9	60.1100	6.9761	100 (160/160)	97.5 (39/40)	1.7907
CEEMD+SVM-RFE+ SVC	5	1	0.2	87.5 (140/160)	95 (38/40)	0.0443
CEEMD+SVM-RFE+GWO-SVC	**6**	79.1905	7.758	99.38 (159/160)	**100 (40/40)**	**1.6551**

Bold values indicate the best experimental results more intuitively.

TABLE 5 | Classification results of different optimization algorithms.

Modeling method	Feature numbers	Best c	Best g	Training set/%	Test set/%	Time (s)
GS-SVC	6	5.6569	8	97.5 (156/160)	97.5 (39/40)	4.0977
GA-SVC	6	43.9056	3.1953	98.75 (158/160)	97.5 (39/40)	3.2152
PSO-SVC	6	5.3156	8.7895	98.13 (157/160)	97.5 (39/40)	4.4063
ABC-SVC	6	85.5963	6.0455	99.38 (159/160)	100 (40/40)	3.5328
CS-SVC	6	72.2167	8.4386	99.38 (159/160)	100 (40/40)	3.1746
FA-SVC	6	82.2227	20.4157	100 (160/160)	97.5 (39/40)	1.8642
GWO-SVC	6	79.1905	7.758	99.38 (159/160)	**100 (40/40)**	**1.6551**

Bold values indicate the best experimental results more intuitively.

classification accuracy of the test set is less than GWO and ABC, and the recognition time is longer than GWO. The accuracy of the test set of GWO, ABC, and CS algorithm is 100%, and all EEG signals are recognized correctly. However, the recognition time of the GWO-SVC model is only 1.6551 s, which is obviously faster than that of ABC-SVC, and CS-SVC model, and 2.7512 s faster than PSO-SVC model which has the slowest recognition speed. Compared with other heuristic intelligent optimization algorithms, the GWO algorithm is more effective and reliable in parameter optimization of the SVC model, where c is 79.1905, $g = 7.758$.

DISCUSSION AND CONCLUSIONS

In this study, we have proposed a new method based on CEEMD combined with iterative feature elimination for EEG of epilepsy aided diagnosis. The CEEMD signal decomposition algorithm was used to decompose the raw EEG signals into the IMF of different orders, and then feature extraction is carried out for the reconstructed signals. The SVM-RFE algorithm was used to reduce secondary features. Finally, the GWO-SVC classification and recognition model was established to realize the accurate and fast identification of Epileptic EEG. From the experimental analysis process and results, we can see that:

(1) CEEMD algorithm based on correlation analysis can make the non-stationary EEG data stable, decompose the complex EEG signals into IMF components with practical physical significance, and solve the problems of mode aliasing. This algorithm is superior to the traditional EMD algorithm in various evaluation indexes.
(2) SVM-RFE is used to filter the features of EEG signals, which can reduce the redundant information acquisition in the EEG data that has no internal relationship with the classification. The useful information of epileptic EEG signals is reflected by fewer features. The complexity of a training model is reduced, and the recognition efficiency and reliability of the classification model are improved.
(3) The normalized data get rid of the influence of the big difference of sample data, speed up the optimal solution process, and improve the classification accuracy. The GWO-SVC epileptic EEG recognition model has a good classification accuracy. Combining CEEMD and SVM-RFE algorithm, it can make the classification accuracy higher than the recognition model of all features, and improve the performance and generalization ability of the model.
(4) The algorithm in this paper can be applied to the aided diagnosis of epileptic EEG. This method can accurately and quickly identify the types of epileptic seizures. It has a certain theoretical guidance and promotion value for doctors to achieve the early diagnosis of epileptic diseases and take a reasonable epileptic treatment plan in the later stage.

The EEG data of epilepsy in the experiment were collected in the laboratory. The collection conditions are better than the actual clinical diagnosis conditions, and the interference is relatively small, but there may be many uncertain factors in the actual EEG analysis. In this study, 200 groups of sample data were tested and analyzed, but the actual clinical diagnosis needs to analyze a large number of data, which brings many challenges to the auxiliary diagnosis of Epilepsy EEG. The results show that although the method proposed in this paper has achieved high recognition accuracy, there are still wrong samples. How to overcome these difficulties will become the focus of the next research, and also the key to improving the recognition rate of epilepsy. We are going to add the disadvantageous factors in the experimental analysis to the future research work, expand the sample size of training data, and constantly improve and optimize the intelligent analysis algorithm to achieve perfect recognition accuracy.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://epileptologie-bonn.de/cms/front_content.php?idcat=193&lang=3&changelang=3.

AUTHOR CONTRIBUTIONS

MZ conceived the study and supervised the study. MZ and KB developed the method and wrote the manuscript. KB and FH implemented the algorithms. KB and WL analyzed the data. All authors read and approved the final manuscript and content of the work.

FUNDING

This work was supported by the major science and technology program of anhui province (Grant No. 201903a07020013).

ACKNOWLEDGMENTS

We would like to thank the department of epileptology at the University of Bonn, Germany for making the database available.

REFERENCES

Amezquita-Sanchez, J. P., Adeli, A., and Adeli, H. (2016). A new methodology for automated diagnosis of mild cognitive impairment (MCI) using magnetoencephalography (meg). *Behav. Brain Res.* 305, 174–180. doi: 10.1016/j.bbr.2016.02.035

Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A., and Kennedy, P. J. (2016). Ensemble feature learning of genomic data using support vector machine. *PLoS ONE* 11:e0157330. doi: 10.1371/journal.pone.0157330

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys. Rev. E* 64:061907. doi: 10.1103/PhysRevE.64.061907

Arunkumar, N., Kumar, K. R., and Venkataraman, V. (2018). Entropy features for focal EEG and non focal EEG. *J. Comput. Sci.* 27, 440–444. doi: 10.1016/j.jocs.2018.02.002

Bajaj, V., and Pachori, R. B. (2013). Epileptic seizure detection based on the instantaneous area of analytic intrinsic mode functions of eeg signals. *Biomed. Eng. Lett.* 3, 17–21. doi: 10.1007/s13534-013-0084-0

Bisdas, S., Shen, H., Thust, S., Katsaros, V., Stranjalis, G., Boskos, C., et al. (2018). Texture analysis- and support vector machine-assisted diffusional kurtosis imagingmay allow *in vivo* gliomas grading and idh-mutation status prediction: a preliminary study. *Sci. Rep.* 8:6108. doi: 10.1038/s41598-018-24438-4

Cao, Y., Guo, Y., Yu, H., and Yu, X. (2017). "Epileptic seizure auto-detection using deep learning method," in *International Conference on Systems and Informatics (ICSAI)* (Hangzhou: IEEE).

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chen, D., Lin, J., and Li, Y. (2018). Modified complementary ensemble empirical mode decomposition and intrinsic mode functions evaluation index for high-speed train gearbox fault diagnosis. *J. Sound Vib.* 424, 192–207. doi: 10.1016/j.jsv.2018.03.018

Chen, E., Sajatovic, M., Liu, H., Bukach, A., Tatsuoka, C., Welter, E., et al. (2018). Demographic and clinical correlates of seizure frequency: findings from the managing epilepsy well network database. *J. Clin. Neurol.* 14, 206–211. doi: 10.3988/jcn.2018.14.2.206

Chen, S.-T., Yu, P.-S., and Tang, Y.-H. (2010). Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *J. Hydrol.* 385, 13–22. doi: 10.1016/j.jhydrol.2010.01.021

Chen, Y.-C., and Hsiao, T.-C. (2018). Towards estimation of respiratory muscle effort with respiratory inductance plethysmography signals and complementary ensemble empirical mode decomposition. *Med. Biol. Eng. Comput.* 56, 1293–1303. doi: 10.1007/s11517-017-1766-z

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., et al. (2020). Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms21072274

Ding, X., Yang, Y., Stein, E. A., and Ross, T. J. (2015). Multivariate classification of smokers and nonsmokers using svm-rfe on structural mri images. *Hum. Brain Mapp.* 36, 4869–4879. doi: 10.1002/hbm.22956

Gandhi, T. K., Chakraborty, P., Roy, G. G., and Panigrahi, B. K. (2012). Discrete harmony search based expert model for epileptic seizure detection in electroencephalography. *Expert Syst. Appl.* 39, 4055–4062. doi: 10.1016/j.eswa.2011.09.093

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for qsar prediction. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009. [Epub ahead of print].

Javidi, S., Mandic, D. P., Took, C. C., and Cichocki, A. (2011). Kurtosis-based blind source extraction of complex non-circular signals with application in EEG artifact removal in real-time. *Front. Neurosci.* 5:105. doi: 10.3389/fnins.2011.00105

Kobow, K., Auvin, S., Jensen, F., Löscher, W., Mody, I., Potschka, H., et al. (2012). Finding a better drug for epilepsy: antiepileptogenesis targets. *Epilepsia* 53, 1868–1876. doi: 10.1111/j.1528-1167.2012.03716.x

Lévesque, M., Shiri, Z., Chen, L.-Y., and Avoli, M. (2017). High-frequency oscillations and mesial temporal lobe epilepsy. *Neurosci. Lett.* 667, 66–74. doi: 10.1016/j.neulet.2017.01.047

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Adv. Eng. Softw.* 69, 46–61. doi: 10.1016/j.advengsoft.2013.12.007

Muñoz-Gutiérrez, P. A., Giraldo, E., Bueno-López, M., and Molinas, M. (2018). Localization of Active brain sources from EEG signals using empirical mode decomposition: a comparative study. *Front. Integr. Neurosci.* 12:55. doi: 10.3389/fnint.2018.00055

Ou-Yang, M., Dung, L. R., Jeng, W. D., Wu, Y. Y., Wu, H. M., Weng, P. K., et al. (2012). Image stitching and image reconstruction of intestines captured using radial imaging capsule endoscope. *Opt. Eng.* 51:057004. doi: 10.1117/1.OE.51.5.057004

Puspita, J. W., Soemarno, G., Jaya, A. I., and Soewono, E. (2017). Interictal epileptiform discharges (IEDs) classification in eeg data of epilepsy patients. *J. Phys.* 943:012030. doi: 10.1088/1742-6596/943/1/012030

Ramakrishnan, T., and Sankaragomathi, B. (2017). A professional estimate on the computed tomography brain, tumor images using SVM-SMO for classification and MRG-GWO for segmentation. *Pattern Recognit. Lett.* 94, 163–171. doi: 10.1016/j.patrec.2017.03.026

Satija, U., Ramkumar, B., and Manikandan, M. S. (2017). Automated ECG noise detection and classification system for unsupervised healthcare monitoring. *IEEE J. Biomed. Health Inform.* 22, 722–732. doi: 10.1109/JBHI.2017.2686436

Shankar, K., Lakshmanaprabu, S. K., Deepak, G., Andino, M., and de Albuquerque, V. H. C. (2018). Optimal feature-based multi-kernel SVM approach for thyroid diseaseclassification. *J. Supercomput*. 76, 1128–1143. doi: 10.1007/s11227-018-2469-4

Sheng, J., Liu, S., Qin, H., Li, B., and Zhang, X. (2018). Drug-resistant epilepsy and surgery. *Curr. Neuropharmacol*. 16, 17–28. doi: 10.2174/1570159X15666170504123316

Tapia, E., Bulacio, P., and Angelone, L. (2012). Sparse and stable gene selection withconsensus svm-rfe. *Pattern Recognit. Lett.* 33, 164–172. doi: 10.1016/j.patrec.2011.09.031

Utkin, L. V., Chekh, A. I., and Zhuk, Y. A. (2016). Binary classification SVM-based algorithms with interval-valued training data using triangular and epanechnikov kernels. *Neural Netw*. 80, 53–66. doi: 10.1016/j.neunet.2016.04.005

Vinton, A., Carino, J., Vogrin, S., Macgregor, L., Kilpatrick, C., Matkovic, Z., et al. (2004). "convulsive" nonepileptic seizures have a characteristic pattern of rhythmic artifact distinguishing them from convulsive epileptic seizures. *Epilepsia* 45, 1344–1350. doi: 10.1111/j.0013-9580.2004.04704.x

Wang, B., Mei, C., Wan, Y., Zhang, J., Chen, P., Xiong, Y., et al. (2019). Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2953908. [Epub ahead of print].

Wang, W., Lu, P., Zhou, Y., Cheng, M.-T., Wang, Y., Zheng, C.-H., et al. (2020). Potential pathogenic genes prioritization based on protein domain interaction network analysis. *IEEE ACM Trans. Comput. Biol. Bioinforma.* doi: 10.1109/TCBB.2020.2983894. [Epub ahead of print].

Wu, Z., and Huang, N. E. (2010). On the filtering properties of the empirical modedecomposition. *Adv. Adapt. Data Anal.* 2, 397–414. doi: 10.1142/S1793536910000604

Wu, Z., and Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.* 1, 1–41. doi: 10.1142/S.1793536909000047

Yamany, W., Emary, E., and Hassanien, A. E. (2015). "New rough set attribute reduction algorithm based on grey wolf optimization," in *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015)* (Beni Suef).

Yang, T., Chen, J., Yan, B., and Zhou, D. (2011). Transcranial ultrasound stimulation:a possible therapeutic approach to epilepsy. *Med. Hypotheses* 76, 381–383. doi: 10.1016/j.mehy.2010.10.046

Yeh, J.-R., Shieh, J.-S., and Huang, N. E. (2010). Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. *Adv. Adapt. Data Anal.* 2, 135–156. doi: 10.1142/S1793536910000422

Yuan, Q., Zhou, W., Liu, Y., and Wang, J. (2012). Epileptic seizure detection with linear and nonlinear features. *Epilepsy Behav*. 24, 415–421. doi: 10.1016/j.yebeh.2012.05.009

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 July 2020
doi: 10.3389/fbioe.2020.00817

Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data

Binsheng He[1*†], *Rongrong Zhu*[2†], *Huandong Yang*[3], *Qingqing Lu*[4], *Weiwei Wang*[4], *Lei Song*[4], *Xue Sun*[4], *Guandong Zhang*[4], *Shijun Li*[5], *Jialiang Yang*[1,4], *Geng Tian*[4*], *Pingping Bing*[1*] *and Jidong Lang*[4*]

[1] Academician Workstation, Changsha Medical University, Changsha, China, [2] Vascular Surgery Department, Tsinghua University Affiliated Beijing Tsinghua Changgung Hospital, Beijing, China, [3] Department of Gastrointestinal Surgery, Yidu Central Hospital of Weifang, Weifang, China, [4] Geneis Beijing Co., Ltd., Beijing, China, [5] Department of Pathology, Chifeng Municipal Hospital, Chifeng, China

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Tianbao Li,
The University of Texas Health Science Center at San Antonio, United States
Cheng Guo,
Columbia University, United States

***Correspondence:**
Binsheng He
hbscsmu@163.com
Geng Tian
tiang@geneis.cn
Pingping Bing
bpping@163.com
Jidong Lang
langjd@geneis.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 April 2020*
Accepted: *26 June 2020*
Published: *30 July 2020*

Citation:
He B, Zhu R, Yang H, Lu Q, Wang W, Song L, Sun X, Zhang G, Li S, Yang J, Tian G, Bing P and Lang J (2020) Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. Front. Bioeng. Biotechnol. 8:817. doi: 10.3389/fbioe.2020.00817

Data quality control and preprocessing are often the first step in processing next-generation sequencing (NGS) data of tumors. Not only can it help us evaluate the quality of sequencing data, but it can also help us obtain high-quality data for downstream data analysis. However, by comparing data analysis results of preprocessing with Cutadapt, FastP, Trimmomatic, and raw sequencing data, we found that the frequency of mutation detection had some fluctuations and differences, and human leukocyte antigen (HLA) typing directly resulted in erroneous results. We think that our research had demonstrated the impact of data preprocessing steps on downstream data analysis results. We hope that it can promote the development or optimization of better data preprocessing methods, so that downstream information analysis can be more accurate.

Keywords: the next generation sequencing, data preprocessing, mutation, cancer, HLA typing

INTRODUCTION

In recent years, sequencing technologies, especially next-generation sequencing (NGS), have been widely used in scientific research and clinical applications. It allows for higher sequencing throughput and lower sequencing costs, and with the development and optimization of experimental and data analysis methods, the subsequent analysis results are increasingly accurate. For example, important techniques for detecting cancer-associated biomarkers using liquid biopsy techniques (Esposito et al., 2017) are essentially done using the NGS technology platform, especially in the detection of cell-free tumor DNA (ctDNA) in plasma, such as Duplex sequencing (Schmitt et al., 2012), Cancer Personalized Profiling by deep Sequencing (CAPP-Seq) (Newman et al., 2014), and Targeted Error Correction Sequencing (TEC-Seq) (Phallen et al., 2017). However, ctDNA sequencing data have strong background noise, contamination of sequencing adapters, unbalanced base distribution, sequencing quality and errors introduced during the experiments; these factors have a crucial impact on the accuracy of detecting low-frequency and even ultra-low-frequency mutations in ctDNA. Therefore, quality control and data preprocessing are especially important for obtaining downstream high-quality and high-confidence analytical data to reduce false positives and false negatives.

Illumina reads are commonly 36–300 nucleotide bases produced by a reversible-terminator cyclic reaction associated to base-specific colorimetric signals within the sequencing machine. Reads can be separated "single-end" or "paired-end" reads, in which case they are representing both extremities of the same nucleotide fragment. These colorimetric signals are translated into base calls by an internal Illumina software (CASAVA), represented in the FASTQ format (Cock et al., 2010), where each nucleotide is associated to an ASCII-encoded quality number corresponding to a PHRED score (Q) (Ewing and Green, 1998), which is in recent Illumina runs ranges from 0 to 41 and the error rate at each position ranges from 7.94e-5 to 1. Whatever the original cause of low quality or high error chance nucleotides, such as air bubbles, spot-specific signal noise, malfunctioning laser or lens, and so on, the *Q* value if encoded and stored together with the sequence information, and this confidence information can be used for subsequent analysis, together with the sequence information itself.

At present, there are many software programs for data quality preprocessing. Cutadapt (Martin, 2011), which is widely used, is the only stand-alone tool that can correctly trim color space reads. It can search for multiple adapters in a single run of the program and removes the best matching one. It can optionally search and remove an adapter multiple times, which is useful when (perhaps accidentally) library preparation has led to an adapter being appended multiple times. It can either trim or discard reads in which an adapter occurs. Reads that are outside a specified length range after trimming can also be discarded. In addition to adapter trimming, low-quality ends of reads can be trimmed using the same algorithm as Burrows-Wheeler Aligner (BWA). FastP (Chen et al., 2018b), as an all-in-one FASTQ preprocessor, provides functions including quality profiling, adapter trimming, read filtering, and base correction. It supports both single-end and paired-end short read data and provides basic support for long-read data, which are typically generated by PacBio and Nanopore sequencers. Trimmomatic (Bolger et al., 2014) includes a variety of processing steps for read trimming and filtering, but the main algorithmic innovations are related to the identification of adapter sequences and quality filtering. Trimmomatic uses a pipeline-based architecture, allowing individual "steps" (adapter removal, quality filtering, and so on) to be applied to each read/read pair in the order specified by the user. Each step can choose to work on the reads in isolation or work on the combined pair, as appropriate. The tool tracks read pairing and stores "paired" and "single" reads separately.

The data preprocessing software and algorithms have shown excellent results in published articles. We have also used them to obtain high-quality clean data to do downstream analysis, such as alignment and mutation detection. Generally, when there are false-positive or false-negative results, we tend to think this may be due to unreasonable parameter settings in the analysis process or other experimental reasons, but this may not always be the case. We analyzed and compared the raw sequencing data with commonly used data preprocessing software, such as Cutadapt, FastP, and Trimmomatic, and found that the data preprocessing results affected the subsequent detection results. Therefore, we realized that in the data preprocessing, we need to choose the software and algorithms carefully, and the data preprocessing algorithms need to be further improved according to actual data features. It is necessary to make different choices according to specific data analysis.

MATERIALS AND METHODS

Sample Collection

HD753, a reference genomic DNA (gDNA), is used as the reference standard (Horizon Diagnostics^{TM}, Waterbeach, United Kingdom) and contains 10 mutation variations: AKT serine/threonine kinase 1 (*AKT1*) p.E17K (5%), phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*) p.E545K (5.6%), epidermal growth factor receptor (*EGFR*) p.745-750del (5.3%), *EGFR* p.V769delinsVASV (5.6%), KRAS proto-oncogene, GTPase (*KRAS*) p.G13D (5.6%), notch receptor 1 (*NOTCH1*) p.P668S (5%), MET proto-oncogene, receptor tyrosine kinase (*MET*) p.V237fs (2.5%), BRCA2 DNA repair associated (*BRCA2*) p.A1689fs (5.6%), *EGFR* p.G719S (5.3%), B-Raf proto-oncogene, serine/threonine kinase (*BRAF*) p.V600E (18.2%), and *PIK3CA* p.H1047R (16.7%). The original HD753 reference has two replicates. We then used the standard sample to do three fivefold dilution experiments and every five-diluted sample has also two replicates, while the negative control sample, a healthy human white blood cells, also has two replicates.

All five human leukocyte antigen (HLA) typing samples and 75 mutation detection samples were obtained from lung cancer patients and informed written consent was obtained from the patients and de-identification. The 80 clinical samples we used were collected from October 2017 to May 2018.

Experiment Workflow

gDNA for NGS-based mutation variations analysis was extracted using the GONOROAD Kit (Qiagen, Hilden, Germany) for formalin-fixed and paraffin-embedded (FFPE) tissue. DNA (200 ng) was used to build the library by using NEBNext Ultra II DNA library Prep Kit for Illumina (96 reactions) (NEB, Ipswich, MA, United States). Integrated DNA technologies (IDT, Skokie, IL, United States) customized probes were used for hybridization capture. We used the *Genesis* 41 gene tumor hotspot mutation customized panel (**Supplementary Sheet 1**) for eight gDNA standard samples and two negative control samples. Quantification was performed with a Library Quantification Kit – Illumina/Universal (Kapa Biosystems, Wilmington, MA, United States) on an ABI 7500 Real Time PCR system (Applied Biosystems, Waltham, MA, United States). A Quality control Agilent 2100 Bioanalyzer with a High Sensitivity DNA Kit was used for quality control (Agilent Technologies, Santa Clara, CA, United States). NGS analysis was performed on a Nextseq500 instrument according to the manufacturer's instructions (Illumina, San Diego, CA, United States). With a NextSeq500/550 High Output V2 kit, Illumina Nextseq500 was used for DNA sequencing in 302 cycles, standing for paired-End 151bp.

The 75 clinical samples' cell-free DNA was extracted using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The obtained DNA (20 ng/sample) was then used to build libraries using Accel-NGS® 2S Plus DNA Library Kits (96 reactions; Swift Biosciences, Ann Arbor, MI, United States). Customized probes were obtained from Integrated DNA technologies (IDT, Skokie, IL, United States) and were used for hybridization capture. All cfDNA libraries utilized the *Genesis* 41 gene tumor hotspot mutation customized panel and were quantified using a Universal Library Quantification Kit (Kapa Biosystems, Wilmington, MA, United States) on an ABI 7500 Real-Time PCR system (Applied Biosystems, Waltham, MA, United States). Sample quality was evaluated using a high sensitivity DNA kit (Agilent Technologies, Santa Clara, CA, United States) with an Agilent 2100 Bioanalyzer per the manufacturer's instructions. NGS with fusion detection was performed using a NextSeq 500/550 High Output v2 kit with a NextSeq 500 sequencer (Illumina, San Diego, CA, United States) for 302 cycles, with standing paired-end reads of 151 bp.

Five DNA samples for HLA typing analysis were extracted from the FFPE tumor tissues using the GeneRead DNA FFPE Kit (Qiagen, Hilden, Germany). DNA samples were normalized to yield a 100 − 250 ng input. Whole genome libraries were prepared using NEBNext® Ultra^{TM} II DNA Library Prep (NEB, Ipswich, MA, United States) and through a series of steps including covaris shearing, end-repair, A-base addition, barcoded adapter ligation, and PCR amplification. Libraries were quantitated using a Qubit dsDNA HS Kit (Invitrogen, Carlsbad, CA, United States) and quality assessed with Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States) as per the manufacturer's protocol. Targeted enrichment was carried out on the prepared libraries to specifically pull down DNA fragments that contained the target site using custom 5′ biotinylated capture probes. Four libraries were then pooled at 125 ng each for a total of 500 ng. Cot-1 DNA (Sigma-Aldrich, MO, United States) and universal blocking oligonucleotides (IDT, Skokie, IL, United States) were added to the pooled libraries and dried in a SpeedVac. The dried mixture was then resuspended in IDT Hybridization Buffer and Hybridization enhancer (IDT, Skokie, IL, United States) and hybridized for 4 h with custom 5′ biotinylated capture IDT probes (IDT, Skokie, IL, United States) and BOKE probes (BOKE, Beijing, China). Streptavidin DynaBeads (Invitrogen, Carlsbad, CA, United States) were used for capture and washes were performed using xGenLockdown-Reagents Kit (IDT, Skokie, IL, United States). The final hybridized product was amplified using KAPA Hifi HotStart Ready Mix (Kapa Biosystems, Wilmington, MA, United States) and Illumina sequencing primers for a total of 15 cycles. Final target capture library quantification was performed using a Qubit dsDNA HS Kit (Invitrogen) and quality assessed with Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States). With a NextSeq500/550 High Output V2 kit, Illumina Nextseq500 (Illumina) was used for DNA sequencing in 302 cycles, standing for paired-End 151 bp.

Mutation Validation

EGFR-T790M (25), *EGFR*-L858R (26), *BRAF*-V600E (5), *PIK3CA*-E545K (6), *KRAS*-G12C (11), and *KRAS*-G12V (2) mutant allele frequencies were determined using a Digital Droplet PCR system (Bio-Rad Laboratories, Inc., Hercules, CA, United States), with a droplet size of 1 nL in a total reaction volume of 20 μL with ∼20 ng of cfDNA library utilized. All primers and probes were synthesized by IDT (Skokie, IL, United States). Droplet counts were determined using the QuantaSoft software (Bio-Rad) (**Supplementary Sheet 7**).

HLA Typing Validation

Human leukocyte antigen typing was performed at the BFR Medical Laboratory (BFR, Beijing, China) by the high–resolution HLA sequence-based typing method (HLA-SBT).

Data Analysis for Mutation Detection

We used Cutadapt (version 1.3, *parameter: -b AGATCGGA AGAGCACACGTCTGAACTCCAGTCAC -b AGATCGGAAGAG CGTCGTGTAGGGAAAGAGTGTA -e 0.01 -m 15*), FastP (version 0.20.0, *parameter: -trim_poly_g*), and Trimmomatic (version 0.39, *parameter: PE -threads 4 -phred33 ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 MINLEN:15*) to preprocess the raw sequencing data (*Fastq*), filtering out the adapter contamination reads, low-quality reads, and unpaired reads to get clean data. We used the Bwa aln (Version: 0.7.12-r1039) algorithm to align the clean data to the human reference genome (hg19) and get the Sequence Alignment/Map format (*sam)* file. For the Binary Alignment/Map format (*bam)* file, the *sam* file was sorted by samtools (Version: 0.1.19-44428cd). According to the bed interval file of the *Genesis* 41 gene tumor hotspot mutation customized panel, we used freebayes (version: v1.0.2-6-g3ce827d, *parameter: -j -m 10 -q 20 -F 0.001 -C 1 -t bed.file –f hg19.fa*) to call single nucleotide polymorphisms (SNPs) and insertions or deletions (indels), and then used ANNOVAR to do the annotation (**Figure 1**).

Data Analysis for HLA Typing

The method and parameters of data preprocessing were consistent with the above. We used Novoalign (version: V3.09.02, parameter: -t 30 -o SAM -r all -l 80 -e 100 -i PE 200 140) to align the clean data to the HLA reference sequence. We then used samtools (version: 1.3.1) to sort the *sam* files to get the *bam* files. We used Athlates (version: 1.0, default parameter) for typing analysis of HLA-A*, HLA-B*, and HLA-C* (**Figure 1**).

RESULTS

No Significant Difference in the Impact on Data Quality After Data Preprocessing

For the 10 standard samples data, calculating the number of reads, GC content, Q20 ratio, average depth, capture efficiency, and duplication rate after data preprocessing (**Figures 2A–F** and **Supplementary Sheet 2**), we found that the data of the three software-processed indicators, except the Q20 ratio,

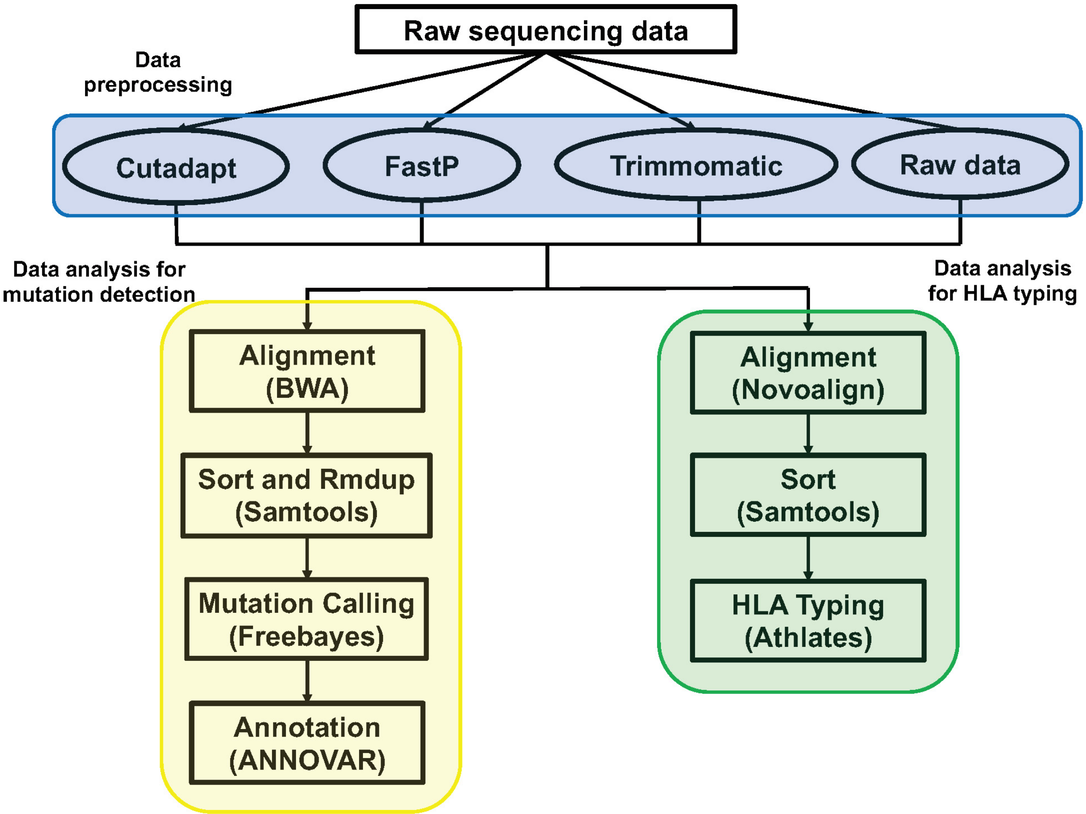


FIGURE 1 | The pipeline of data analysis. The blue section was the three method of data preprocessing: Cutadapt, FastP, Trimmomatic, and raw sequencing data; the yellow section was the pipeline of data analysis for mutation detection; the green section was the pipeline of data analysis for HLA typing.

showed no significant difference (the *p*-value of the two-tailed heteroscedastic *T*-test was > 0.05). The Q20 ratio after FastP treatment was significantly improved, and the two-tailed heteroscedastic *T*-test *p*-values were 0.036 (vs. Raw data), 0.040 (vs. Cutadapt), and 0.026 (vs. Trimmomatic).

For the 75 clinical samples data, we also found the same conclusion that the data, except the Q20 ratio, showed no significant difference (**Supplementary Figure 1** and **Supplementary Sheet 6**). The Q20 ratio after FastP treatment was significantly improved, and the two-tailed heteroscedastic *T*-test *p*-values were 1.69476E-10 (vs. Raw data), 3.05502E-10 (vs. Cutadapt), and 2.24745E-11 (vs. Trimmomatic).

Frequency of Mutations Detected After Data Preprocessing May Be Affected

For the 10 standard samples data, we found that all of the hotspot mutations were detected in raw data, Cutadapt, FastP, and Trimmomatic preprocessing data in the two replicate reference standard gDNAs and the fivefold diluted HD753 specimens (**Supplementary Sheet 3**), while false positive results of *EGFR* p.G719S were found in all the negative control samples (HD753-NB). It may have been caused by sequencing errors or contamination introduced during the experiment. We found that the four preprocessing data analysis results had lower mutation frequencies than the expected frequencies of HD753-0A and HD753-0B (**Figure 3A**), which may be related to the experimental capture operation. There was no statistical difference between the distribution of frequencies of the four data types (the *p*-values of the two-tailed heteroscedastic *T*-test were > 0.05). But for the repeated dilution samples, the detected mutation frequency fluctuated greatly (**Figures 3B,C**). We assumed that a mutation frequency greater than 1% was used as a threshold for positive result for the FFPE or tissue samples. For a hotspot mutation *AKT1* p.E17K in HD753-1A, the detection results after Cutadapt and FastP data pretreatment were positive, and the detection frequencies were 1.06% (41/3869) and 1.00% (38/3785), respectively. Meanwhile, the raw data and Trimmomatic treatments were negative, with detection frequencies of 0.95% (34/3579) and 0.96% (34/3549), respectively. For *EGFR* p. 745_750del of HD753-2A, the results were negative after pretreatment with Cutadapt and FastP data, and the detection frequencies were 0.97% (20/2055) and 0.98% (20/2051), respectively. The results of the raw data and Trimmomatic treatment were positive, and the detection frequencies were 1.05% (20/1900) and 1.06% (20/1889), respectively. While the mutation of *NOTCH1* p.P668S in HD753-2A was detected as a positive result, which was preprocessed by FastP data, the detection frequency was 1.12% (32/2860). The data preprocessed by Cutadapt, Trimmomatic, and the raw data were negative,

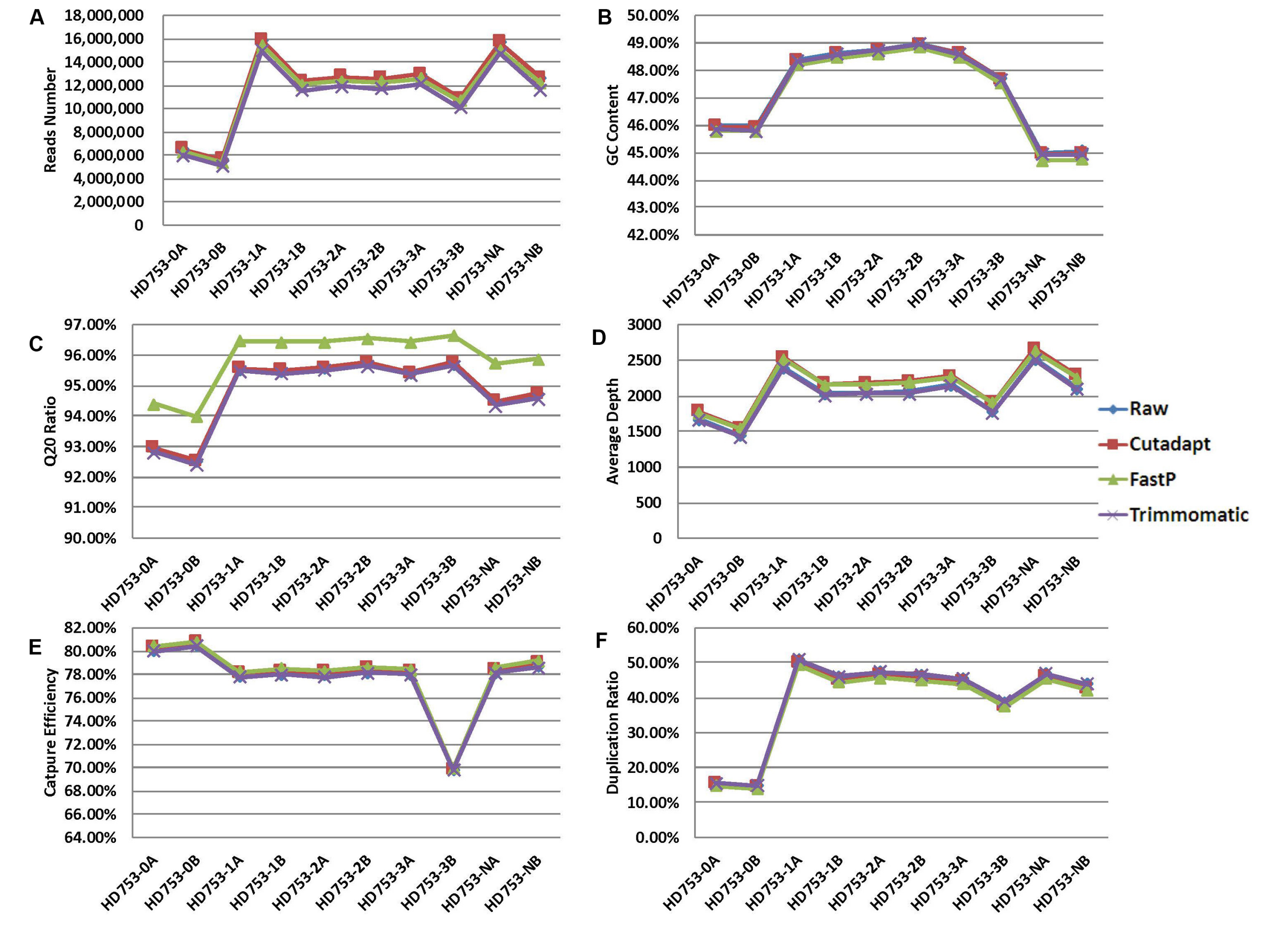


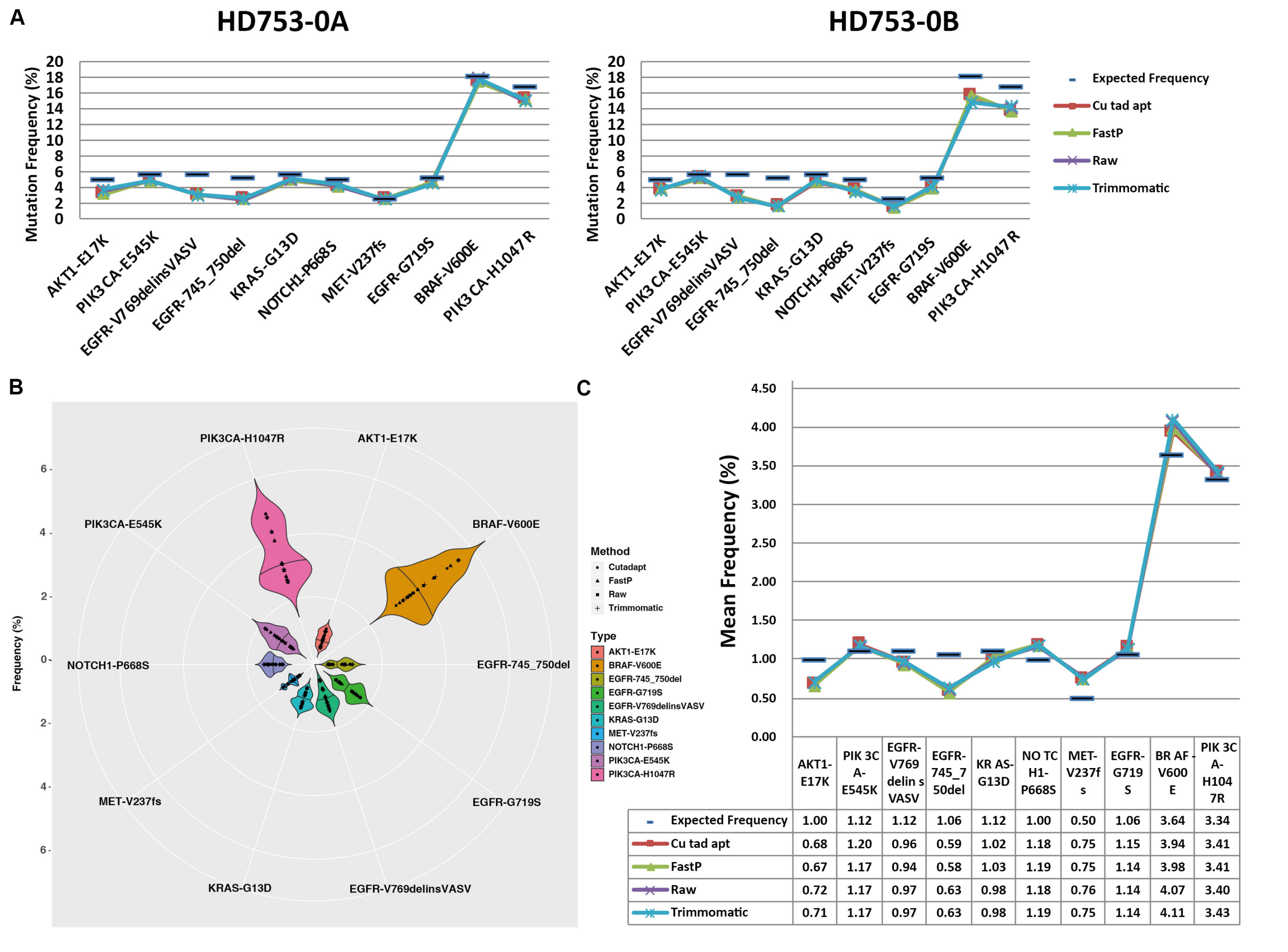
FIGURE 2 | Quality control statistical distribution of Cutadapt, FastP, and Trimmomatic preprocessed data and raw sequencing data. **(A)** Statistical distribution of the number of reads, **(B)** statistical distribution of the GC content, **(C)** statistical distribution of the Q20 ratio, **(D)** statistical distribution of the average depth, **(E)** statistical distribution of the capture efficiency, and **(F)** statistical distribution of the duplication rate.

and the detection frequency was 0.98% (28/2857), 0.98% (26/2659), and 0.97% (26/2686), respectively. We also found that the results of Trimmomatic data pretreatment were basically consistent with the raw sequencing data, and the results of Cutadapt and FastP data pretreatment were consistent. It showed that there were no significant differences in the four methods of mutation support reads number. However, the sequencing depth of the mutation sites were quite different. That may be because Cutadapt and FastP only trim the reads, making the sequence shorter and achieving multi-alignment in the alignment process, which could be increasing the probability of the alignment. In contrast, in our paired-end preprocessing data, Trimmomatic retained the sequenced full-length sequence to be consistent with the raw sequencing data (**Supplementary Sheet 3**).

For the 75 clinical samples data, we used digital droplet PCR system to determine the mutations' frequency, and the frequency limit was 0.1%. Consistent with the results found in the results of the standard, there was some fluctuation in the detection frequency after 4 data preprocessing method, but the coefficient of determination R^2 of raw data, Cutadapt, Fastp, and Trimmomatic was 0.9386, 0.9381, 0.9416, and 0.9416, respectively. Compared with the result of ddPCR, the detection rate of data preprocessing by raw data, Cutadapt, Fastp, and Trimmomatic was 94.67% (71/75), 100% (75/75), 98.67% (74/75), and 96.00% (72/75), respectively (**Supplementary Sheet 7**). We found that the false negatives had a low mutation frequency (**Table 1**), and the effect of Cutadapt was the best compared to the other three methods.

HLA Typing Data After Preprocessing Had a Significant Impact on Data Quality

We calculated the reads number, GC content, and Q20 ratio for 10 HLA typing samples after data preprocessing (**Figures 4A–C** and **Supplementary Sheet 4**). We found that the data distribution after the three software treatments had significant fluctuations. The Q20 ratio was statistically significant based on the two-tailed heteroscedasticity test. The number of reads after Trimmomatic data preprocessing was significantly



	AKT1-E17K	PIK 3CA-E545K	EGFR-V769 delin sVASV	EGFR-745_750del	KR AS-G13D	NO TCH1-P668S	MET-V237fs	EGFR-G719S	BR AF -V600E	PIK 3CA-H1047R
Expected Frequency	1.00	1.12	1.12	1.06	1.12	1.00	0.50	1.06	3.64	3.34
Cu tad apt	0.68	1.20	0.96	0.59	1.02	1.18	0.75	1.15	3.94	3.41
FastP	0.67	1.17	0.94	0.58	1.03	1.19	0.75	1.14	3.98	3.41
Raw	0.72	1.17	0.97	0.63	0.98	1.18	0.76	1.14	4.07	3.40
Trimmomatic	0.71	1.17	0.97	0.63	0.98	1.19	0.75	1.14	4.11	3.43

FIGURE 3 | Distribution of hotspot mutation detection in the reference standard samples: **(A)** frequency distribution of the hotspot mutation detection in the two experimental replicates of the reference standard samples, **(B)** frequency distribution of the hotspot mutation detection in each experimental repeat of the fivefold dilution standard samples, and **(C)** average frequency distribution of the hotspot mutation detection in each experimental repeat of the fivefold dilution standard samples.

TABLE 1 | Compared with ddPCR results, the false negative results of data preprocessing by raw data, Cutadapt, Fastp, and Trimmomatic.

Sample_ID	Mutation_Type	ddPCR (%)	Raw_data (%)	Cutadapt (%)	Fastp (%)	Trimmomatic (%)
T790M-sample21	EGFR:p.T790M	0.17	0.00	0.14	0.00	0.00
L858R-sample19	EGFR:p.L858R	0.32	0.00	0.15	0.15	0.15
V600E-sample8	BRAF:p.V600E	0.31	0.00	0.26	0.26	0.00
E545K-sample10	PIK3CA:p.E545K	0.24	0.00	0.16	0.14	0.00

**Mutations' frequency limit was 0.1%.*

different from the distribution of the remaining three data types (**Figure 4D**).

Data Preprocessing May Affect HLA Typing Analysis

We performed HLA typing data analysis on five samples captured by BOKE and IDT probes. We obtained incorrect typing results with the data after pretreatment of Cutadapt and FastP (**Tables 2a, 2b** and **Supplementary Sheet 5**). After pretreatment with Cutadapt and FastP data, the sample NZTD181200662 showed errors in the typing analysis of HLA-A and HLA-C, whether it was the BOKE probe capture or IDT probe capture, which was inconsistent with the validation results (such as **Table 1**, the red background was shown). The results of the raw data and Trimmomatic data preprocessing were consistent with the validation results. Since the NZTD181200690 sample was classified incorrectly in the four analysis results, it may have been caused by experiments or sequencing errors. Therefore, for the overall result of the BOKE probe capture, the accuracy

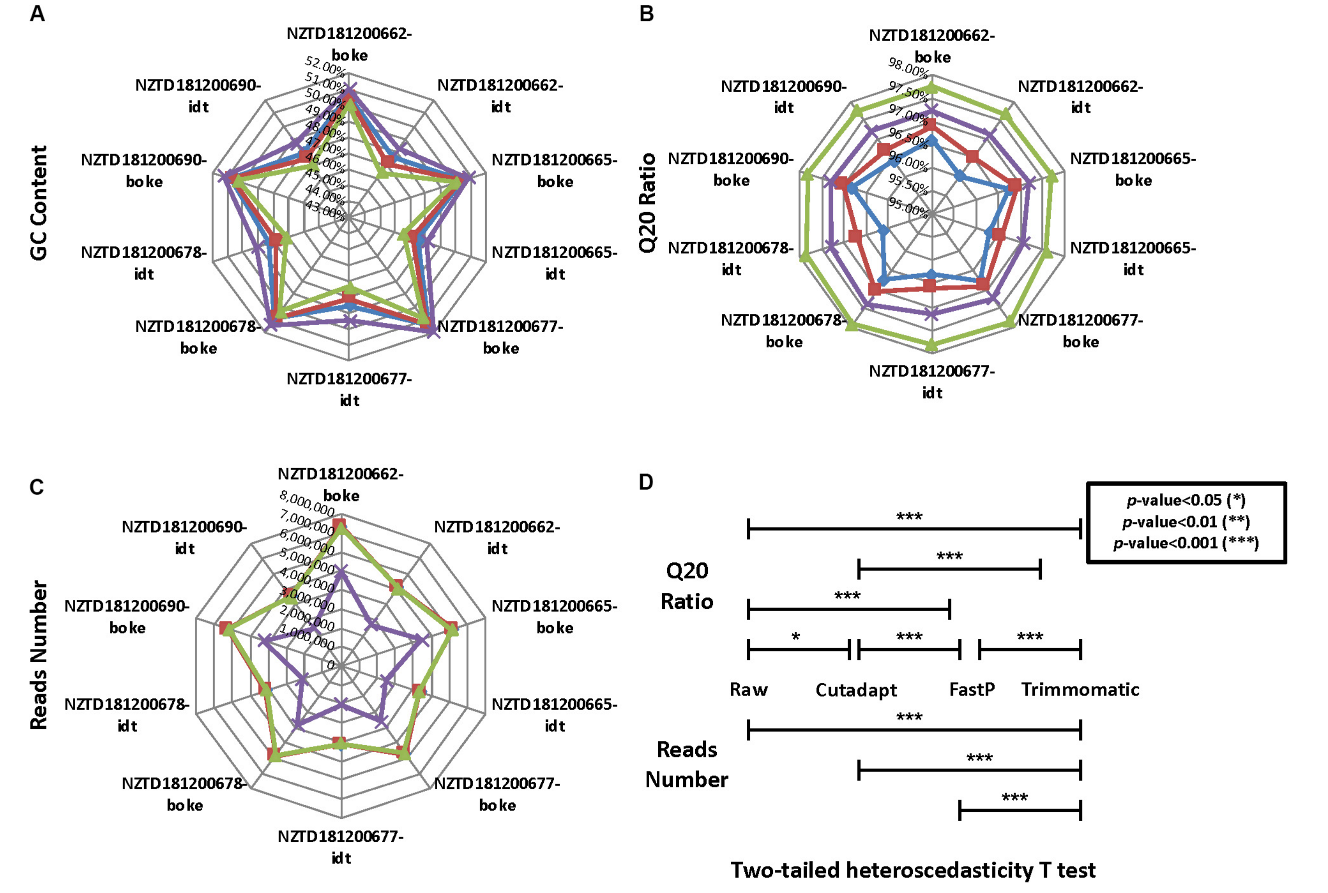


FIGURE 4 | Quality control statistical distribution of Cutadapt, FastP, Trimmomatic preprocessed data, and raw sequencing data for 10 HLA typing samples. **(A)** Statistical distribution of the GC content, **(B)** statistical distribution of the Q20 ratio, **(C)** statistical distribution of the reads number, and **(D)** significant level for the four data types in the Q20 ratio and the reads number.

after treatment with Cutadapt, FastP, raw data, and Trimmomatic was 86.67, 80.00, 93.33, and 93.33%, respectively. For the overall results of the IDT probe capture, the accuracy rates after treatment with Cutadapt, FastP, raw data, and Trimmomatic were 86.67, 86.67, 93.33, and 93.33%, respectively.

For the NZTD181200662 sample, we extracted the reads ID of the sequencing data captured by the BOKE probe capture and the IDT probe (**Figures 5A,B**), and we wanted to know if the sequence reads causing the typing error had certain characteristics. We found that the three preprocessed data were highly consistent with the raw sequencing data, but the Trimmomatic preprocessed data also had many specific reads, accounting for 32.64 and 47.52% of the FastP preprocessing data captured by the BOKE and IDT probes, respectively. This phenomenon was basically the same in the remaining samples (**Supplementary Figures 3–6**). Due to the high accuracy of the raw data and Trimmomatic preprocessed data, we assumed that the read features that caused the incorrect HLA typing data of the Cutadapt and FastP data were in their specific reads compared with the raw data and Trimmomatic data. We extracted this part of the read and analyzed the length distribution of the reads. We found that the length of the read from 143 bp to 149 bp was significantly reduced (**Figure 5C**). Therefore, we extracted the 143–149 bp reads from the NZTD181200662's BOKE and IDT probes captured data processed by Cutadapt and FastP for HLA-A and HLA-C typing, respectively. The results were consistent with the validation results for the BOKE's capture probes. The HLA-A typing results of Cutadapt and FastP processed data were A*02:01:01/A*02:01:01 and the HLA-C typing results were C*08:22/C*08:22. For the IDT's capture probe, the HLA-A typing results of Cutadapt and FastP processed data were A*02:01:01/A*02:01:01 and the HLA-C typing results were C*08:22/C*08:22.

DISCUSSION

Data quality control and preprocessing play an important role in data analysis in scientific research and clinical fields, and it is often the first step in data analysis. We believe that it can help us evaluate the experimental steps or problems in the sequencing process, and also reduce the sequence of low-quality or adapter contamination, reduce the computational cost, and allow us to obtain high-quality sequencing sequences for downstream

TABLE 2a | Summary of HLA typing results in the four data types with BOKE capture probes.

Sample-BOKE	Validation			Cutadapt			FastP			Raw			Trimmomatic		
	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C
NZTD181200662	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22
	A*02:01:01	B*81:02	C*08:22	A*02:01:01	B*81:02	C*08:01:01	A*02:53N/A*02:96/. . .	B*81:02	C*08:01:01	A*02:01:01	B*81:02	C*08:22	A*02:01:01	B*81:02	C*08:22
NZTD181200665	A*11:01:01	B*52:01:01:02	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01	A*24:02:01	B*54:01:01	C*01:02:01
	A*24:02:01	B*54:01:01	C*12:02:02:01	A*24:02:01	B*52:01:01:02	C*12:02:02	A*24:02:01	B*52:01:01:02	C*12:02:02	A*24:02:01	B*52:01:01:02	C*12:02:02	A*11:01:01	B*52:01:01:02	C*12:02:02
NZTD181200677	A*03:01:01:01	B*40:01:02	C*03:04:01	A*03:01:01:01	B*40:01:02	C*05:01:01:02	A*03:01:01:01	B*40:01:02	C*05:01:01:02	A*03:01:01:01	B*40:01:02	C*05:01:01:02	A*03:01:01:01	B*40:01:02	C*05:01:01:02
	A*11:01:01	B*44:02:01	C*05:01:01:02	A*11:01:01	B*44:02:01	C*03:04:01	A*11:01:01	B*44:02:01	C*03:04:01	A*11:01:01	B*44:02:01	C*03:04:01	A*11:01:01	B*44:02:01	C*03:04:01
NZTD181200678	A*02:01:01	B*41:01:01:01	C*15:02:01:01	A*02:01:01	B*51:01:01	C*17:01:01	A*02:01:01	B*51:01:01	C*17:01:01	A*02:01:01	B*51:01:01	C*17:01:01	A*02:01:01	B*51:01:01	C*17:01:01
	A*03:01:01:01	B*51:01:01	C*17:01:01:05	A*03:01:01:01	B*41:01	C*15:02:01	A*03:01:01:01	B*41:01	C*15:02:01	A*03:01:01:01	B*41:01	C*15:02:01	A*03:01:01:01	B*41:01	C*15:02:01
NZTD181200690	A*11:02:01	B*27:04:01	C*12:02:02:01	A*11:77	B*27:04:01	C*12:02:02	A*11:77	B*27:04:01	C*12:02:02	A*11:02:01	B*27:04:01	C*12:02:02	A*11:02:01	B*27:04:01	C*12:02:02
	A*11:02:01	B*51:01:01:01	C*14:02:01	A*11:02:01	B*51:01:01	C*14:02:01	A*11:02:01	B*51:01:01	C*14:02:01	A*11:77	B*51:01:01	C*14:02:01	A*11:77	B*51:01:01	C*14:02:01

The color values represented that the predicted results were inconsistent with the validated results.

TABLE 2b | Summary of HLA typing results in the four data types with IDT capture probes.

Sample-IDT	Validation			Cutadapt			FastP			Raw			Trimmomatic		
	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C	HLA-A	HLA-B	HLA-C
NZTD181200662	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22	A*02:01:01	B*40:06:01:01	C*08:22
	A*02:01:01	B*81:02	C*08:22	A*02:338	B*81:02	C*08:22	A*02:338	B*81:02	C*08:22	A*02:01:01	B*81:02	C*08:22	A*02:01:01	B*81:02	C*08:22
NZTD181200665	A*11:01:01	B*52:01:01:02	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01	A*24:02:01	B*54:01:01	C*01:02:01	A*11:01:01	B*54:01:01	C*01:02:01
	A*24:02:01	B*54:01:01	C*12:02:02:01	A*24:02:01	B*52:01:01:02	C*12:02:02	A*24:02:01	B*52:01:01:02	C*12:02:02	A*11:01:01	B*52:01:01:02	C*12:02:02	A*24:02:01	B*52:01:01:02	C*12:02:02
NZTD181200677	A*03:01:01:01	B*40:01:02	C*03:04:01	A*11:01:01	B*40:01:02	C*03:04:01	A*11:01:01	B*40:01:02	C*05:01:01:02	A*03:01:01:01	B*40:01:02	C*03:04:01	A*03:01:01:01	B*40:01:02	C*03:04:01
	A*11:01:01	B*44:02:01	C*05:01:01:02	A*03:01:01:01	B*44:02:01	C*05:01:01:02	A*03:01:01:01	B*44:02:01	C*03:04:01	A*11:01:01	B*44:02:01	C*05:01:01:02	A*11:01:01	B*44:02:01	C*05:01:01:02
NZTD181200678	A*02:01:01	B*41:01:01:01	C*15:02:01:01	A*02:01:01	B*41:01	C*17:01:01	A*02:01:01	B*51:01:01	C*17:01:01	A*02:01:01	B*51:01:01	C*15:02:01	A*02:01:01	B*51:01:01	C*17:01:01
	A*03:01:01:01	B*51:01:01	C*17:01:01:05	A*03:01:01:01	B*51:01:01	C*15:02:01	A*03:01:01:01	B*41:01	C*15:02:01	A*03:01:01:01	B*41:01	C*17:01:01	A*03:01:01:01	B*41:01	C*15:02:01
NZTD181200690	A*11:02:01	B*27:04:01	C*12:02:02:01	A*11:02:01	B*27:04:01	C*12:02:02	A*11:02:01	B*27:04:01	C*12:02:02	A*11:02:01	B*27:04:01	C*14:02:01	A*11:02:01	B*27:04:01	C*12:02:02
	A*11:02:01	B*51:01:01:01	C*14:02:01	A*11:77	B*51:01:01	C*14:02:01	A*11:77	B*51:01:01	C*14:02:01	A*11:77	B*51:01:01	C*12:02:02	A*11:126	B*51:01:01	C*14:02:01

The color values represented that the predicted results were inconsistent with the validated results.

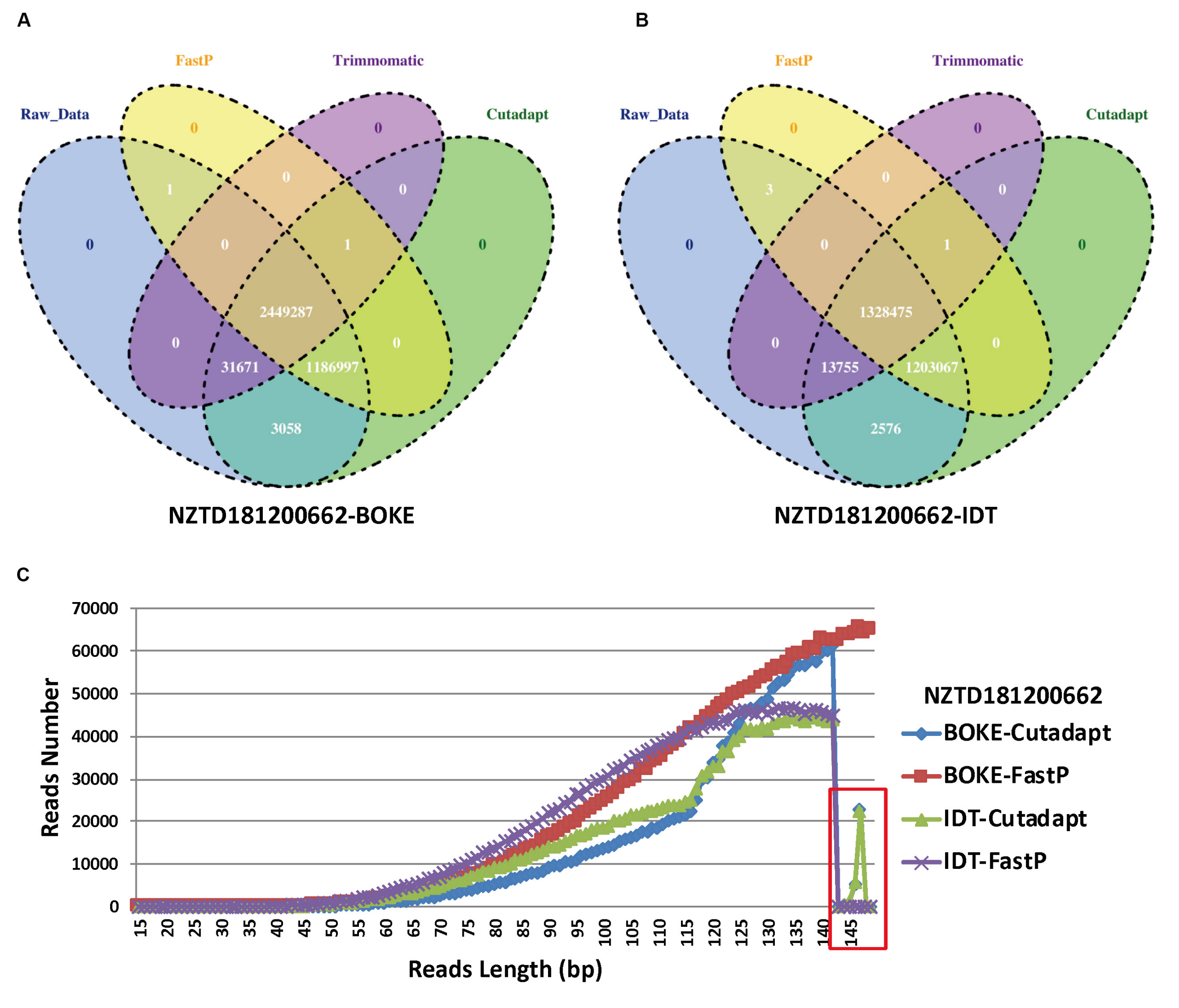


FIGURE 5 | Statistics for the reads ID in the four data preprocessing types of sample NZTD181200662. **(A)** Four kinds of reads ID difference by BOKE probe capture data processing, **(B)** four kinds of reads ID difference by IDT probe capture data processing, and **(C)** the length distribution in the Cutadapt preprocessing data for the BOKE and IDT probes and the FastP preprocessing data for the BOKE and IDT probes.

analysis, making the analysis results more reliable. When false-positive or false-negative results are obtained, we usually think it is caused by (i) experimental factors, such as errors introduced by PCR or sample contamination; (ii) sequencing factors, such as sequencing quality and data contamination caused by index hopping when splitting data; or (iii) analysis software parameters setting factors, such as alignment software or specific parameter adjustment of downstream personalized analysis. There have been some studies that have done some comparisons of data preprocessing methods, for example, Del Fabbro et al. (2013) evaluated nine different trimming algorithms in four datasets and three common NGS-based applications (RNA-Seq, SNP calling, and genome assembly) (Chen et al., 2018a). But until now, we still did not notice that the data preprocessing step may also have a certain impact on the analysis results. We may even consider the notion that the sole purpose of data preprocessing is to reduce the downstream computing consumption and describe the quality of the sequencing data, as the actual importance and meaning have been previously neglected.

In this study, we compared commonly used data preprocessing software and found differences in the detection of hotspot mutations and HLA typing. Although the detection results may be affected by the three factors described above, for the different processing of the same data and the subsequent set of analysis processes, this could reflect the difference between the different pretreatment methods and the impact on the detection results. For the current "liquid biopsy" method, the sample testing requirements are to detect ctDNA mutations in the plasma to guide subsequent targeted drug therapy or real-time monitoring, but the ctDNA's content in the plasma is very small (Bettegowda et al., 2014). For the accuracy of detecting mutations, each step in the experiment and analysis process should require strict quality control. Each step plays an important role in the detection results and cannot be ignored. Particularly for the detection of low-frequency or ultra-low-frequency mutations such as hotspot mutations, we showed that if the sequencing depth and mutation support reads number changes, it may directly lead to false-positive or false-negative results, which has a huge impact on clinical testing.

Currently, there are many available data quality control and preprocessing software programs, in addition to the three methods described in the article, such as FASTQC (Andrews, 2010), SOAPnuke (Chen et al., 2018c), and NGSQC (Dai et al., 2010). But most methods for the strategy of data preprocessing are to cut off all subsequent bases as long as the average quality of the bases in a certain bin or consecutive bases is below a certain threshold to reduce memory consumption and I/O reading, increasing the speed of operation. They do not notice the distribution of the actual low-mass bases in the sequence, which could result in many short sequences and may reduce the accuracy of downstream alignment and increase the sequencing depth of some reference sites. Thus, the analysis results may be inaccurate, and the effect may not be as good as the result of not doing data preprocessing, which was also confirmed in our analysis results. As the sequencing throughput becomes higher and higher, the sequencing read length becomes longer and longer, but the longer the sequencing read length, the worse the sequencing quality. Therefore, data preprocessing becomes increasingly important in data analysis. Existing principles and methods of data preprocessing for the long sequencing read length are worth considering. Our research explains the impact of data preprocessing steps on downstream data analysis results. We hope that our study can promote the development or optimization for the data preprocessing methods, so that downstream information analysis can be more accurate.

DATA AVAILABILITY STATEMENT

The standard sample and 5 HLA-typing clinical sample FASTQ data files are available from the NCBI Sequence Read Archive (SRA) database (BioProject ID: PRJNA556054). The 75 clinical sample FASTQ data files are available from the NCBI Sequence Read Archive (SRA) database (BioProject ID: PRJNA562379).

AUTHOR CONTRIBUTIONS

JL, BH, PB, and GT designed the study, collected, analyzed, and interpreted the data, and wrote the manuscript. HY, JY, QL, WW, LS, XS, GZ, and RZ did the experiment. RZ and SL reviewed the manuscript. All authors approved the final version of the manuscript.

FUNDING

This study was partially funded by the Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105), the Natural Science Foundation of China (Nos. 61803151, 81560405, and 81960449), the Natural Science Foundation of Hunan Province (Nos. 2018JJ2461, 2018JJ2463, and 2018JJ3570), the Project of Scientific Research Fund of Hunan Provincial Education Department (Nos. 19A060 and 19C0185), the Talents Science and Technology Program of Changsha (No. kq1907035), the project to introduce intelligence from oversea experts to the Changsha City (No. 2089901), and Beijing Tsinghua Changgung Hospital Fund (No. 12016C1009).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00817/full#supplementary-material

REFERENCES

Andrews, S. (2010). *A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., et al. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* 6:224ra224.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chen, S., Liu, M., and Zhou, Y. (2018a). "Bioinformatics analysis for cell-free tumor DNA sequencing data," in *Computational Systems Biology. Methods in Molecular Biology*, ed. T. Huang (Totowa, NJ: Humana Press), 67–95. doi: 10.1007/978-1-4939-7717-8_5

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018b). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018c). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 7, 1–6.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/nar/gkp1137

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., et al. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* 11(Suppl. 4):S7. doi: 10.1186/1471-2164-11-S4-S7

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* 8:e85024. doi: 10.1371/journal.pone.0085024

Esposito, A., Criscitiello, C., Trapani, D., and Curigliano, G. (2017). The emerging role of "Liquid Biopsies," circulating tumor cells, and circulating cell-free tumor DNA in lung cancer diagnosis and identification of resistance mutations. *Curr. Oncol. Rep.* 19:1.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces usingPhred. II. error probabilities. *Genome Res.* 8, 186–194. doi: 10.1101/gr.8.3.186

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12.

Newman, A. M., Bratman, S. V., To, J., Wynne, J. F., Eclov, N. C. W., Modlin, L. A., et al. (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* 20, 548–554. doi: 10.1038/nm.3519

Phallen, J., Sausen, M., Adleff, V., Leal, A., Hruban, C., White, J., et al. (2017). Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* 9:eaan2415.

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., Loeb, L. A., et al. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14508–14513.

Conflict of Interest: JY, QL, WW, LS, XS, GZ, GT, and JL were employed by the company Geneis Beijing Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 31 July 2020
doi: 10.3389/fbioe.2020.00913

MicroRNA-126 Modulates Palmitate-Induced Migration in HUVECs by Downregulating Myosin Light Chain Kinase via the ERK/MAPK Pathway

*Yi Wang[1,2]†, Mei Wang[3]†, Pei Yu[2], Li Zuo[2], Qing Zhou[2], Xiaomei Zhou[3] and Huaqing Zhu[2]**

[1] Department of Biological Engineering, School of Life Sciences, Anhui Medical University, Hefei, China, [2] Laboratory of Molecular Biology and Department of Biochemistry, Anhui Medical University, Hefei, China, [3] General Department of Hyperbaric Oxygen, Hefei Hospital Affiliated to Anhui Medical University, Hefei, China

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Mei Zhu,
Anhui Agricultural University, China
Jidong Lang,
Geneis (Beijing) Co., Ltd., China

***Correspondence:**
Huaqing Zhu
aydzhq@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

***Received:** 29 May 2020*
***Accepted:** 15 July 2020*
***Published:** 31 July 2020*

Citation:
Wang Y, Wang M, Yu P, Zuo L, Zhou Q, Zhou X and Zhu H (2020) MicroRNA-126 Modulates Palmitate-Induced Migration in HUVECs by Downregulating Myosin Light Chain Kinase via the ERK/MAPK Pathway. Front. Bioeng. Biotechnol. 8:913. doi: 10.3389/fbioe.2020.00913

MicroRNA-126 (miR-126) is an endothelial-specific microRNA that has shown beneficial effects on endothelial dysfunction. However, the underlying molecular mechanism is unclear. The present study evaluated the effects of miR-126 on the cell migration and underlying mechanism in HUVECs treated with palmitate. The present results demonstrated that overexpression of miR-126 was found to decrease cell migration in palmitate-treated HUVECs, with decreased MLCK expression and subsequent decreased phosphorylated MLC level. miR-126 also decreased the phosphorylation of MYPT1 in palmitate-treated HUVECs. In addition, it was demonstrated that miR-126 decreases expression of the NADPH oxidase subunits, p67 and Rac family small GTPase 1 with a subsequent decrease in cell apoptosis. Moreover, the phosphorylation of ERK was reduced by miR-126 in palmitate-induced HUVECs. Taken together, the present study showed that the effect of miR-126 on cell migration and cell apoptosis is mediated through downregulation of MLCK via the ERK/MAPK pathway.

Keywords: microRNA-126, cell migration, myosin light chain kinase, endothelial dysfunction, ERK/MAPK pathway

INTRODUCTION

Atherosclerosis (AS) is a chronic progressive pathological process characterized by multiple factors. Specifically, endothelial dysfunction is the earliest step in the pathogenesis of AS (Gimbrone and García-Cardeña, 2016). Palmitate, a main component of saturated fat, is associated with increased cardiovascular disease risk. Furthermore, clinical and experimental studies have demonstrated that high concentrations of free fatty acid (FFA) in the plasma, promotes endothelial dysfunction (Arijit et al., 2017). Palmitate increased monocyte expression of CD11b, which was associated with increased adhesion to rat aortic endothelium and CD36 expression, which promoted oxidized LDL uptake (Gao et al., 2012).

The phosphorylation of myosin regulatory light chain (MLC) has an essential role in the control of actomyosin contractility participates in cell contraction, cell adhesion, cell migration and epithelial barrier formation. Myosin light chain kinase (MLCK) induced the phosphorylation of MLC which activated by Ca^{2+}-calmodulin, is important in stress fiber formation and cell contractility. Aberrant expression of MLCK was shown to promote the progression of numerous inflammatory diseases, including pancreatitis, respiratory diseases, cardiovascular diseases, cancer and inflammatory bowel disease (Kim et al., 2012; Yu et al., 2018; Wang et al., 2019).

MicroRNAs (miRNAs) are a class of small 18–22 nucleotide, non-coding, single-stranded RNA molecules that regulate gene expression at the post-transcriptional level by binding to target mRNA. It was well-known that abnormal expression of miRNAs have been closely linked to the progression of AS by regulating endothelial cell function, lipid accumulation and vascular cells proliferation (Wojciechowska et al., 2017; Hung et al., 2018). Furthermore, miR-126, a miRNA specific for endothelial cells mediates vascular development and angiogenesis. Circulating level of miR-126 was decreased in the coronary artery disease (CAD) patients compared with healthy control (Wang et al., 2017). Previous study has shown that miR-126 play a protective role in human cardiac microvascular endothelial cells from hypoxia/reoxygenation-induced injury and inflammatory response by increasing NO secretion (Yang et al., 2017). In addition to vascular changes, miR-126 also modulates inflammation, and regulate lipid metabolism in endothelial cells (Yuan et al., 2016). Our previous study showed that miR-126 serves an anti-apoptotic role in palmitate-treated human umbilical vein endothelial cells (HUVECs) by decreasing the production of reactive oxygen species (ROS) (Wang et al., 2015). However, the role of miR-126 in the MLCK expression is unclear. Therefore, this study was aimed to analyze the effect exerted by miR-126 in MLCK expression in palmitate treated HUVECs, as well as the underlying mechanisms.

MATERIALS AND METHODS

Reagents and Antibodies

Palmitate and oleate were obtained from Sigma-Aldrich. DMEM medium was obtained from Gibco. FBS was purchased from the Zhejiang Tianhang Biological Technology. Anti-MLCK, anti-MLC, anti-MYPT1, anti-p-MYPT1, anti-ERK, anti-p-ERK and anti-β-actin antibodies were purchased from Santa Cruz Biotechnology. Anti-pMLC was obtained from Cell Signaling Technology. Anti-NOXA2/p67phox and anti-Rac family small GTPase 1 were purchased from Abcam.

Cell Culture and Transfection

HUVECs were cultured at 37°C in DMEM supplemented with 10% FBS at 5% CO_2 incubator. miR-126 mimic, miR-126antagomir and a scrambled oligonucleotide (Qiagen GmbH) were transfected with TransMessenger Transfection Reagent (Qiagen GmbH) according to the instructions. After 24 h, cell medium was changed with palmitate or oleate supplementation for 24 h. Oleate, as an unsaturated fatty acid exerts beneficial effects on endothelial dysfunction, was used as a control (Grenon et al., 2012). Each experiment was repeated a minimum of three times.

miR-126 Expression Assay

Total RNA from HUVECs was extracted by using TRIzol reagent (Invitrogen; Thermo Fisher Scientific, Inc.). A miRNA plate assay kit (Signosis, Inc.) and an oligo mix specific for miR-126 (Signosis, Inc.) were used to detect miR-126 expression following the manufacturer's protocol. The U6 small nucleolar RNA was chosen as an endogenous reference of miR-126. miR-126 expression was also assessed by reverse transcription-quantitative polymerase chain reaction (RT-qPCR), the sequences of the primers were as follows: miR-126, forward 5′-UCGUACCGUGAGUAAUAAUGCG-3′, reverse 5′-CAUUAUUACUCACGGUACGAUU-3′, and U6, forward 5′-CTCGCTTCGGCAGCACA-3′, and reverse 5′-AACGCTTCACGAATTTGCGT-3′. The reaction procedure was as follows: pre-denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 15 s, 65°C for 15 s, and 72°C for 15 s.

Scratch Healing Assay

HUVECs were seeded in 12-well plates and transfected as indicated, a total of 24 h later, palmitate or oleate were incubated with HUVECs. When the cells reached confluence, a sterile 200-μl pipette tip was used to create a horizontal wound in the confluent monolayer. Photographs of scratch wounds were captured prior to stimulation (0 h) and 24 h after incubation. The initial distance (0 h) and the distance traveled by cells after 24 h as detected using a microscope (Olympus). The percentage of wound healing was calculated using Image J (National Institutes of Health).

Cell Apoptosis

Cells underwent transfection for 48 h, followed by digestion and centrifugation to remove the supernatant. The cells were then washed with PBS and centrifuged again. The apoptotic rate of HUVECs was detected using annexin V-allophycocyanin apoptosis detection kit (Beyotime) following the manufacturer's instructions.

Western Blot Analysis

Protein samples were extracted from cultured cells. Cells were lysed using RIPA buffer. The protein concentrations were determined using a BCA protein assay kit (Beyotime Institute of Biotechnology). The protein was separated by SDS-PAGE, then transferred to polyvinylidene fluoride membranes. The membrane was blocked with 5% non-fat dry milk solution at room temperature for 2 h, and followed by incubation with the following primary antibodies (anti-MLCK, anti-MLC, anti-pMLC, anti-MYPT1, anti-pMYPT1, anti-ERK, anti-p-ERK, anti-p67phox, anti-Rac1, and anti-β-actin, all were used at a dilution of 1:1,000.) overnight. The 2 day, the membranes incubation with secondary antibodies at room temperature for 2 h and visualized with a Super Signal West Pico kit (Thermo Fisher

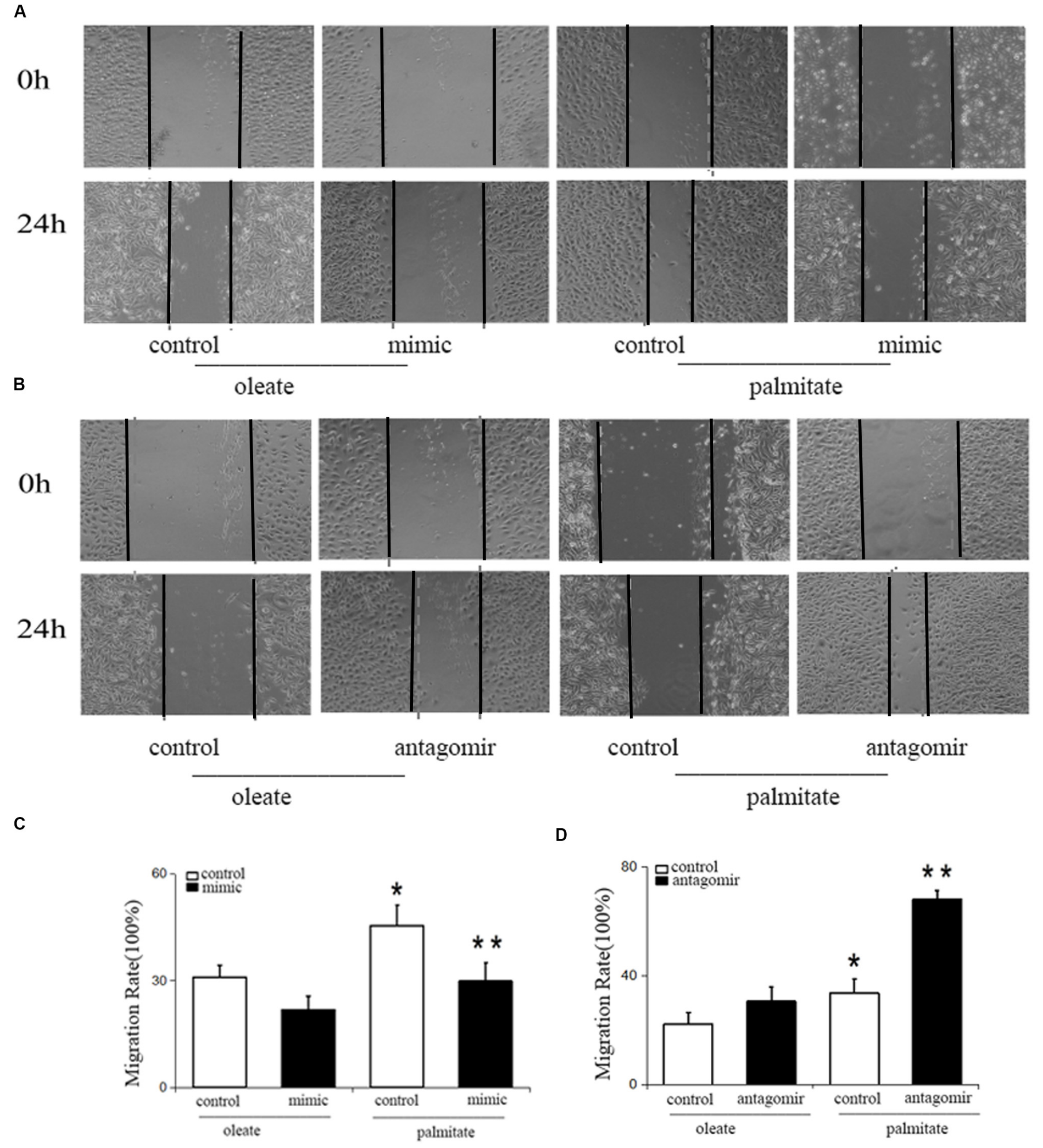


FIGURE 1 | Effect of miR-126 on the migration of palmitate-treated HUVECs. **(A,C)** The migration rate was increased in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 mimic inhibited the migration rate of palmitate-treated HUVECs. **(B,D)** miR-126 antagomir increased the migration rate of palmitate-treated HUVECs. *$P < 0.05$ vs oleate-treated HUVECs; **$P < 0.05$ vs palmitate-treated HUVECs.

Scientific, Inc.). Data were quantified through densitometry using Quantity One software.

Statistical Analysis

All data are reported in the form of mean ± standard deviation. The significant differences among groups were carried out with ANOVA followed by Newman-Keuls test for multiple comparisons. *P*-values below 0.05 were considered to indicate a statistically significant difference. Statistical analyses were carried out using SPSS 17.0.

RESULTS

miR-126 Reduces Palmitate-Induced Migration (motility) in HUVECs

To study the influence of miR-126 on cell migration in palmitate-treated HUVECs, the present study performed wound healing scratch assays. Following treatment with palmitate, the migration distances of HUVECs were found to be significantly longer compared with the oleate-treated HUVECs 24 h after injury. However, miR-126 mimic decreased the migration distances in

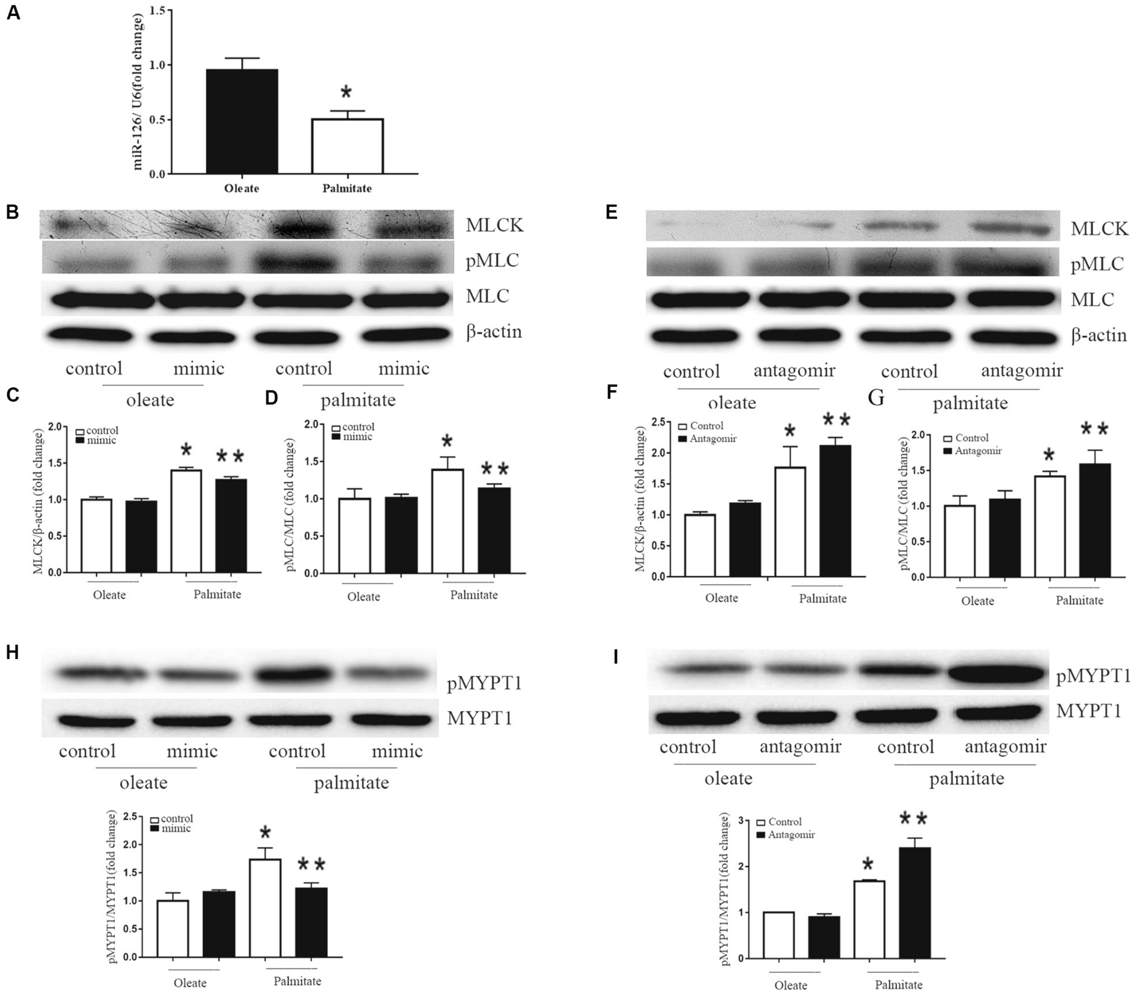


FIGURE 2 | Effect of miR-126 on the expression of MLCK, p-MLC and p-MYPT1 in palmitate-treated HUVECs. **(A)** Palmitate inhibited the expression of miR-126 in HUVECs compared with oleate. **(B–D)** The expression level of MLCK was upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs, and the p-MLC/MLC ratio in palmitate-treated HUVECs was upregulated, which all inhibited by miR-126 mimic. **(E–G)** The expression of level of MLCK and p-MLC/MLC ratio in palmitate-treated HUVECs was enhanced by miR-126 antagomir. **(H)** The pMYPT1/MYPT1 ratio was upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs which inhibited by miR-126 mimic. **(I)** The pMYPT1/MYPT1 ratio in palmitate-treated HUVECs was enhanced by miR-126 antagomir. *n* = 3. **P* < 0.05 vs oleate-treated HUVECs; ***P* < 0.05 vs palmitate-treated HUVECs.

palmitate-treated HUVECs compared with the control group (transfected with a scrambled oligonucleotide). By contrast, miR-126 antagomir further increased the migration distances in palmitate-treated HUVECs compared with the control group (**Figure 1**). The results indicated that miR-126 reduced cell migration in palmitate-treated HUVECs.

miR-126 Reduces MLC Phosphorylation by Regulating the Expression of MLCK and MYPT1 Phosphorylation in Palmitate-Treated HUVECs

To investigate the effect of palmitate on miR-126 expression in HUVECs, HUVECs were incubated with palmitate or oleate (0.1 mM) as a control. After 24 h, miR-126 expression was determined in HUVECs using the miRNA plate assay. As presented in **Figure 2A** that palmitate significantly decreased miR-126 expression. To measure the influence of miR-126 on MLCK expression, the level of MLCK was analyzed in palmitate-treated HUVECs transfected with a miR-126 mimic or miR-126 antagomir. Palmitate induced a markedly increased MLCK protein expression in HUVECs compared with oleate-treated cells. Furthermore, overexpression of miR-126 decreased MLCK protein expression in palmitate-treated HUVECs, while downregulation of miR-126 had the opposite effect (**Figures 2B,C**). In addition, MLCK is known to catalyze MLC phosphorylation. The present results demonstrated that the expression of MLC in palmitate-treated HUVECs was not significantly different compared with oleate-treated HUVECs, However, the phosphorylated MLC/MLC ratio in palmitate-treated HUVECs was increased compared with oleate-treated HUVECs, andmiR-126 mimic

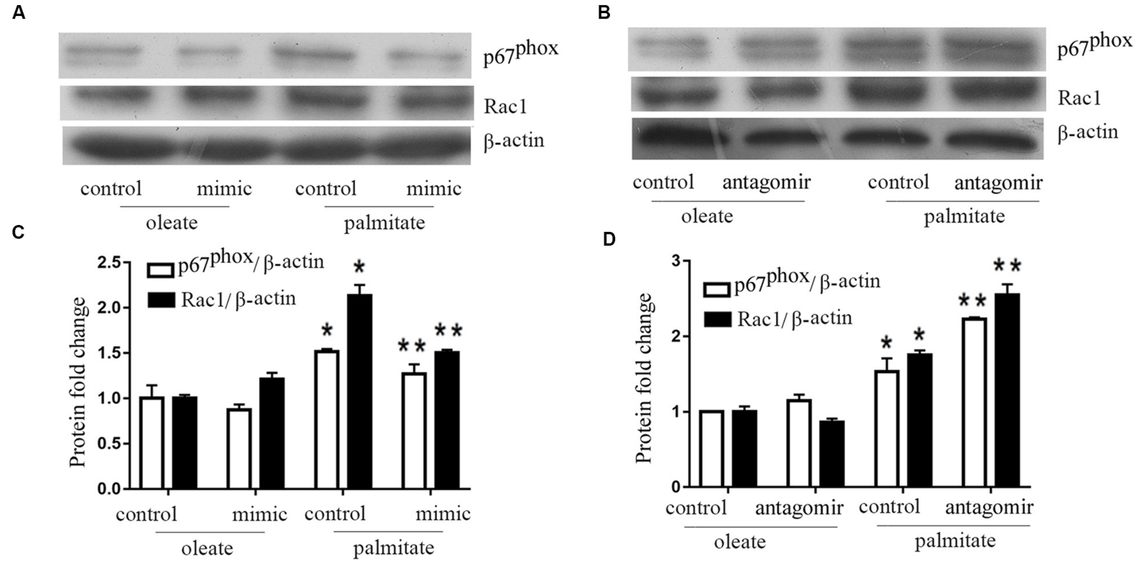


FIGURE 3 | miR-126 modulates the expression of p67phox and Rac1 in palmitate-treated HUVECs. **(A,C)** The expression level of p67phox and Rac1 were upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 mimic inhibited the expression level of p67phox and Rac1 in palmitate-treated HUVECs. **(B,D)** miR-126 antagomir enhanced the expression level of p67phox and Rac1 in palmitate-treated HUVECs. *n* = 3. **P* < 0.05 vs oleate-treated HUVECs; ***P* < 0.05 vs palmitate-treated HUVECs.

significantly decreased MLC phosphorylation (**Figures 2B,D**). By contrast, miR-126 antagomir significantly increased MLC phosphorylation (**Figures 2E,G**). To determine whether miR-126 regulates MLC phosphorylation through the phosphorylation of MYPT1, the levels of pMYPT1 and MYPT1 were quantified. The results demonstrated that miR-126 mimic significantly increased MYPT1 phosphorylation, whereas miR-126 antagomir significantly decreased MYPT1 phosphorylation (**Figures 2H,I**). These results indicated that miR-126 reduced the phosphorylation of MLC by regulate the expression of MLCK and the phosphorylation of MYPT1 in palmitate-treated HUVECs.

miR-126 Reduces NADPH Oxidase Subunits Rac1 and P67phox Expression in Palmitate-Treated HUVECs

Previous studies have shown that NADPH oxidase activity was is influenced by MLCK. Therefore, NADPH oxidase expression levels were quantified in the present study. The results revealed that palmitate increased the expression levels of Rac1 and p67phox (subunits of NADPH oxidase 2) in HUVECs compared with oleate-treated HUVECs (**Figure 3**). Furthermore, overexpression of miR-126 decreased Rac1 and p67phox protein expression in palmitate-treated HUVECs, while downregulation of miR-126 had the opposite effect. These results indicated that miR-126 reduced the expression of Rac1 and p67phox in palmitate-treated HUVECs.

miR-126 Reduces Apoptosis in Palmitate-Treated HUVECs

We investigated the effect of miR-126 on the regulation of HUVECs apoptosis induced by palmitate. As presented in **Figure 4**, the number of apoptotic cells in palmitate-treated HUVECs was higher than those observed in oleate-treated HUVECs. However, upregulation of miR-126 significantly alleviated apoptosis in palmitate-treated HUVECs, while downregulation of miR-126 further increased the number of apoptotic cells in palmitate-treated HUVECs. The data indicated that miR-126 inhibited apoptosis in palmitate-treated HUVECs.

miR-126 Attenuates Activation of ERK in Palmitate-Treated HUVECs

The ERK/MAPK pathway has been reported to be associated with endothelial dysfunction. Next, we investigated whether miR-126 participates in regulating the ERK/MAPK pathway in palmitate-induced HUVECs. miR-126 mimic, antagomir or a scrambled oligonucleotide were transfected into HUVECs, respectively, a total of 24 h later, HUVECs were exposed to palmitate or oleate (0.1 mM) for a further 24 h. As shown in **Figure 5**, phosphorylated ERK/ERK ratio was upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs, overexpression of miR-126 attenuated the phosphorylated ERK/ERK ratio in palmitate-treated HUVECs, whereas downregulation of miR-126 increased the phosphorylated ERK/ERK ratio in palmitate-treated HUVECs. The results indicated that miR-126 modulate activation of ERK in palmitate-treated HUVECs.

DISCUSSION

The endothelial cells (ECs) dysfunction in AS is characterized by increased cellular migration, apoptosis and enhanced permeability of the endothelial cell monolayer, allowing passage

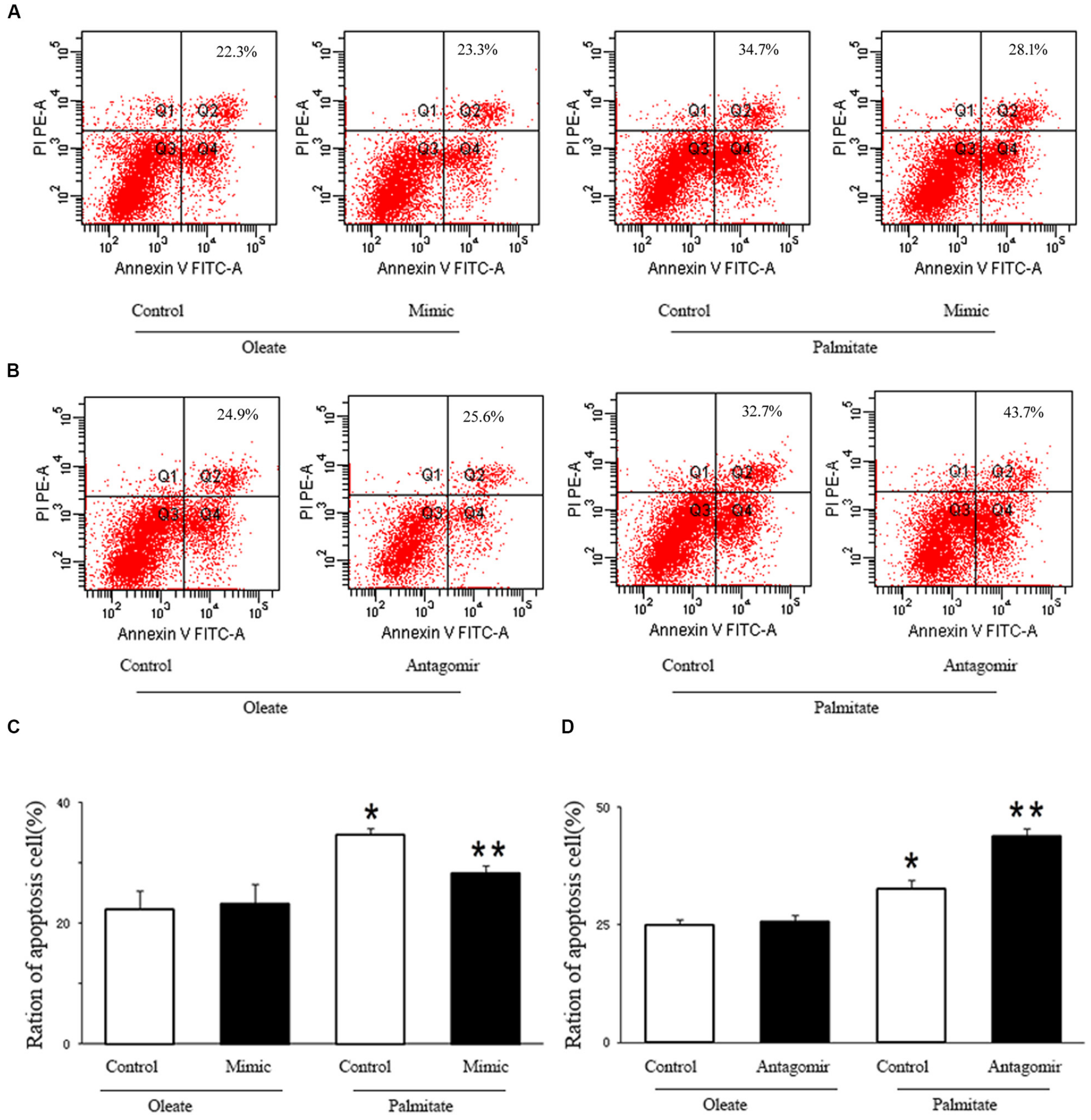


FIGURE 4 | miR-126 affected palmitate-induced HUVECs apoptosis as assessed by FACS. **(A,C)** FACS was used to detect HUVECs apoptosis, HUVECs apoptosis rate was increased in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 mimic decreased cell apoptosis in palmitate-treated HUVECs. **(B,D)** FACS was used to detect HUVECs apoptosis, HUVECs apoptosis rate was increased in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 antagomir promoted cell apoptosis in palmitate-treated HUVECs. *n* = 3. **P* < 0.05 vs oleate-treated HUVECs; ***P* < 0.05 vs palmitate-treated HUVECs.

of lipids and inflammatory factors (Chistiakov et al., 2016; Hu et al., 2020). miR-126 is the most prominent miRNA in ECs, abnormal expression of miR-126 may contributing to the pathogenesis of AS. *In vivo* study showed that miR-126 is essential for maintaining vascular integrity by involved in endothelial cell migration, disruption of cytoskeletal structure and cell apoptosis (Huveneers et al., 2015). The present study, demonstrated that the presence of miR-126 ameliorated cell migration and cell apoptosis, and subsequently reduced the expression of MLCK in HUVECs that had previously been treated with palmitate.

Furthermore, the endothelial cytoskeleton contractile machinery has an important role in maintaining barrier properties of the endothelium (Kása et al., 2015). Additionally, MLCK is a specifically enzyme via catalyze the phosphorylation of MLC mediate the reorganization of the cytoskeleton, leading to the disruption of endothelial barrier integrity (Rossi et al., 2011).

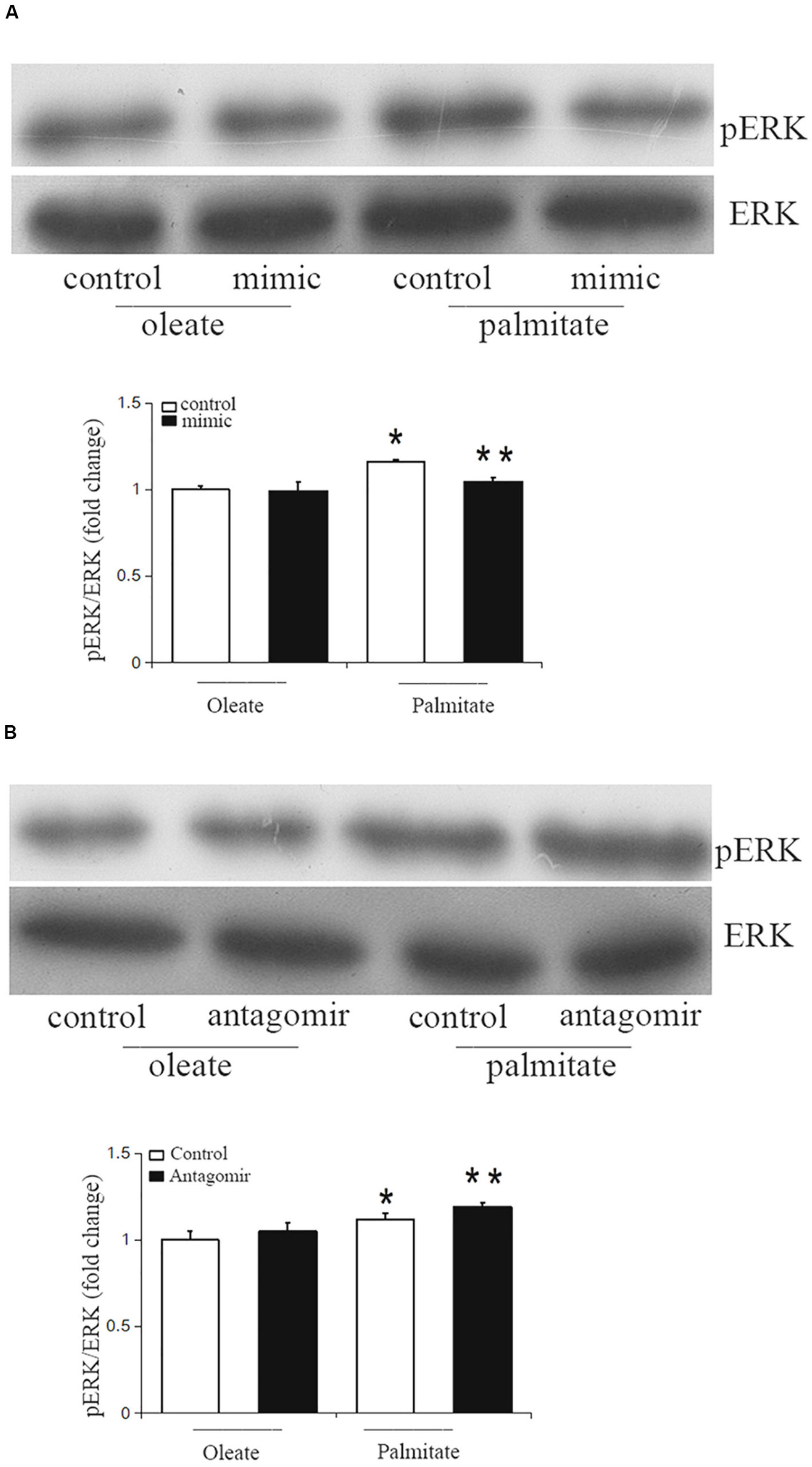


FIGURE 5 | miR-126 modulates the phosphorylation of ERK in palmitate-treated HUVECs. **(A)** Western blot results showed that phosphorylated ERK/ERK ratio was upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 mimic decreased the phosphorylated ERK/ERK ratio in palmitate-treated HUVECs. **(B)** Western blot results showed that phosphorylated ERK/ERK ratio was upregulated in palmitate-treated HUVECs compared with oleate-treated HUVECs, miR-126 antagomir increased the phosphorylated ERK/ERK ratio in palmitate-treated HUVECs. *n* = 3. *$P < 0.05$ vs oleate-treated HUVECs; **$P < 0.05$ vs palmitate-treated HUVECs.

The compromised endothelial barrier become more permeable to lipids and immune cells, ultimately leading to AS lesion formation (Schnittler, 2016). The present results indicated that miR-126 attenuates the expression of MLCK and decreases phosphorylation of MLC in palmitate-treated HUVECs. Dephosphorylation of MLC is accomplished by MLCP, which is consisting of MYPT1, a myosin target subunit, and a subunit with uncertain function. Dephosphorylation of MLC results in cell relaxation. The catalytic activity of MLCP are inhibited by the phosphorylation of MYPT1 at multiple sites by several kind of kinases, therefore, led to decreased dephosphorylated MLC and thus, vascular smooth muscle contraction (Qiao et al., 2014; Chang et al., 2016; Gao et al., 2017; Deng et al., 2020). The current study found that miR-126 decreased phosphorylation levels of MYPT1 in palmitate-treated HUVECs, which may account for the observed decrease in MLC phosphorylation. Therefore, miR-126 may modulate the phosphorylation of MLC via the regulation of MLCK expression and MLCP activity.

Previous studies have found that knockout of MLCK reduces the level of oxidized low-density lipoprotein (oxLDL)-induced endothelial hyper-permeability and also reduced the size of aortic lesions. Notably, it was also identified that MLCK acts through both MLC phosphorylation-coupled and -uncoupled pathways (Shen et al., 2010). Furthermore, Usatyuk et al. demonstrated that upregulation of MLCK in human pulmonary artery endothelial cells enhances the activation of endothelial NADPH oxidase and enhances ROS production (Usatyuk et al., 2012). Alternatively, downregulation of MLCK significantly inhibits NADPH oxidase, resulting in subsequently reduced ROS production. NADPH oxidase are important enzymes that regulate ROS generation in the vasculature. NADPH oxidase inhibitors have been shown to attenuate palmitate-induced excessive production of ROS within animal models (Li et al., 2019). Therefore, it was postulated that miR-126 could regulate NADPH oxidase expression by the regulation of MLCK. The current study examined the level of cellular apoptosis, as well as the expression of NADPH oxidase 2 subunits, such as p67phox and Rac1. The results indicated that miR-126 ameliorated cell apoptosis and reduced expression of p67phox and Rac1 in palmitate-treated HUVECs. These results were consistent with our previous study showing that miR-126 reduced ROS production in palmitate-treated HUVECs (Wang et al., 2015).

The MAPK signaling pathway has a vital function in the pathogenesis of AS. Chandra S et al. reported that high glucose induced endothelial dysfunction are mediated by the ERK/MAPK pathway (Chandra et al., 2019). Recently, a study have demonstrated that the ERK/MAPK pathway contribute to the activation of MLCK (Tan et al., 2014). Therefore, a previous study has shown that overexpression of miR-126 suppressed ERK pathway activity in glioma cells and resulted in the inhibition of glioma cell proliferation and invasion (Li et al., 2017; Wang et al., 2020). In the current study, we found that miR-126 could affect the activity of MLCK contributed to endothelial dysfunction via the ERK/MAPK pathway. The present data revealed that ERK activation was decreased in the presence of miR-126 mimic and increased by miR-126 antagomir. These results suggest that miR-126 may modulate MLCK expression through the ERK/MAPK signaling pathway.

In conclusion, the present results suggest a role for miR-126 in palmitate-treated HUVECs cell migration by causing a downregulation in MLCK via ERK activation. Understanding how miR-126 regulate endothelial cell migration and cell apoptosis provide important insights into the development of AS.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

YW and HZ study conception and design. YW, PY, and QZ acquisition of data. YW, MW, XZ, and LZ analysis and interpretation of the data. YW and HZ manuscript drafting and revision. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Key Research and Development Project in Anhui Province (No. 202004b11020025) and the Grant for Scientific Research of BSKY (No. XJ201617) from Anhui Medical University and the National Natural Science Foundation of China (Nos. 81570419 and 81270372).

REFERENCES

Arijit, G., Gao, L., Thakur, A., Parco, M. S., and Lai, C. W. K. (2017). Role of free fatty acids in endothelial dysfunction. *J. Biomed. Sci.* 24:50. doi: 10.1186/s12929-017-0357-5

Chandra, S., Fulton, D. J. R., Caldwell, R. B., Caldwell, R. W., and Toque, H. A. (2019). Hyperglycemia-impaired aortic vaso relaxation mediated through arginase elevation: role of stress kinase pathways. *Eur. J. Pharmacol.* 844, 26–37. doi: 10.1016/j.ejphar.2018.11.027

Chang, A. N., Kamm, K. E., and Stull, J. T. (2016). Role of myosin light chain phosphatase in cardiac physiology and pathophysiology. *J. Mol. Cel. Cardiol.* 101, 35–43. doi: 10.1016/j.yjmcc.2016.10.004

Chistiakov, D. A., Orekhov, A. N., and Bobryshe, Y. V. (2016). The role of miR-126 in embryonic angiogenesis, adult vascular homeostasis, and vascular repair and its alterations in atherosclerotic disease. *J. Mol. Cell. Cardiol.* 97, 47–55. doi: 10.1016/j.yjmcc.2016.05.007

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., and Chen, P. (2020). Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms2107 2274

Gao, D., Pararasa, C., Dunston, C. R., Bailey, C. J., and Griffiths, H. R. (2012). Palmitate promotes monocyte atherogenicity via de novo ceramide synthesis. *Free. Radic. Biol. Med.* 53, 796–806. doi: 10.1016/j.freeradbiomed.2012. 05.026

Gao, N., Tsai, M. H., Chang, A. N., He, W., Chen, C. P., Zhu, M., et al. (2017). Physiological vs. Pharmacological signalling to myosin phosphorylation in airway smooth muscle. *J. Physiol.* 595, 6231–6247. doi: 10.1113/JP274715

Gimbrone, M. A. Jr., and García-Cardeña, G. (2016). Endothelial cell dysfunction and the pathobiology of atherosclerosis. *Circ. Res.* 118, 620–636. doi: 10.1161/CIRCRESAHA.115.306301

Grenon, S. M., Aguado-Zuniga, J., Hatton, J. P., Owens, C. D., Conte, M. S., and Hughes-Fulford, M. (2012). Effects of fatty acids on endothelial cells: inflammation and monocyte adhesion. *J. Surg. Res.* 177, 35–43. doi: 10.1161/ATVBAHA.117.309017

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for QSAR prediction. *IEEE. J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009

Hung, J., Miscianinov, V., Sluimer, J. C., Newby, D. E., and Baker, A. H. (2018). Targeting non-coding RNA in vascular biology and disease. *Front. Physiol.* l9:1655. doi: 10.3389/fphys.2018.01655

Huveneers, S., Daemen, M. J., and Hordijk, P. L. (2015). Between Rho (k) and a hard place: the relation between vessel wall stiffness, endothelial contractility, and cardiovascular disease. *Circ. Res.* 116, 895–908. doi: 10.1161/CIRCRESAHA.116.305720

Kása, A., Csortos, C., and Verin, A. D. (2015). Cytoskeletal mechanisms regulating vascular endothelial barrier function in response to acute lung injury. *Tissue Barriers* 3, 1–2. doi: 10.4161/21688370.2014.974448

Kim, K. M., Csortos, C., Czikora, I., Fulton, D., Umapathy, N. S., Olah, G., et al. (2012). Molecular characterization of myosin phosphatase in endothelium. *J. Cell Physiol.* 227, 1701–1708. doi: 10.1002/jcp.22894

Li, Y., Cifuentes-Pagano, E., DeVallance, E. R., DeJesus, D. S., Sahoo, S., Meijles, D. N., et al. (2019). NADPH oxidase 2 inhibitors CPP11G and CPP11H attenuate endothelial cell inflammation & vessel dysfunction and restore mouse hind-limb flow. *Redox. Biol.* 22:101143. doi: 10.1016/j.redox.2019.101143

Li, Y., Li, Y., Ge, P., and Ma, C. (2017). miR-126 regulates the ERK pathway via targeting KRAS to inhibit the glioma cell proliferation and invasion. *Mol. Neurobiol.* 54, 137–145. doi: 10.1007/s12035-015-9654-8

Qiao, Y. N., He, W. Q., Chen, C. P., Zhang, C. H., Zhao, W., Wang, P., et al. (2014). Myosin phosphatase target subunit 1 (MYPT1) regulates the contraction and relaxation of vascular smooth muscle and maintains blood pressure. *J. Biol. Chem.* 289, 22512–22523. doi: 10.1074/jbc.M113.525444

Rossi, J. L., Ralay, R. H., and Patel, F. (2011). Albumin causes increased myosin light chain kinase expression in astrocytes via p38 mitogen-activated protein kinase. *J. Neurosci. Res.* 89, 852–861. doi: 10.1002/jnr.22600

Schnittler, H. (2016). Contraction of endothelial cells: 40 years of research, but the debate still lives. *Histochem. Cell Biol.* 146, 651–656. doi: 10.1007/s00418-016-1501-0

Shen, Q., Rigor, R. R., Pivetti, C. D., Wu, M. H., and Yuan, S. Y. (2010).). Myosin light chain kinase in microvascular endothelial barrier function. *Cardiovasc. Res* 87, 272–280. doi: 10.1093/cvr/cvq144

Tan, J., Wang, Y., Xia, Y., Zhang, N., Sun, X., Yu, T., et al. (2014). Melatonin protects the esophageal epithelial barrier by suppressing the transcription, expression and activity of myosin light chain kinase through ERK1/2 signal transduction. *Cell Physiol. Biochem.* 34, 2117–2127. doi: 10.1159/000369656

Usatyuk, P. V., Singleton, P. A., Pendyala, S., Kalari, S. K., He, D., Gorshkova, I. A., et al. (2012). Novel role for non-muscle myosin light chain kinase (MLCK) in hyperoxia-induced recruitment of cytoskeletal proteins, NADPH oxidase activation, and reactive oxygen species generation in lung endothelium. *J. Biol. Chem.* 12, 9360–9375. doi: 10.1074/jbc.M111.294546

Wang, B., Wang, L., Zheng, C. H., and Xiong, Y. (2019). "Imbalance data processing strategy for protein interaction sites prediction," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Piscataway, NJ: IEEE.

Wang, W., Zhou, Y., Cheng, M. T., Wang, Y., Zheng, C. H., Xiong, Y., et al. (2020). "Potential pathogenic genes prioritization based on protein domain interaction network analysis," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Piscataway, NJ: IEEE.

Wang, X., Lian, Y., Wen, X., Guo, J., Wang, Z., Jiang, S., et al. (2017). Expression of miR-126 and its potential function in coronary artery disease. *Afr. Health Sci.* 17, 474–480. doi: 10.4314/ahs.v17i2.22

Wang, Y., Wang, F., Wu, Y., Zuo, L., Zhang, S., and Zhou, Q. (2015). microRNA-126 attenuates palmitate-induced apoptosis by targeting TRAF7 in HUVECs. *Mol. Cell Biochem.* 1, 123–130. doi: 10.1007/s11010-014-2239-4

Wojciechowska, A., Braniewska, A., and Kozar-Kamińska, K. (2017). MicroRNA in cardiovascular biology and disease. *Adv. Clin. Exp. Med.* 5, 865–874. doi: 10.17219/acem/62915

Yang, H. H., Chen, Y., Gao, C. Y., Cui, Z. T., and Yao, J. M. (2017). Protective effects of MicroRNA-126 on human cardiac microvascular endothelial cells against Hypoxia/Reoxygenation-Induced injury and inflammatory response by activating PI3K/Akt/eNOS signaling pathway. *Cell Physiol .Biochem.* 42, 506–518. doi: 10.1159/000477597

Yu, M., Wang, Q., Ma, Y., Li, L., Yu, K., and Zhang, Z. (2018). ArylHydrocarbon receptor activation modulates intestinal epithelial barrier function by maintaining tight junction integrity. *Int. J. Biol. Sci.* 14, 69–77. doi: 10.7150/ijbs.22259

Yuan, X., Chen, J., and Dai, M. (2016). Paeonol promotes microRNA-126 expression to inhibit monocyte adhesion to ox-LDL-injured vascular endothelial cells and block the activation of the PI3K/Akt/NF-κB pathway. *Int. J. Mol. Med.* 38, 1871–1878. doi: 10.3892/ijmm.2016.2778

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fbioe.2020.00701

A Deep Learning Framework to Predict Tumor Tissue-of-Origin Based on Copy Number Alteration

Ying Liang [1†], Haifeng Wang [2†], Jialiang Yang [3†], Xiong Li [4], Chan Dai [3], Peng Shao [1], Geng Tian [3], Bo Wang [3] and Yinglong Wang [1]*

[1] College of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang, China, [2] Department of Urology, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China, [3] Geneis (Beijing) Co. Ltd., Beijing, China, [4] School of Software, East China Jiaotong University, Nanchang, China

Cancer of unknown primary site (CUPS) is a type of metastatic tumor for which the sites of tumor origin cannot be determined. Precise diagnosis of the tissue origin for metastatic CUPS is crucial for developing treatment schemes to improve patient prognosis. Recently, there have been many studies using various cancer biomarkers to predict the tissue-of-origin (TOO) of CUPS. However, only a very few of them use copy number alteration (CNA) to trance TOO. In this paper, a two-step computational framework called CNA_origin is introduced to predict the tissue-of-origin of a tumor from its gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a convolution neural network (CNN). Based on real datasets released from the public database, CNA_origin had an overall accuracy of 83.81% on 10-fold cross-validation and 79% on independent datasets for predicting tumor origin, which improved the accuracy by 7.75 and 9.72% compared with the method published in a previous paper. Our results suggested that the autoencoder model can extract key characteristics of CNA and that the CNN classifier model developed in this study can predict the origin of tumors robustly and effectively. CNA_origin was written in Python and can be downloaded from https://github.com/YingLianghnu/CNA_origin.

Keywords: tumor, tissue-of-origin, copy number alteration, autoencoder, convolution neural network

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Guohua Huang,
Shaoyang University, China
Khanh N. Q. Le,
Taipei Medical University, Taiwan
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China

****Correspondence:***
Ying Liang
aliang1229@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *29 March 2020*
Accepted: *04 June 2020*
Published: *05 August 2020*

Citation:
Liang Y, Wang H, Yang J, Li X, Dai C, Shao P, Tian G, Wang B and Wang Y (2020) A Deep Learning Framework to Predict Tumor Tissue-of-Origin Based on Copy Number Alteration. Front. Bioeng. Biotechnol. 8:701. doi: 10.3389/fbioe.2020.00701

1. INTRODUCTION

Cancer metastasis is the process in which tumor cells fall off from the primary site, enter the circulatory system, transfer to other parts of the body, and continue to grow. In about 3–5% of metastatic tumors, the sites of origin cannot be found, and this is known as cancer of unknown primary site (CUPS). Patients diagnosed with CUPS are treated with broad-spectrum anticancer drugs and have a low median survival time of 9–12 months. Precise diagnosis of the tissue of origin for metastatic CUP is essential for deciding on the treatment scheme to improve the patient's prognosis (Chen et al., 2017). Clinical, imaging and pathological examination are used to detect the tissue of origin, but these approaches can only determine the tissue of origin in about 50–80% of CUP patients.

Recently, a large number of studies have tried to use cancer biomarkers to predict the primary tumor site for CUPs so as to provide much-needed guidelines for timely patient care and cancer therapy (Liang et al., 2016; Grewal et al., 2019; Wang et al., 2019; Zheng et al., 2019). The gene expression patterns in tumors have high specificity, and so these the most widely used biomarkers for tumor classification (Bloom et al., 2004; Tothill et al., 2005; Staub et al., 2010; Wu et al., 2010; Handorf et al., 2013; Xu et al., 2016; Wang et al., 2018; Li et al., 2019). For example, Li used the within-sample relative gene expression orderings of gene pairs within individual samples to identify a prediction signature (Li et al., 2019). Wang proposed a general framework to identify a subset of genes for each tumor subtype and presented a corresponding classification model for distinguishing different tumor subtypes (Wang et al., 2018). Xu established a comprehensive database integrating microarray- and sequencing-based gene expression profiles of 16,674 tumor samples covering 22 common human tumor types to discriminate the origins of tumor tissue, which will be an additional useful tool for determining the tumor origin (Xu et al., 2016).

DNA methylation and miRNA regulate the expression of genes involved in numerous biological processes (Rosenfeld et al., 2008; Rosenwald et al., 2010; Ferracin et al., 2011; Mueller et al., 2011; Søkilde et al., 2014). Tang developed a user-friendly webserver to predict tumor origin by identifying highly tissue-specific CpG sites and miRNA expression (Tang et al., 2017). Bae tried to discover tissue-specific methylation markers and predicted the tissue-of-origin in CUPS (Bae et al., 2018). Yang proposed an inverse space sparse representation model to distinguish tumor origins considering the characteristics of gene-based tumor data (Yang et al., 2019). Visual imagery is one of the main methods used by pathologists to assess the stage, type, and subtype of tumors (Shi et al., 2016; Coudray et al., 2018; Mohsen et al., 2018). Coudray employed visual inspection of histopathology slides to classify lung adenocarcinoma, lung squamous cell carcinoma, and normal lung tissue, which achieved performance comparable to that of pathologists (Coudray et al., 2018). Ultrasound imaging can also be used for tumor detection and diagnosis with a deep polynomial network algorithm (Shi et al., 2016).

As yet, few studies have investigated the roles of genome variants on tissue-of-origin in CUPS. Genome variants include mutation, small insertion, and deletion (INEDL) and copy number alteration (CNA). CNA is amplification and deletion of genomic sequences ranging from kilobases (Kb) to megabases (Mb) in size, which covers 360 Mb and encompasses hundreds of genes, disease loci, and functional elements (Redon et al., 2006). As the main genetic marker of the genome, CNA can affect the gene function through gene dose, gene breakage, gene fusion, and position effects and is closely related to the occurrence and development of tumor (Poduri et al., 2013). CNA also plays an increasingly important role in targeted therapy, personalized treatment, and prognosis judgment for tumors. Marquard developed a tool named TumorTracer by using publicly available somatic mutation data to train random forest classifiers and thus to identify the tissue of origin. This was demonstrated to be accurate enough to aid in the clinical diagnosis of cancers with unknown primary origin (Marquard et al., 2015). Zhang conducted a comprehensive genome-wide analysis of CNAs from six cancer types and selected 19 discriminative genes for tumor classification, but their overall prediction accuracy was about 75% (Zhang et al., 2016). In the current study, a computational method called CNA_origin is proposed to predict the tissue of origin with the information of gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a convolution neural network (CNN). This predictor successfully learned the inherent information of gene copy number and exhibited superior performance to classical algorithms for the same benchmark datasets.

TABLE 1 | Number of samples per tissue for CNA profiles.

Primary site	Histology	CNA datasets
Breast	BRCA (Breast invasive carcinoma)	847
Colorectal	COADREAD (Colorectal adenocarcinoma)	575
Brain	GBM (Glioblastoma multiforme)	563
Kidney	KIRC (Kidney renal clear cell carcinoma)	490
Ovarian	OV (Ovarian serous cystadenocarcinoma)	562
Uterine	UCEC (Uterine Corpus Endometrial Carcinoma)	443

2. MATERIALS AND METHODS

2.1. Datasets

The copy number signal was produced by Affymetrix SNP 6.0 arrays for the set of samples in the cancer genome atlas (TCGA) study, as generated with the Firehose analysis pipeline. The preprocessing analysis of the dataset was performed with GISTIC (Beroukhim et al., 2007). These datasets were from primary solid tumor samples released by MSKCC in 2013 that could be downloaded from http://cbio.mskcc.org/cancergenomics/pancan_tcga/. The datasets with a sample size greater than 400 were selected. The details of all tissue samples, including tumor status, histopathology details, and sample sizes, are summarized in **Table 1**.

Each sample had 24,174 genes with discrete copy number values denoted as "–2," "–1," "0," "1," "2," where "–2" was homozygous deletion, "–1" was heterozygous loss, "0" was diploid, "1" was one copy gain and "2" was high-level amplification or multiple-copy gain (Ciriello et al., 2013). The CNA values were scaled to [–1, 1] with Equation (1).

$$x' = \frac{x}{|x|_{max}} \tag{1}$$

where x was the CNA value of the gene, $|x|_{max}$ was the maximum absolute value of CNA among samples, and x' was the value after correction.

2.2. Feature Extraction

Each sample had 24,174 gene-level CNA values. High dimensionality and small sample sizes have seriously obscured the intrinsic nature of CNA data. In this paper, CNA_origin applied a stacked autoencoder (SAE) to extract the features of CNA values, which converted the high-dimensional data into low-dimensional codes by training a multilayer neural network with small central layers to reconstruct high-dimensional input vectors (Hinton and Salakhutdinov, 2006). The SAE consisted of an adaptive multilayer "encoder" network and an asymmetric "decoder" network, and high-dimensional abstraction whilst maintaining the key information was achieved for feature reduction with the help of hidden nodes in the code layer, as illustrated in **Figure 1A**.

In the encoder network, the 24,174 gene-level CNA values used as inputs were mapped to the latent representation of next layer using Equation (2).

$$X^{[i]} = f(W_i X^{[i-1]} + b_i) \tag{2}$$

where $f(x) = max\,(0, x)$ was ReLU activation function, b_i was the bias of layer i, and W_i was the weight between layer i-1 and i. In the decoder network, the code layer was used to reconstruct the input by a reverse mapping using Equation (3).

$$X^{[i]} = f(W_i{}' X^{[i-1]} + b_i{}') \tag{3}$$

where $W_i{}' = W_i^T$. The tanh activation function $f\,(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ was added to predict the final value, and the dimensionality of the final output layer was the same as that of the input layer. To determine the optimized parameters of W and b, layer-by-layer pretraining was used to minimize the error between the input X and output X'. The middle features were extracted through hidden nodes in the code layer.

CNA_origin was implemented in Python 3.7.3 using Keras (2.24) with the backend of TensorFlow (1.14.0). For the feature extraction of gene CNA, the neuron numbers in symmetrical hidden layers were set at 4,096, 1,024, 256, 100, 256, 1,024, and 4,096, respectively. The middlemost 100 neurons represented the extracted features, as it was found that features with more than 100 dimensions were not helpful to improve the classifier performance. The initial learning rate was set to 0.01, batch size to 64, and epochs to 16. This autoencoder was optimized using the Adam algorithm to learn the model parameters, and the loss function was mean square error.

2.3. Classifier Construction

The fully connected layer learns the global patterns in feature space, but convolution layer applies filters in the form of convolution operations to learn local patterns from the image (Baek et al., 2018). Inspired by the visual world, CNN has two interesting properties, translation invariant and spatial hierarchies of patterns, which allow a convolution network to efficiently learn increasingly complex and abstract visual concepts (Chollet, 2015, 2017). These properties are specialized for image data and also show outstanding performance in sequence processing (Le et al., 2017, 2019b). The same input transformation was performed on every subsequence; a pattern learned at a certain position in a sequence was later recognized at a different position, making 1D convnets translation invariant. A 1D convolution layer could catch local patterns in a sequence, making it competitive with recurrent neural networks (RNN) on sequence-processing at a considerably cheaper computational cost.

CNA_origin reshaped the 100 features of the sample into a 100 × 1 vector; each input tensor was 100 in width, 1 in height, and 1 in depth. The 1D convolution was used to extract local subsequences with D filters, and each filter was of $k \times 1$ in size, which means the filter was k in width and 1 in height. CNA_origin utilized multi-scale convolution kernels, such as 1 × 1, 3 × 1, 5 × 1, 7 × 1, and 9 × 1, to extract high-order features of different levels and increase the diversity of feature extraction. Among them, the 1 × 1 convolution kernel changed the number of channels, increased the non-linear transformation of features, and improved the generalization ability of the network. The number 48 or 64 in parentheses behind $k \times 1$ meant convolution with 48 or 96 filters. CNA_origin padded the features by adding k/2 columns with elements being zero to the head and tail of the sequence; therefore, the width of the new sequence after convolution with stride 1 was still the same.

The Concat operation in **Figure 1** meant that the layer stacked features from each branch together. Different convolution layers and max-pooling layers concatenated like the Inception module, which increased the depth of the network and improved the robustness of the CNN. At the beginning of the network, a larger convolution kernel was used to reduce the number of parameters and computation, as illustrated in **Figure 1B**. In the last, the network connected two full connection layers, with a dropout layer to avoid overfitting. Usually, the number of hidden units was far larger than the obtained data, resulting in overfitting. The dropout layer helped alleviate this problem by removing some of the connections in the network (Baek et al., 2018). Output such as 50 × 1 × 128 meant that the feature maps were 50 in width, 1 in height, and 128 in depth. The final result was the probability that the sample belonged to each class and was found with the "softmax" activation function, which is often used in solving multi-classification problems. It was defined as Equation (4).

$$P_k = \frac{exp(\alpha_k)}{\sum_{i\,=\,1}^{m} exp(\alpha_i)} \tag{4}$$

P_k was the probability that the sample belonged to class k. exp(x) represented an exponential function, α_k was the input value of class k, and m was the number of tumor classes. The categorical cross-entropy loss corresponding with the "softmax" activation function was used, which was a variant of binary cross-entropy and was defined as Equation (5).

$$loss = -\sum_{i\,=\,1}^{n} y_{i1} logP_{i1} + y_{i2} logP_{i2} + \cdots + y_{im} logP_{im} \tag{5}$$

P_{im} was the predicted probability, n was the number of samples, and y_{im} was the true label.

For the classification learning, the number of multi-scale convolution kernels was set to 64, batch size to 16, and epochs to 12. The learning rate was dynamically adjusted according to the loss value of the test dataset, and the initial value was 0.01. The dropout rate was set to 0.4, and the loss function was sparse categorical crossentropy.

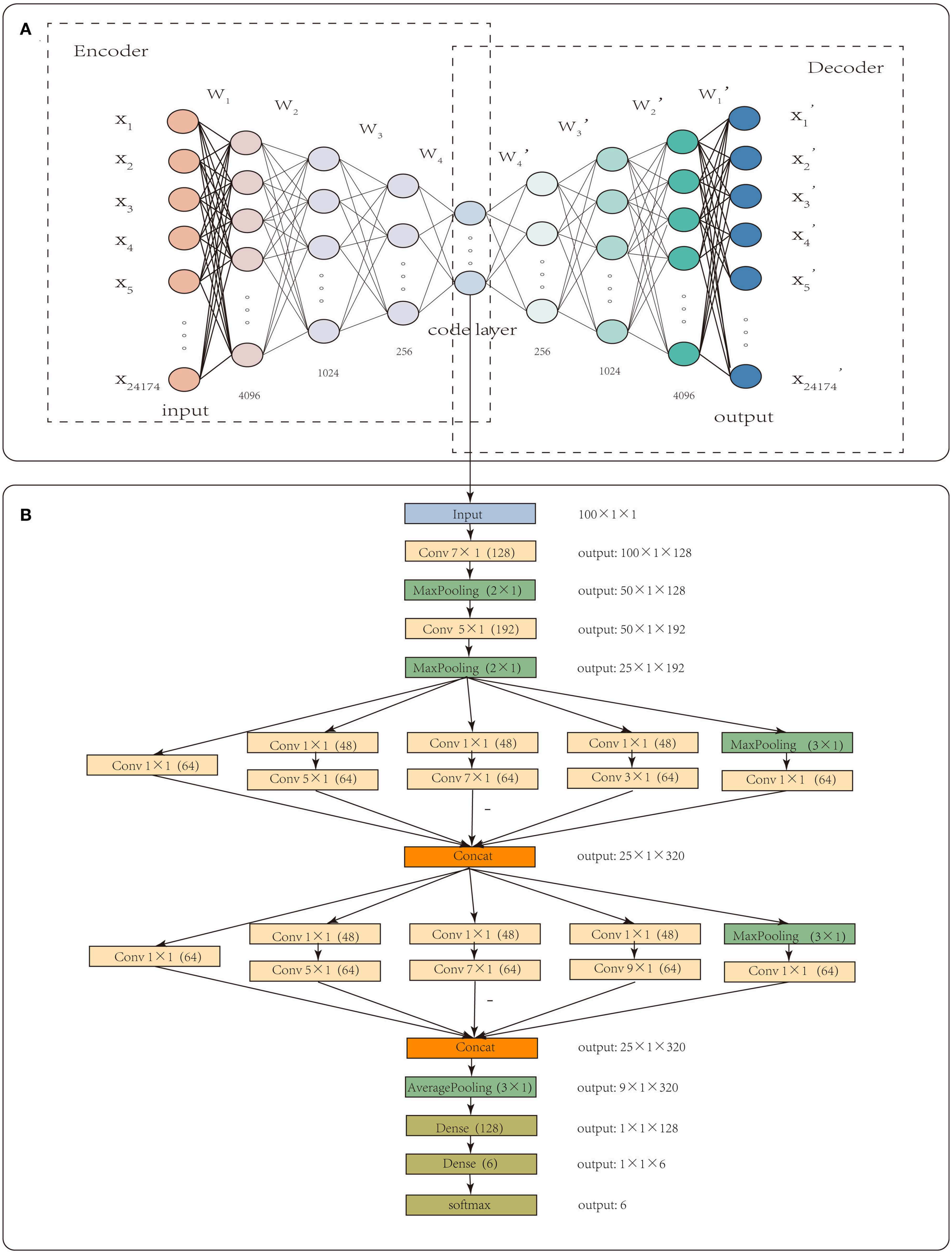


FIGURE 1 | The workflow of CNA_origin. CNA_origin applied a stacked autoencoder to extract the feature of CNA values, which was composed of a symmetrical encoder and decoder network, and 4,096, 1,024, and 256 were the neuron numbers in symmetrical hidden layers **(A)**. A 1D CNN with multi-scale convolution kernels (1 × 1, 3 × 1, 5 × 1, 7 × 1, 9 × 1) was used to construct a classifier model, and the number 48 or 64 in parenthesis behind *k* × 1 meant convolution with 48 or 96 filters. The Concat layer stacked features from each branch together; the output denoted the dimensions of feature maps for each layer **(B)**.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation Metrics

The six tumor datasets were used to train CNA_origin. To understand the generalization performance, CNA_origin was also tested by independent datasets. In this work, the precision (P), recall (R), accuracy (ACC), and F1-score were adopted to assess the performance of the corresponding method; they have been used as measurement metrics in previous works (Le et al., 2018, 2019a). They are defined as Equation (6).

$$\begin{aligned} P &= \frac{T_P}{T_P + F_P} \\ R &= \frac{T_P}{T_P + F_n} \\ ACC &= \frac{T_P + T_n}{T_P + F_p + F_n + T_n} \\ F1 - score &= \frac{2 \times P \times R}{P + R} \end{aligned} \qquad (6)$$

where T_P, T_n, F_P, and F_n were the numbers of true positives, true negatives, false positives, and false negatives, respectively. $P \in [0, 1]$, $R \in [0, 1]$, $ACC \in [0, 1]$, and $F1 - score \in [0, 1]$. P = 0 indicated that all predicted positive results were actually negative. When all results were incorrect, $T_P = 0$ and $T_n = 0$; therefore, P = 0, R = 0, ACC = 0, and F1-score = 0. When all results were correct, $F_P = 0$ and $F_n = 0$; therefore, P = 1, R = 1, ACC = 1, and F1-score = 1. Precision and recall are two contradictory metrics. Generally speaking, when the precision is high, the recall is often low, while when the recall is high, the precision is often low.

3.2. CNA_Origin Performance

Ten-fold cross-validation was utilized to evaluate our algorithm with the extracted 100-dimensional features. The datasets were randomly divided into ten subsets of approximately equal size. Our network was trained 10 times; nine of the 10 subsets were used as the training datasets, and the remaining one was the test dataset. All of the above evaluation indices of our algorithm, that is, P, R, ACC, and F1-score, were calculated according to the results in our work. The average values of four metrics P, R, ACC, and F1-score defined in Equation (6) over ten test datasets are listed in **Table 2**.

TABLE 2 | CNA_origin performance measured by three metrics via 10-fold cross-validation.

Cancer	Precision	Recall	F1-score
BRCA	0.8750	0.9231	0.8984
COADREAD	0.8158	0.7381	0.7750
GBM	0.9310	0.8438	0.8852
KIRC	0.8889	0.9600	0.9231
OV	0.8980	0.8672	0.8800
UCEC	0.6792	0.7200	0.6990

3.3. Performance Comparison With Other Algorithms

The performance of our algorithm was compared with four other classical classification algorithms with the same benchmark datasets. Random forest (RF) is an ensemble classifier that produces multiple decision trees using a randomly selected subset of training samples and variables (Liu et al., 2019). XGBoost is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning and has been used in many bioinformatics fields (Chen and Guestrin, 2016; Deng et al., 2020; Hu et al., 2020). Long Short-Term Memory (LSTM) is an artificial RNN architecture that is well-suited to classifying, processing, and making predictions based on time series data (Hochreiter and Schmidhuber, 1997). Zhang proposed a method to computationally classify cancer types by using CNA

TABLE 3 | Comparison of CNA_origin predictions with those of other algorithms.

Cancer	Predictor	Precision	Recall	F1-score
BRCA	CNA_origin	**0.8750**	**0.9231**	**0.8984**
	LSTM	0.8713	0.8462	0.8585
	RF	0.8556	0.8645	0.8601
	XGboost	0.8214	0.8846	0.8519
	CNA_zhang	0.7916	0.8735	0.8306
COADREAD	CNA_origin	0.8158	0.7381	0.7750
	LSTM	**0.8571**	**0.8077**	**0.8317**
	RF	0.7659	0.6923	0.7272
	XGboost	0.7959	0.7500	0.7723
	CNA_zhang	0.6000	0.7346	0.6605
GBM	CNA_origin	0.9310	0.8438	0.8852
	LSTM	0.8913	0.8913	0.8913
	RF	0.8627	0.8627	0.8627
	XGboost	**0.9535**	**0.8913**	**0.9213**
	CNA_zhang	0.8870	0.8593	0.8730
KIRC	CNA_origin	0.8889	**0.9600**	**0.9231**
	LSTM	0.8837	0.9268	0.9048
	RF	**0.9056**	0.8571	0.8807
	XGboost	0.8780	0.8780	0.8780
	CNA_zhang	0.8085	0.9268	0.8636
OV	CNA_origin	**0.8980**	0.8627	**0.8800**
	LSTM	0.7843	**0.9091**	0.8421
	RF	0.7826	0.9000	0.8372
	XGboost	0.7551	0.8409	0.7957
	CNA_zhang	0.8461	0.7586	0.8000
UCEC	CNA_origin	0.6792	**0.7200**	**0.6990**
	LSTM	0.6897	0.6557	0.6723
	RF	0.6451	0.6060	0.6250
	XGboost	0.7407	0.6557	0.6957
	CNA_zhang	**0.7419**	0.4693	0.5750

The bold values are the best performance among counterparts.

level values; this was denoted as CNA_zhang here because the authors did not give the method a name (Zhang et al., 2016). CNA_zhang used minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) to select features and the Dagging algorithm to give the final classification. The input of LSTM, RF, and XGboost was the extracted features from the autoencoder, and the GridSearchCV function in the sklearn package was used to select the optimal super-parameters that, were promised in the best condition.

Table 3 shows that the performance of CNA_origin was superior to LSTM, RF, XGboost, and CNA_zhang for BRCA, KIRC, OV, and UCEC. For BRCA, compared with LSTM and CNA_zhang, the F1-score was increased by 4.6 and 8.1%, respectively, and the recall (R) was increased by 9.08 and 5.67%, respectively. For GBM, CNA_origin performed slightly worse than the best, XGboost, with reductions of 2.35% in precision, 5.32% in recall, and 3.91% in F1-score. For KIRC, compared with LSTM and CNA_zhang, the F1-score was increased by 2.02 and 6.88%, respectively, and the recall was increased by 3.58%. For UCEC, compared with LSTM and CNA_zhang, the F1-score was increased by 3.97 and 21.56%, respectively, and the recall was increased by 9.80 and 53.41%, respectively. For COADREAD, CNA_origin performed slightly worse than the best LSTM algorithm, with reductions of 4.81% in precision, 8.61% in recall, and 6.81% in F1-score, respectively. For OV, the F1-score of CNA_origin was increased by 4.50% and 10.00% compared with LSTM and CNA_zhang; the recall was worse than the best, LSTM, by 5.10%, and precision was better than LSTM and CNA_zhang by 14.49 and 6.13%, respectively. CNA_origin exhibited perfect performance for the tumor classification.

The macro-averages of precision, F1-score, recall, and accuracy of six types of tumors were utilized to evaluate our predictor. Ten-fold cross-validation was run 100 times to test CNA_origin, LSTM, RF, XGboost, and CNA_zhang. For precision, CNA_origin had a mean value of 0.8369, which was increased by 0.70 and 6.87% compared with LSTM and CNA_zhang. For recall, the mean value of CNA_origin was 0.8345, which was increased by 0.91 and 8.68% compared with LSTM and CNA_zhang, respectively. For the F1-score, the mean value of CNA_origin was 0.8339, which was increased by 0.77 and 8.22% compared with LSTM and CNA_zhang, respectively. For accuracy, the CNA_origin had a mean value of 0.8381, which was increased by 0.92 and 7.75% compared with LSTM and CNA_zhang, respectively. The results are shown in **Figure 2**.

The results showed that the sensitivity, accuracy, and specificity of UCEC were significantly lower than those of other tumors. The results of UCEC were further analyzed, and it was found that about 48–76% of UCEC samples were predicted

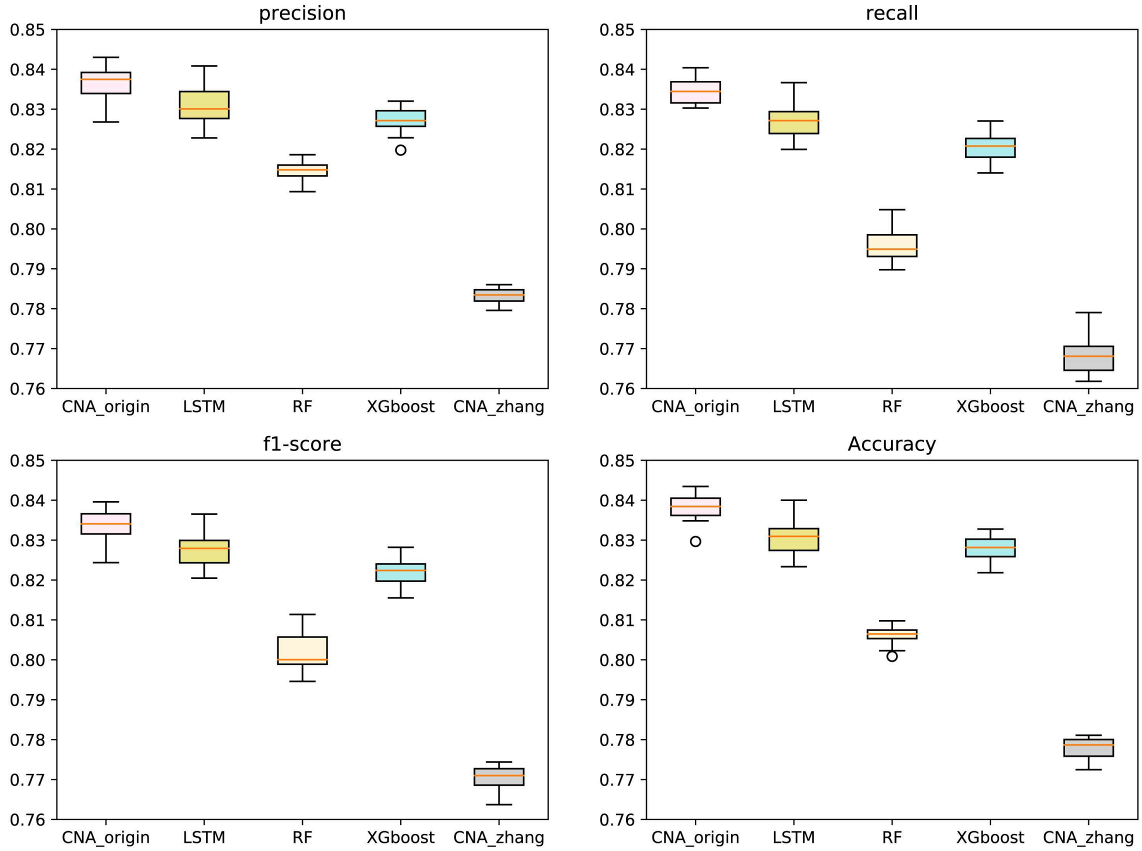


FIGURE 2 | Performance comparison between CNA_origin and other algorithms (basic LSTM, RF, XGboost, and CNA_zhang) for the macro-averages of precision, F1-score, recall, and accuracy from 10-fold cross-validation 100 times.

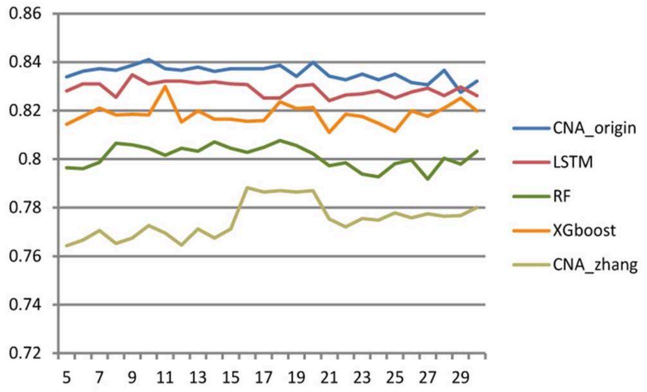


FIGURE 3 | Effect of cross-validation fold k value on classifier performance. When the value of k became larger, the performance of classifiers was improved, but a small sample size of the test set had a negative impact on model evaluation.

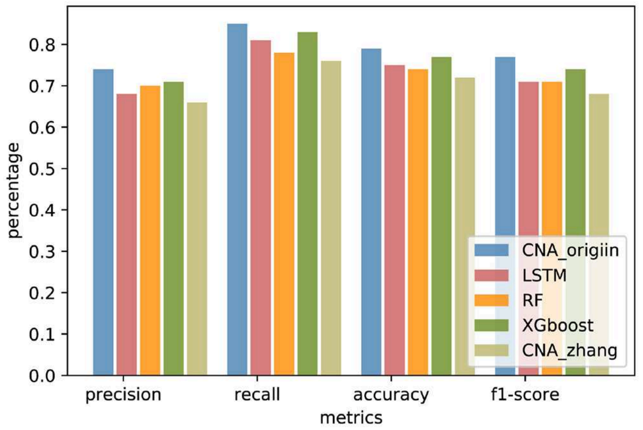


FIGURE 4 | Performance comparison of CNA_origin and other algorithms (basic LSTM, RF, XGboost, and CNA_zhang) for independent datasets from the TCGA.

to be OV, while 24–52% of UCEC samples were predicted to be BRCA. This may be because BRCA, OV, and UCEC are hormone-dependent tumors, which have a close relationship in tumorigenesis. Many reports have pointed out that BRCA, OV, and UCEC are related to changes in estrogen and estrogen receptors (Rodriguez et al., 2019; Scherbakov et al., 2019; Sehouli et al., 2019). Moreover, the physical location of ovary and uterus is very close, which may lead to contamination of tissue samples and difficulty in distinguishing UCEC from OV samples.

3.4. Impact of Sample Size

Different cross-validation fold k values were used to study the effect of sample number on the performance of the classifier. The larger k was, the more samples there were in the training set, and then the fewer samples there were in the test set, and vice versa. The range of k ranged from 5 to 30 with step size = 1, and **Figure 3** shows the accuracy of CNA_origin, LSTM, RF, XGboost, and CNA_origin with the different fold k values. With increasing k value, the performance of CNA_origin was gradually improved at first, which could be due to a bigger k including more training samples. But, as k became larger, the number of samples in the test set became smaller, and the performance of the classifiers was weakened. The results indicated that the performance of CNA_origin would be further improved if the training samples were expanded and that sufficient test samples were also very important for model evaluation.

3.5. Performance Comparison of Independent Datasets

In order to compare generalization performance on the independent data, experiments were performed with CNA datasets released by TCGA in 2016 downloaded from http://gdac.broadinstitute.org/. The TCGA datasets had 1080 BRCA samples, 611 COADRAD samples, 577 GBM samples, 528 KIRC samples, 552 OV samples, and 533 UCEC samples, respectively. The preprocessing analysis of 24776 gene CNA values was performed with GISTIC2 (Mermel et al., 2011). The TCGA datasets were reasonably independent of the training data because of preprocessing analyses such as quality control, alignment, and variation detection, which had a different systematic bias. The genes involved in both MSKCC datasets and TCGA datasets were selected, and the TCGA samples existing in MSKCC datasets were removed. There were 19895 common genes present in the MSKCC and TCGA datasets, and the independent datasets contained 234 BRCA samples, 50 COADRAD samples, 25 GBM samples, 41 KIRC samples, 21 OV samples, and 99 UCEC samples (see **Supplementary Material** for details). The independent datasets were used to evaluate the performance of CNA_origin. As shown in **Figure 4**, the overall performance of CNA_origin in terms of precision, recall, accuracy, and F1-score was the highest among the tools, at 0.74, 0.85, 0.79, and 0.77, respectively (see **Supplementary Material** for details). According to the results shown in **Figure 4**, it was concluded that CNA_origin performed successfully in the independent datasets.

4. CONCLUSIONS

Patients with CUPS often have a low median survival time of 9–12 months. Precise diagnosis of the tissue origin for metastatic CUPS is essential for determining the treatment scheme to improve patient prognosis. A lot of studies have tried to use cancer biomarkers to predict the primary tumor site for CUPS so as to provide important guidelines for timely patient care and cancer therapy. CNA provides a new way to identify and classify tumor types. In this study, a computational method, CNA_origin, was proposed to predict the tissue of origin from information on gene CNA levels. CNA_origin set up an intellectual deep-learning network mainly composed of an autoencoder and a CNN. This predictor successfully learned the inherent information of gene copy number and exhibited superior performance to the classical algorithms on k-fold cross-validations and independent datasets.

At present, the accuracy of using only CNA as the biomarker for tumor traceability is not very high. Integrating multiple biomarkers, such as CNA and DNA methylation or gene expression data, to trace tumor is our future goal.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YL conceived of the algorithm, develop the program, and wrote the manuscript. JY, BW, and GT helped with manuscript editing, designed, and performed experiments. PS and YW prepared the datasets. XL and CD carried out analyses and helped with the program design. HW designed of the work and participated in revising articles. All authors read and approved the final manuscript.

FUNDING

This study was supported by Jiangxi Provincial Department of Education Science and Technology plan Project (Grant No. GJJ180197) and Jiangxi Provincial Natural Science Fund (No. 20192ACB21004).

ACKNOWLEDGMENTS

The authors thank the reviewers for their useful feedback and constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00701/full#supplementary-material

REFERENCES

Bae, J. M., Kim, K., Chae, H. J., Wen, X., Kim, K. Y., Gwon, H. K., et al. (2018). Abstract 3312: Identification of tissue-of-origin in cancer of unknown primary site (cups) using methylation-specific targeted resequencing: a pilot study. *Cancer Res.* 78(13 Suppl.), 3312–3312. doi: 10.1158/1538-7445.AM2018-3312

Baek, J., Lee, B., Kwon, S., and Yoon, S. (2018). LncRNAnet: long non-coding RNA identification using deep learning . *Bioinformatics* 34, 3889–3897. doi: 10.1093/bioinformatics/bty418

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20007–20012. doi: 10.1073/pnas.0710052104

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., et al. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am. J. Pathol.* 164, 9–16. doi: 10.1016/S0002-9440(10)63090-8

Chen, F., Zhang, Y., Bossé, D., Lalani, A.-K. A., Hakimi, A. A., Hsieh, J. J., et al. (2017). Pan-urologic cancer genomic subtypes that transcend tissue of origin. *Nat. Commun.* 8, 1–15. doi: 10.1038/s41467-017-00289-x

Chen, T., and Guestrin, C. (2016). "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA).

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258.

Chollet, F. (2015). *keras*. GitHub repository. Available online at: https://github.com/keras-team/keras

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567. doi: 10.1038/s41591-018-0177-5

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., et al. (2020). Developing computational model to predict protein-protein interaction sites based on the xgboost algorithm. *Int. J. Mol. Sci.* 21:2274. doi: 10.3390/ijms21072274

Ferracin, M., Pedriali, M., Veronese, A., Zagatti, B., Gafà, R., Magri, E., et al. (2011). Microrna profiling for the identification of cancers with unknown primary tissue-of-origin. *J. Pathol.* 225, 43–53. doi: 10.1002/path.2915

Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., et al. (2019). Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* 2:e192597. doi: 10.1001/jamanetworkopen.2019.2597

Handorf, C. R., Kulkarni, A., Grenert, J. P., Weiss, L. M., Rogers, W. M., Kim, O. S., et al. (2013). A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. *Am. J. Surg. Pathol.* 37:1067. doi: 10.1097/PAS.0b013e31828309c4

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for qsar prediction. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977009. [Epub ahead of print].

Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2018). Classifying the molecular functions of rab gtpases in membrane trafficking using deep convolutional neural networks. *Anal. Biochem.* 555, 33–41. doi: 10.1016/j.ab.2018.06.011

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., Chua, M. C. H., and Yeh, H.-Y. (2019a). Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture. *Comput. Struct. Biotechnol. J.* 17, 1245–1254. doi: 10.1016/j.csbj.2019.09.005

Le, N. Q. K., Yapp, E. K. Y., and Yeh, H.-Y. (2019b). Et-gru: using multi-layer gated recurrent units to identify electron transport proteins. *BMC Bioinform.* 20:377. doi: 10.1186/s12859-019-2972-5

Li, M., Li, H., Hong, G., Tang, Z., and Guo, S. (2019). Identifying primary site of lung-limited cancer of unknown primary based on relative gene expression orderings. *BMC Cancer* 19:67. doi: 10.1186/s12885-019-5274-4

Liang, Y., Qiu, K., Liao, B., Zhu, W., Huang, X., Li, L., et al. (2016). Seeksv: an accurate tool for somatic structural variation and virus integration detection. *Bioinformatics* 33, 184–191. doi: 10.1093/bioinformatics/btw591

Liu, X., Liu, X., Lai, Y., Yang, F., and Zeng, Y. (2019). "Random decision dag: An entropy based compression approach for random forest," in *International Conference on Database Systems for Advanced Applications* (Springer), 319–323.

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2015). Tumortracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med. Genomics* 8:58. doi: 10.1186/s12920-015-0130-0

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Mohsen, H., El-Dahshan, E.-S. A., El-Horbaty, E.-S. M., and Salem, A.-B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Comput. Inform. J.* 3, 68–71. doi: 10.1016/j.fcij.2017.12.001

Mueller, W. C., Spector, Y., Edmonston, T. B., Cyr, B. S., Jaeger, D., Lass, U., et al. (2011). Accurate classification of metastatic brain tumors using a novel microrna-based test. *Oncologist* 16, 165–174. doi: 10.1634/theoncologist.2010-0305

Poduri, A., Evrony, G. D., Cai, X., and Walsh, C. A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science* 341:1237758. doi: 10.1126/science.1237758

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

Rodriguez, A. C., Blanchard, Z., Maurer, K. A., and Gertz, J. (2019). Estrogen signaling in endometrial cancer: a key oncogenic pathway with several open questions. *Horm. Cancer* 10, 51–63. doi: 10.1007/s12672-019-0358-9

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.* 26, 462–469. doi: 10.1038/nbt1392

Rosenwald, S., Gilad, S., Benjamin, S., Lebanony, D., Dromi, N., Faerman, A., et al. (2010). Validation of a microrna-based qrt-pcr test for accurate identification of tumor tissue origin. *Mod. Pathol.* 23, 814–823. doi: 10.1038/modpathol.2010.57

Scherbakov, A., Shestakova, E., Galeeva, K., and Bogush, T. (2019). Brca1 and estrogen receptor α expression regulation in breast cancer cells. *Mol. Biol.* 53, 442–451. doi: 10.1134/S0026893319030166

Sehouli, J., Braicu, E. I., Richter, R., Denkert, C., Jank, P., Jurmeister, P. S., et al. (2019). Prognostic significance of ki-67 levels and hormone receptor expression in low-grade serous ovarian carcinoma: an investigation of the tumor bank ovarian cancer network. *Hum. Pathol.* 85, 299–308. doi: 10.1016/j.humpath.2018.10.020

Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., and Wang, T. (2016). Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing* 194, 87–94. doi: 10.1016/j.neucom.2016.01.074

Søkilde, R., Vincent, M., Møller, A. K., Hansen, A., Høiby, P. E., Blondal, T., et al. (2014). Efficient identification of mirnas for classification of tumor origin. *J. Mol. Diagn.* 16, 106–115. doi: 10.1016/j.jmoldx.2013.10.001

Staub, E., Buhr, H., and Gröne, J. (2010). Predicting the site of origin of tumors by a gene expression signature derived from normal tissues. *Oncogene* 29, 4485–4492. doi: 10.1038/onc.2010.196

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2017). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., Van Laar, R. K., et al. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res.* 65, 4031–4040. doi: 10.1158/0008-5472.CAN-04-3617

Wang, A., An, N., Chen, G., Liu, L., and Alterovitz, G. (2018). Subtype dependent biomarker identification and tumor classification from gene expression profiles. *Knowl.Based Syst.* 146, 104–117. doi: 10.1016/j.knosys.2018.01.025

Wang, Q., Xu, M., Sun, Y., Chen, J., Chen, C., Qian, C., et al. (2019). Gene expression profiling for diagnosis of triple-negative breast cancer: a multicenter, retrospective cohort study. *Front. Oncol.* 9:354. doi: 10.3389/fonc.2019.00354

Wu, A. H., Drees, J. C., Wang, H., VandenBerg, S. R., Lal, A., Henner, W. D., et al. (2010). Gene expression profiles help identify the tissue of origin for metastatic brain cancers. *Diagn. Pathol.* 5:26. doi: 10.1186/1746-1596-5-26

Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., et al. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Mod. Pathol.* 29, 546–556. doi: 10.1038/modpathol.2016.60

Yang, X., Wu, W., Chen, Y., Li, X., Zhang, J., Long, D., et al. (2019). An integrated inverse space sparse representation framework for tumor classification. *Pattern Recogn.* 93, 293–311. doi: 10.1016/j.patcog.2019.04.013

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860(11 Pt B), 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zheng, Y., Ding, Y., Wang, Q., Sun, Y., Teng, X., Gao, Q., et al. (2019). 90-gene signature assay for tissue origin diagnosis of brain metastases. *J. Transl. Med.* 17:331. doi: 10.1186/s12967-019-2082-1

Conflict of Interest: JY, CD, GT, and BW were employed by the company Geneis Beijing Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 August 2020
doi: 10.3389/fgene.2020.00960

Prediction of Radiosensitivity in Head and Neck Squamous Cell Carcinoma Based on Multiple Omics Data

Jie Liu[1†], *Mengmeng Han*[1†], *Zhenyu Yue*[2,3], *Chao Dong*[1], *Pengbo Wen*[4,5], *Guoping Zhao*[4], *Lijun Wu*[1,4], *Junfeng Xia*[1] *and Yannan Bin*[1*]

[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institutes of Physical Science and Information Technology, Anhui University, Hefei, China, [2] Anhui Provincial Engineering Laboratory of Beidou Precision Agricultural Information, Anhui Agricultural University, Hefei, China, [3] School of Information and Computer, Anhui Agricultural University, Hefei, China, [4] Key Laboratory of High Magnetic Field and Ion Beam Physical Biology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, [5] Department of Bioinformatics, School of Medical Informatics, Xuzhou Medical University, Xuzhou, China

Head and neck squamous cell carcinoma (HNSCC) is a malignant tumor. Radiotherapy (RT) is an important treatment for HNSCC, but not all patients derive survival benefit from RT due to the individual differences on radiosensitivity. A prediction model of radiosensitivity based on multiple omics data might solve this problem. Compared with single omics data, multiple omics data can illuminate more systematical associations between complex molecular characteristics and cancer phenotypes. In this study, we obtained 122 differential expression genes by analyzing the gene expression data of HNSCC patients with RT (N = 287) and without RT (N = 189) downloaded from The Cancer Genome Atlas. Then, HNSCC patients with RT were randomly divided into a training set (N = 149) and a test set (N = 138). Finally, we combined multiple omics data of 122 differential genes with clinical outcomes on the training set to establish a 12-gene signature by two-stage regularization and multivariable Cox regression models. Using the median score of the 12-gene signature on the training set as the cutoff value, the patients were divided into the high- and low-score groups. The analysis revealed that patients in the low-score group had higher radiosensitivity and would benefit from RT. Furthermore, we developed a nomogram to predict the overall survival of HNSCC patients with RT. We compared the prognostic value of 12-gene signature with those of the gene signatures based on single omics data. It suggested that the 12-gene signature based on multiple omics data achieved the best ability for predicting radiosensitivity. In conclusion, the proposed 12-gene signature is a promising biomarker for estimating the RT options in HNSCC patients.

Keywords: head and neck squamous cell carcinoma, radiotherapy, multiple omics data, radiosensitivity, gene signature

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Xinguo Lu,
Hunan University, China
Yanwen Chong,
Wuhan University, China

***Correspondence:**
Yannan Bin
ynbin@ahu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 June 2020*
Accepted: *31 July 2020*
Published: *18 August 2020*

Citation:
Liu J, Han M, Yue Z, Dong C, Wen P, Zhao G, Wu L, Xia J and Bin Y (2020) Prediction of Radiosensitivity in Head and Neck Squamous Cell Carcinoma Based on Multiple Omics Data. Front. Genet. 11:960. doi: 10.3389/fgene.2020.00960

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common malignancy in the world, and nearly 60% of newly diagnosed HNSCC is locally advanced disease (Alsahafi et al., 2019; van der Heijden et al., 2019; Wang et al., 2019). Radiotherapy (RT) is a commonly used adjuvant therapy for HNSCC in addition to surgical treatment (Jemal et al., 2011; Suh et al., 2015). But each of HNSCC patients receiving the same dose of RT has different responses due to the complexity and heterogeneity of tumor, and some patients even have RT injury and secondary cancer (Scaife et al., 2015). Globally, the prognosis of HNSCC patients receiving RT remains a challenge. Therefore, prognostic biomarkers of radiosensitivity prediction for HNSCC are needed to improve RT options and predict treatment response.

In recent years, several studies have suggested that miRNAs, lncRNAs, and some of their target genes were correlated with the RT outcomes in HNSCC patients (Leucci et al., 2016; Weng et al., 2016; Chen et al., 2018; Han et al., 2018). For instance, the upregulation of miR-494-3p expression can enhance the radiosensitivity of HNSCC (Weng et al., 2016), and the upregulation of LINC00473 promotes the radioresistance of HNSCC cells (Han et al., 2018). Furthermore, some gene expression-based signatures have been constructed to predict the survival rate of HNSCC patients with RT. For example, Eschrich et al. (2009) developed the radiation sensitivity index, which was used to predict survival probability in HNSCC patients receiving concurrent chemoradiotherapy. Ma et al. (2019) identified a 4-gene methylation signature to predict the survival rate of HNSCC patients with RT. However, these studies only used single omics data which could not draw more comprehensive associations between complex molecular characteristics and cancer phenotypes. By contrast, multiple omics data involve multidimensional studies of cancer cells, potentially revealing the molecular mechanisms behind different phenotypes of cancer, such as metastasis and recurrence (Chakraborty et al., 2018; Xi et al., 2018; Wang et al., 2020). Therefore, a model based on multiple omics data could be an effective method for radiosensitivity prediction of HNSCC patients.

In this work, in order to construct reliable biomarkers for predicting RT response in HNSCC, we used the gene expression data and copy number variation (CNV)/single nucleotide variation (SNV) data from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) to develop a gene signature by the two-stage regularization (2SR) (Lin et al., 2015; Hu et al., 2019) and multivariable Cox regression (Gui and Li, 2005; Benner et al., 2010) models. Then, we evaluated the ability of the gene signature for predicting radiosensitivity by Kaplan–Meier survival analysis (Tripepi and Catalano, 2004). Furthermore, we constructed a nomogram based on the gene signature and clinical variables to facilitate a more intuitive prediction of 3-year and 5-year survival rates for HNSCC patients receiving RT.

MATERIALS AND METHODS

Figure 1 illustrated the workflow of the proposed signature for predicting RT response in HNSCC. Firstly, R package DESeq2 (Love et al., 2014) was used to identify differential expression genes (DEGs) between HNSCC patients receiving RT and without RT. Secondly, the 2SR and multivariable Cox regression models were used to construct a gene signature associated with the radiosensitivity prediction of HNSCC patients. Finally, Kaplan–Meier survival analysis and time-dependent receiver operating characteristic (ROC) curves (Heagerty et al., 2000; Kamarudin et al., 2017) were used to evaluate the performance of the gene signature. And a nomogram based on the gene signature and clinical variables was constructed to predict the 3-year and 5-year survival rates. The major procedures were described in the following sections.

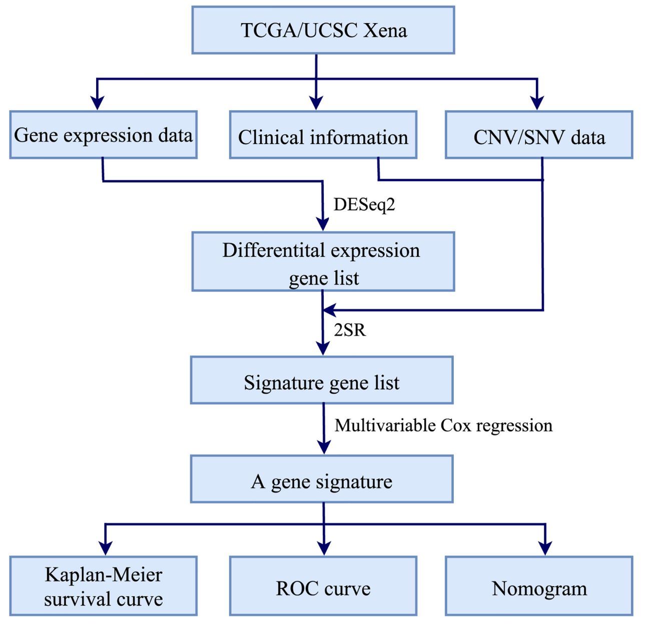


FIGURE 1 | Workflow of constructing a gene signature for predicting RT response in HNSCC.

Data Processing

We downloaded the transcriptomic gene expression data and the clinical follow-up data of HNSCC patients from TCGA (Tomczak et al., 2015). Meanwhile, we collected the genomic CNV/SNV data from UCSC Xena platform (Goldman et al., 2019). Firstly, we abandoned samples with overall survival (OS) less than 30 days to avoid the impact of deaths with unrelated causes (Chen et al., 2018), and a total of 476 HNSCC patients were analyzed. Furthermore, the available clinical variables included gender, age, OS, vital status, T stage, N stage, clinical stage, tumor grade, and RT options. Secondly, we obtained 20,557 common genes from the gene expression data and the CNV/SNV data, which were used to screen the gene signature associated with the radiosensitivity prediction of HNSCC in subsequent analysis. Thirdly, DESeq2 was used in the normalization of gene expression data and the detection of DEGs between 287 HNSCC patients with RT and 189 patients without RT. As a consequence, genes with $|\log_2 \text{ fold change}| \geq 1$ and false discovery rate (FDR) < 0.01 were defined as DEGs for the further analysis. Finally, the CNV/SNV data of DEGs were converted into a sample-by-gene matrix. If there were one or more mutations within the gene for each sample, the gene-level mutation status value in sample-by-gene matrix was defined as 1; otherwise, it was defined as 0.

Establishment of the Gene Signature

Based on DEGs, we used the 2SR and multivariable Cox regression models to construct a gene signature that can predict the radiosensitivity of HNSCC patients. The 2SR model could integrate multiple layer omics data to identify signature genes. Specifically, on the first layer, we predicted gene expression values using the CNV/SNV data of DEGs. On the second layer, we used a regularization methodology to regress the predicted gene expression on the first layer and performed signature genes selection and estimation. Multivariable Cox regression model was used to estimate regression coefficients for the identified signature genes. From this model, gene score of HNSCC patients was described as the sum of the products of individual gene expression levels and the estimated regression coefficients. The detailed processes were elaborated in the following.

Firstly, the 287 patients with RT were randomly divided into a training set ($N = 149$) and a test set ($N = 138$) (**Table 1**). Secondly, on the training set, we input gene expression data, clinical data, and the sample-by-gene matrix of DEGs into the 2SR model to select the OS-related signature genes. The 2SR model was used with default parameters, and the output of this model was a list of genes and their correlated coefficients with OS in patients with RT. When the correlation between genes and OS reached 80% and above, these genes were identified as signature genes. Finally, we calculated the coefficients of these signature genes using multivariable Cox regression model and constructed a gene signature according to the expression levels of these genes, which can stratify HNSCC patients into the high- and low-score groups with the median score of the gene signature on the training set as cutoff value.

Statistical Analysis

In the work, the ROC curve was performed via the R package survivalROC (Heagerty et al., 2013), and the area under the ROC curve (AUC) was used to assess the overall performance of radiosensitivity prediction. Kaplan–Meier survival curves were used to further evaluate the significance difference of OS between different groups. A two-tailed P-value (P) < 0.05 was considered statistically significant in all analyses. The nomogram and the calibration plot were established using R package rms (Harrell et al., 2019) and were used to predict OS of HNSCC patients with RT.

RESULTS AND DISCUSSION

Identification of a 12-Gene Signature

Based on the gene expression data from 287 patients with RT and 189 patients without RT, 122 DEGs with $|\log_2$ fold change$| \geq 1$ and FDR < 0.01 were obtained. The gene expression, clinical, and CNV/SNV data of these DEGs on the training set were imported into 2SR model. With the probability related to OS should be >80% of HNSCC patients receiving RT, 12 genes were picked out as the signature genes. Next, we calculated the coefficients of these signature genes using multivariable Cox regression model. Finally, in order to predict the radiosensitivity of HNSCC patients, we constructed a gene signature on the training set according to the expression levels of these 12 genes as follows: gene score = *TDRD9* × 6.950E-6 + *CELF3* × 1.106E-2 + *FGF19* × 1.937E-5 + *KCNB2* × 5.388E-3 + *CLDN6* × 1.334E-4 − *BEST2* × 6.053E-4 − *DDX25* × 1.802E-3 − *TMPRSS15* × 4.378E-4 − *ALPI* × 2.134E-3 − *FABP7* × 1.754E-3 − *IL17REL* × 2.132E-3 − *RORB* × 1.182E-3. The details of these 12 signature genes were shown in **Supplementary Table S1**. Then, based on the gene scores, the patients were divided into the high- and low-score groups, where the cutoff value of -0.06338 was derived from the median score of gene scores on the training set samples. Specifically, on the training (149 HNSCC patients receiving RT) and test (139 patients receiving RT) sets, HNSCC patients with a gene score ≥-0.06338 were divided into the high-score group, while those with a gene score <-0.06338 were divided into the low-score group.

Radiosensitivity Prediction by the 12-Gene Signature

To assess the radiosensitivity prediction ability of the 12-gene signature on the HNSCC patients, Kaplan–Meier survival and ROC curves were performed on the training and test sets,

TABLE 1 | Clinical variables of the 287 HNSCC patients with RT.

Variable	Subgroup	Total		Training set		Test set		P*
		N	%	*N*	%	*N*	%	
Age (year)	≤60	156	54.36	87	58.39	69	50.0	0.22
	>60	131	45.64	62	41.61	69	50.0	
Gender	Male	224	78.05	123	82.55	101	73.19	0.15
	Female	63	21.95	26	17.45	37	26.81	
T stage	1	14	4.88	9	6.04	5	3.62	0.34
	2	66	22.30	27	18.12	39	28.26	
	3	79	27.53	45	30.20	34	24.64	
	4	122	42.51	63	42.28	59	42.75	
N stage	0	113	39.37	57	38.26	56	40.58	0.35
	1	56	19.51	26	17.45	30	21.74	
	2	105	36.59	59	39.60	46	33.33	
	3	3	1.05	0	0	3	2.26	
Clinical stage	1	9	3.14	2	1.34	7	5.07	0.30
	2	23	8.01	12	8.05	11	7.97	
	3	48	16.72	21	14.09	27	19.57	
	4	207	72.13	114	76.51	93	67.39	
Tumor grade	1	25	8.71	11	7.38	14	10.14	0.20
	2	168	58.54	82	55.03	86	62.32	
	3	75	26.13	47	31.54	28	20.29	
	4	2	0.70	2	1.34	0	0	
Survival time (month)		21.77		20.73		23.62		0.32
Vital status	Death	110	38.33	57	38.26	53	38.41	0.12
	Alive	177	61.67	92	61.74	85	61.59	

**The difference between the test set and training set was calculated in terms of clinical pathologic factors. Specifically, age was compared with Wilcoxon rank-sum test; gender, clinical T, N, stage, and grade were compared with the chi-squared test; survival time and the status difference were assessed with the log-rank test.*

respectively. On the training set, there was a significant difference on radiosensitivity between the high- and low-score groups (P = 0.0011, **Figure 2A**). As we can see, patients in the high-score group were associated with poor radiosensitivity, while patients in the low-score group showed good radiosensitivity. In the light of the time-dependent ROC curves of 3-year and 5-year survival (**Figure 2B**), the survival time prediction accuracy of the 12-gene signature for HNSCC patients had AUC of 0.705 at 3 years and 0.697 at 5 years. Furthermore, the prediction performance of the 12-gene signature was evaluated on the test set. As seen in **Figure 2C**, the survival rate was significantly higher in the low-score group than that in the high-score group (P = 0.00031), which was similar with the result of the training set. And the 3-year and 5-year prediction accuracy achieved 0.661 and 0.584, respectively (**Figure 2D**). The performance on the test set was similar with that on the training set. It indicated the generalization ability of the 12-gene signature was good, and this gene signature could provide a method to predict the radiosensitivity for HNSCC patients. However, the radiosensitivity prediction accuracy of 5-year survival on the test set was lower than that on the training set (**Figures 2B,D**). There were two possible reasons to explain the causes of the difference. First, the follow-up times were relatively short for HNSCC cohort in TCGA, and the OS of most patients was also less than 5 years. Second, the intrinsic genetic heterogeneity of the tumor could lead to different OS in HNSCC patients with same therapeutic method. The longer the OS, the larger the effect of the intrinsic genetic heterogeneity, and the more difficultly we evaluated this effect with RT.

Assessment of the 12-Gene Signature in All HNSCC Patients

Because of the complexity and heterogeneity of tumors, some HNSCC patients are treated by RT with good outcomes, and some patients may have a resistance to RT, even get worse. Therefore, the 12-gene signature is important for the radiosensitivity prediction of HNSCC patients to improve RT options. In addition, we also assessed the prognostic value of the 12-gene signature in a total of 476 HNSCC patients. As seen in **Figure 3A**, it indicated that patients in the low-score group generally had a higher 5-year survival rate than that in the high-score group ($P < 0.0001$). And there was a significant difference between patients with RT and without RT in the low-score groups (P = 0.0033, **Figure 3B**); however, in the high-score group, the difference was insignificant (P = 0.95, **Figure 3C**). It suggested that patients in the high-score group might have the radioresistance and did not benefit from RT.

In addition, based on different clinical variables such as age, gender, T stage, N stage, clinical stage and tumor grade, HNSCC patients were further stratified into different subgroups (**Table 1**). Then we evaluated the prognostic value of the 12-gene signature on these different subgroups between HNSCC patients in the high- and low-score groups. On the subgroups of clinic T (**Figures 4A,B**), clinic N (**Figures 4C,D**), stage 3-4 (**Figure 4F**), and grade (**Figures 4G,H**), the survival rates of patients in the low-score group were significantly higher than those in the high-score group. While there was no statistically significant difference of survival rate between patients in the high- and low-score groups in the subgroup of stage 1-2 (**Figure 4E**). It was mainly due to the small number of patients in this clinical phase, which accounted for only 11% of all HNSCC patients. Furthermore, on the subgroups of age (**Supplementary Figures S1A,B**) and gender (**Supplementary Figure S1C**), the survival rates of patients in the low-score group also were significantly longer than those patients in the high-score group. However, the survival rate between female patients (**Supplementary Figure S1D**) in the high- and low-score groups did not have statistically significant difference, which was similar with that on the stage 1-2 subgroup. Taken together, these results suggested that this 12-gene signature could serve as a novel and reliable biomarker for the radiosensitivity prediction of HNSCC patients.

Prediction of OS in HNSCC With RT by Nomogram

As a visual tool, nomogram has been widely used in predicting the prognosis of cancers (Lubsen et al., 1978; Gorlia et al., 2008). In this work, the nomogram was used to construct the OS prediction model for HNSCC patients with RT. As shown in **Figure 5A**, age, T stage, N stage, clinical stage, tumor grade, and gene score were considered as relevant variables for the nomogram construction. Since the points of male and female in the nomogram were similar and closed to zero, gender was not shown here. In addition, the contribution of gene score was very important, and it played a crucial role in survival estimation of HNSCC patients with RT. In order to evaluate the predicted outcomes of nomogram, the calibration plots on the training and test sets were exhibited in **Figure 5B**. It was found that the predicted results of the nomogram showed good agreements with the actual situations, especially on the test set. Furthermore, according to the time-dependent ROC curves (**Figure 5C**), the nomogram achieved 0.701 and 0.641 of AUC for 3-year OS on the training and test sets, respectively. It revealed that the nomogram could be used as a promising tool to predict the OS of HNSCC patients with RT.

Comparison With the Gene Signatures Based on Single Omics Data

Besides the 12-gene signature established using multiple omics data, we also assessed the radiosensitivity prediction ability of the gene signatures based on single omics data, such as gene expression data or CNV/SNV data. Since the 2SR model requires multiple omics data as the input files, we used differential expression analysis, univariable Cox proportional hazards regression analysis (David, 1972), classical LASSO regression model (Tibshirani, 1996, 1997; Segal, 2006), and multivariate Cox regression model to select the most important biomarkers and take their linear combination as a predictor of radiosensitivity based on single omics data. Firstly, differential expression analysis was used to identify DEGs by analyzing gene expression profiles of HNSCC patients with RT and without RT. Of note, the construction of the gene

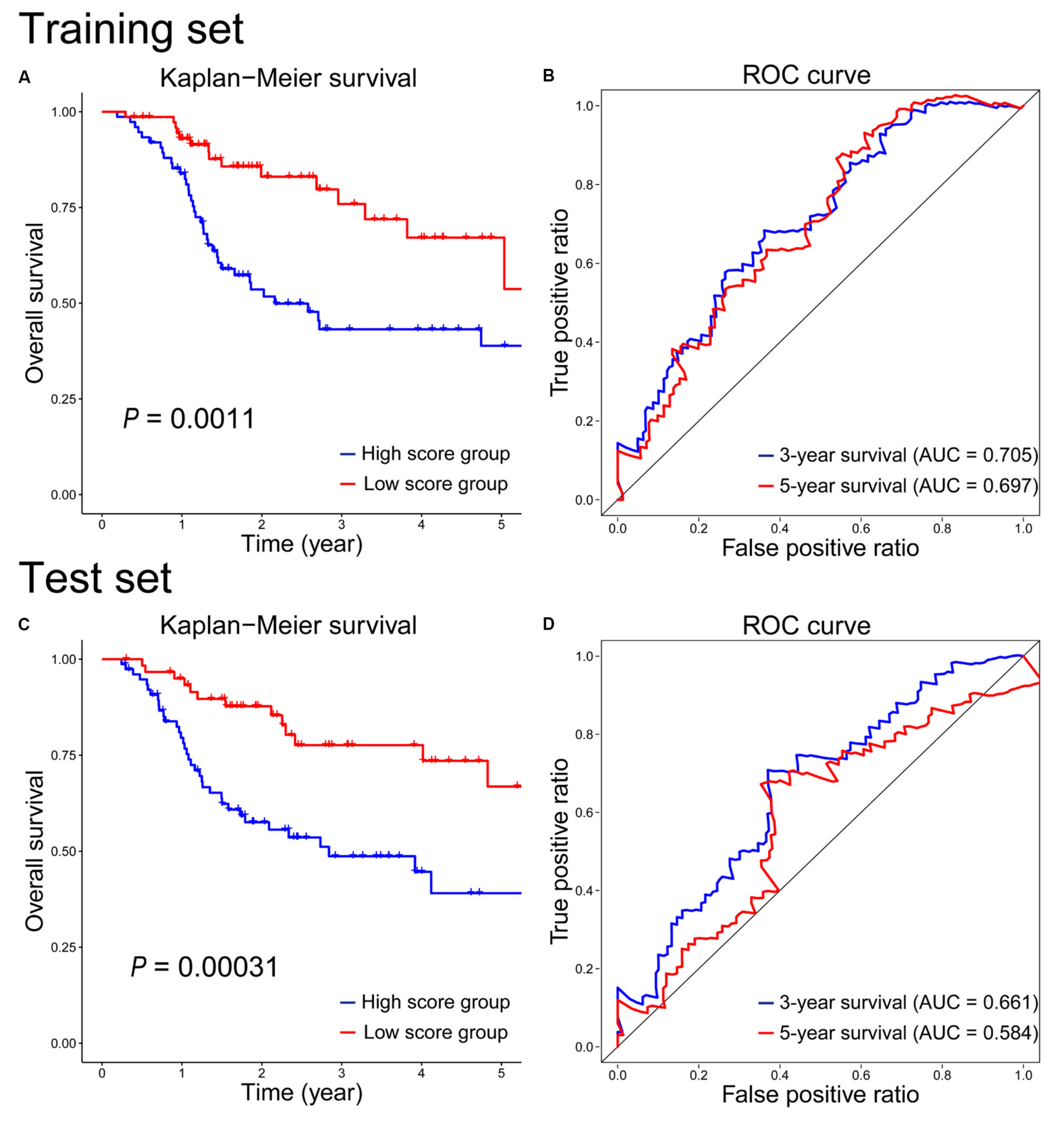


FIGURE 2 | Kaplan–Meier survival and time-dependent ROC curves on the training **(A,B)** and test **(C,D)** sets according to the 12-gene signature.

signature based on the CNV/SNV data did not use differential expression analysis. Secondly, univariate Cox proportional hazards regression analysis and LASSO logistic regression model were used to screen out the characteristic genes associated with survival. Thirdly, multivariate Cox regression model was used to establish a gene signature for radiosensitivity prediction. Then, HNSCC patients receiving RT were divided into the high- and low-score groups according to the median score on the training set patients. Finally, Kaplan–Meier survival analysis and ROC curves were conducted to observe the difference of survival rate between the patients in the high- and low-score groups.

Based on gene expression data, a 7-gene signature was constructed for radiosensitivity prediction, and the gene

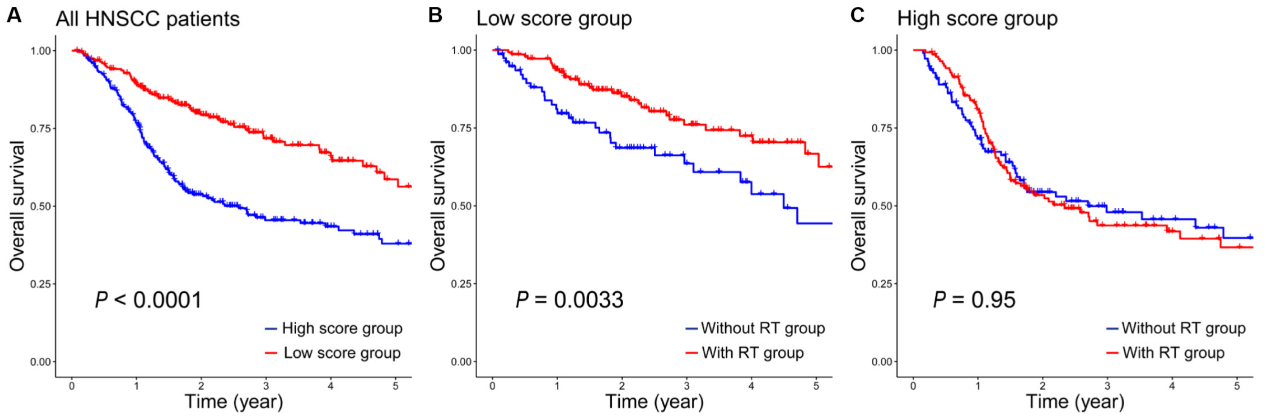


FIGURE 3 | The prognostic values of the 12-gene signature in all HNSCC patients. **(A)** Kaplan–Meier analysis of overall survival in 476 patients according to the 12-gene signature. **(B)** Kaplan–Meier survival curves of patients with/without RT in the low-score group. **(C)** Kaplan–Meier survival curves of patients with/without RT in the high-score group.

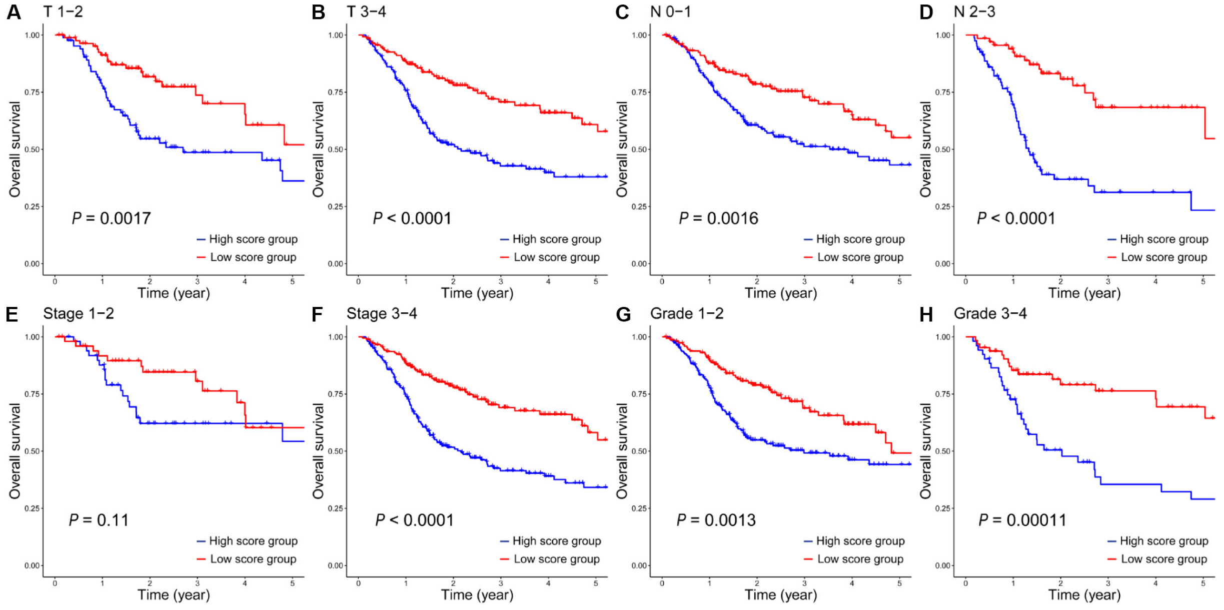


FIGURE 4 | The Kaplan–Meier survival analysis of the 12-gene signature in all HNSCC patients in the high- and low-score groups on clinical subgroups of T 1-2 **(A)**, T 3-4 **(B)**, N 0-1 **(C)**, N 2-3 **(D)**, Stage 1-2 **(E)**, Stage 3-4 **(F)**, Grade 1-2 **(G)** and Grade 3-4 **(H)**.

score = *CHGB* × 1.121E-4 + *ODAM* × 3.603E-5 + *RP11-169K17.3* × 8.518E-2 − *ZNF541* × 8.229E-5 − *CLGN* × 2.952E-3 − *AC011747.3* × 1.407E-1 − *RP11-203B7.2* × 1.357E-1. The details of these 7 signature genes were shown in **Supplementary Table S2**. The survival analysis results on the test set were shown in **Figure 6A**. Obviously, the difference between the high- and low-score groups was significant (P = 0.029), which was similar with that on the training set (P = 0.011, **Supplementary Figure S2A**). The prognostic accuracy of the 7-gene signature for HNSCC patients receiving RT was 0.62 at 3 years and 0.575 at 5 years (**Figure 6B**), which were both lower than the performances of the 12-gene signature based on multiple omics data.

Based on CNV/SNV data, a 3-gene signature was developed for radiosensitivity prediction, and the gene score = − *BCLAF1* × 8.084E-1 − *ABCB9* × 3.330E-1 − *MIS18BP1* × 3.697E-1. The details of the 3-gene signature were shown in **Supplementary Table S3**. The survival analysis on the test set showed that there was a significant difference between the high- and low-score groups (P = 0.018, **Figure 6C**), but it was inconsistent with the result on the training set (P = 0.0068, **Supplementary Figure S2C**). The 3-year and 5-year survival prognostic accuracy of the 3-gene signature were 0.387 and 0.345 on the test set, respectively (**Figure 6D**), which were far less than the performances of the 12-gene signature using multiple omics data. As single omics data, the sample-by-gene

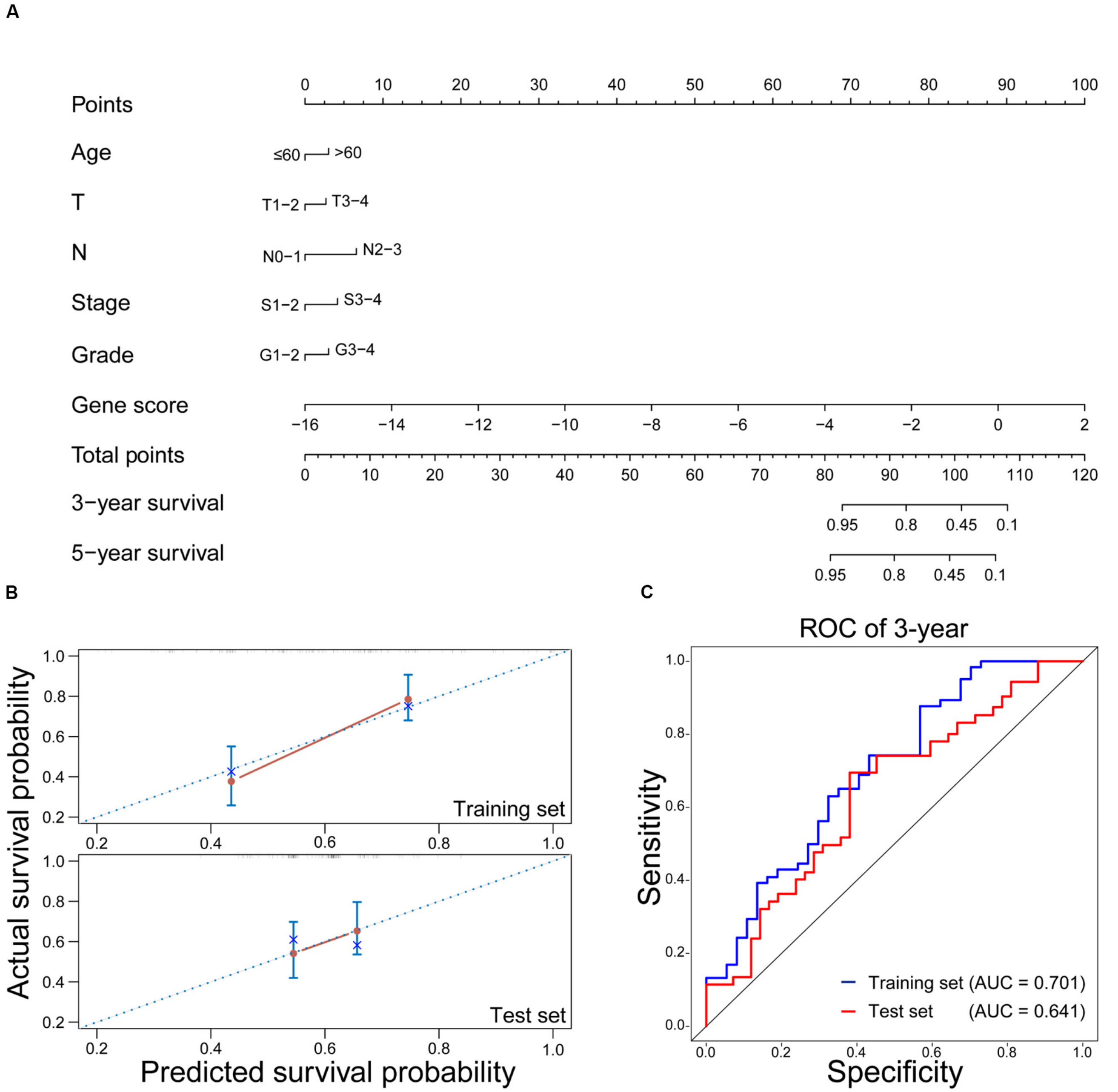


FIGURE 5 | Evaluation of the nomogram on predicting the OS of HNSCC patients with RT. **(A)** Nomogram for predicting the 3-year and 5-year OS in HNSCC with RT. **(B)** Calibration plots of the nomogram on the training and test sets. The 45-degree line represents the real outcomes. **(C)** Time-dependent ROC curves of 3-year OS prediction using the nomogram on the training and test sets.

matrix based on CNV/SNV data was sparse, and the genetic information extracted from the sparse matrix was very limited. So the radiosensitivity prediction accuracy of the gene signature based on CNV/SNV data was not good.

At first, we compared the 12-gene signature with those gene signatures (7-gene and 3-gene signatures) based on single omics data. The results showed that the 12-gene signature achieved the highest separation ability, and it significantly stratified patients into the low- and high-score groups on the test set (P = 0.00031 vs. P = 0.029 vs. P = 0.018). It also had the highest accuracy of survival estimation among these gene signatures (3-year survival: 0.661 vs. 0.620 vs. 0.387; 5-year survival: 0.584 vs. 0.575 vs. 0.345) on the test set. In addition, we performed GO and KEGG enrichment analyses, and the result showed there were no significant enrichments. Moreover, on the GO and KEGG terms (Jiao et al., 2012; Xi et al., 2020), there are also no correlations between these signature genes and radiation. Maybe these genes

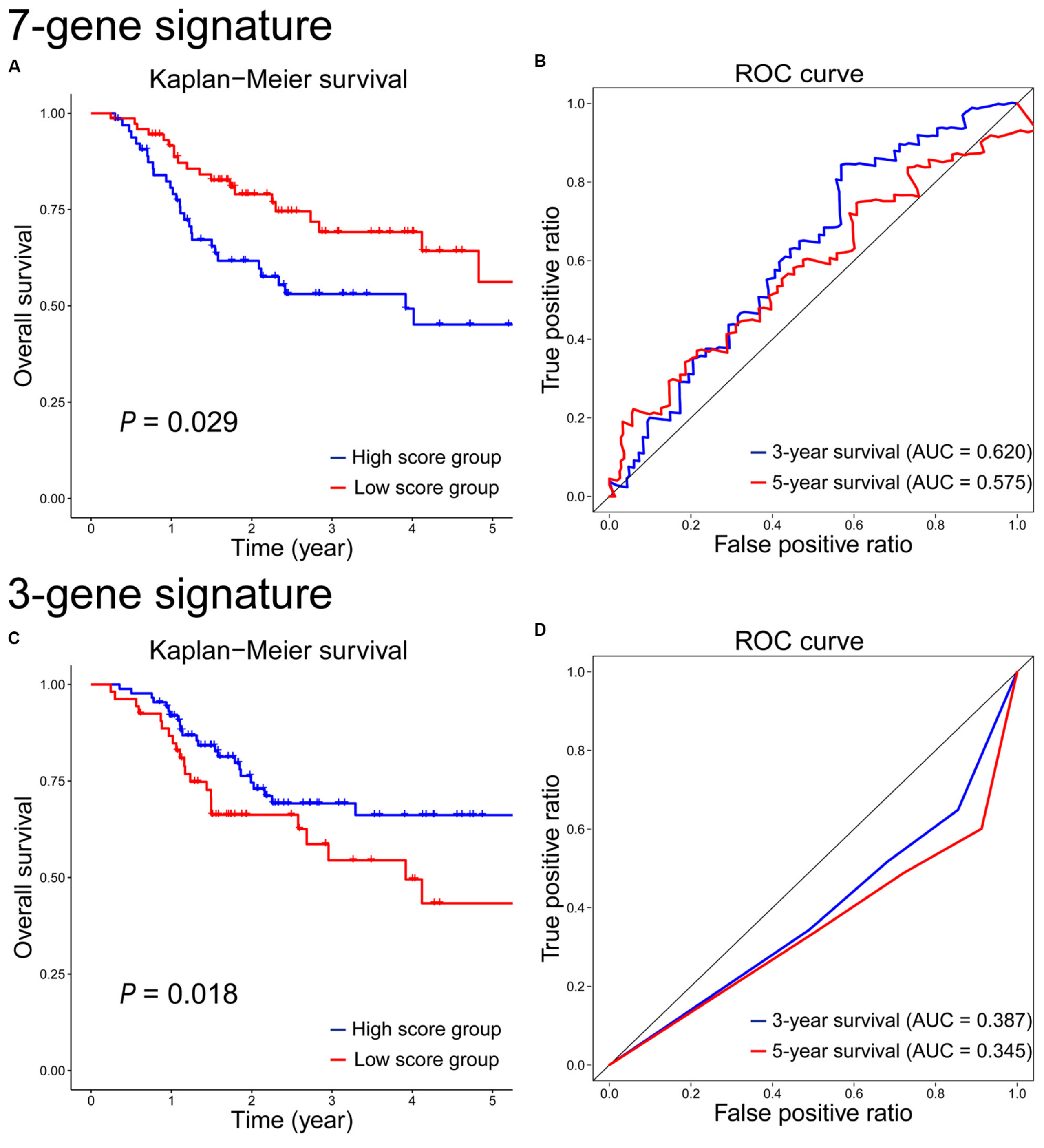


FIGURE 6 | Kaplan–Meier survival and time-dependent ROC curves on the test set according to the 7-gene signature **(A,B)** and the 3-gene signature **(C,D)**.

were novel candidate targets and biomarkers correlated with radiation. However, given the performance of 12-gene signature was better than those of 7-gene and 3-gene signatures, the genes in 12-gene signature were more important on radiation response and radiosensitivity prediction. Nevertheless, their roles need to be proved by biological and clinical experiments. Furthermore, we evaluated the performance of the 5-miRNA signature (Chen et al., 2018) on the test set. As shown in **Supplementary Figure S3**, there is no significant different between the high- and low-score groups based on 5-miRNA signature, and the items of AUC (0.493 and 0.450) on 3-year and 5-year survivals were lower than those (0.661 and 0.584) based on 12-gene signature. In a word, the 12-gene signature based on multiple omics data was a relatively reliable biomarker to predict whether the HNSCC patient benefit from the treatment of RT.

CONCLUSION

In this study, we used the gene expression, clinical, and CNV/SNV data to develop and validate the 12-gene signature, which may serve as a promising prognostic biomarker for the radiosensitivity prediction of HNSCC patients. Furthermore, we constructed a nomogram based on gene score and clinical variables, which might be a useful tool on the survival estimation of HNSCC patients receiving RT. Finally, we systemically compared the prognosis ability of the gene signatures based on multiple and single omics data, and the results showed that the 12-gene signature based on multiple omics data was more accurate in predicting radiotherapy response and survival rate of HNSCC patients. However, this study also has some limitations. First, these signature genes as biomarkers for radiosensitivity in HNSCC deserve further biological and clinical verification. Second, gene expression signatures are subject to sampling bias caused by the complexity and heterogeneity of tumors, so we will consider the subtypes of tumor in the future study.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JX, LW, GZ, MH, JL, and ZY conceived and designed the study. MH and JL performed the experiments. MH, JL, and CD wrote the manuscript. JX, YB, and PW reviewed the manuscript. All authors read and approved the final version of this manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (11835014, 61672037, and 21601001) and the Recruitment Program for Leading Talent Team of Anhui Province (2019-16).

ACKNOWLEDGMENTS

The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00960/full#supplementary-material

FIGURE S1 | The Kaplan-Meier survival analysis of the 12-gene signature in all HNSCC patients in the high and low score groups on age ≤ 60, age > 60, male, female subgroups.

FIGURE S2 | Kaplan-Meier survival and time-dependent ROC curves on the training set according to the 7-gene signature and the 3-gene signature.

TABLE S1 | The information of genes in the 12-gene signature.

TABLE S2 | The information of genes in the 7-gene signature.

TABLE S3 | The information of genes in the 3-gene signature.

REFERENCES

Alsahafi, E., Begg, K., Amelio, I., Raulf, N., Lucarelli, P., Sauter, T., et al. (2019). Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death Dis.* 10:540. doi: 10.1038/s41419-019-1769-1769

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., and Mansmann, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. *Biomed. J.* 52, 50–69. doi: 10.1002/bimj.200900064

Chakraborty, S., Hosen, M. I., Ahmed, M., and Shekhar, H. U. (2018). Onco-multi-OMICS approach: a new frontier in cancer research. *Biomed Res. Int.* 2018:9836256. doi: 10.1155/2018/9836256

Chen, L., Wen, Y., Zhang, J., Sun, W., Lui, V. W. Y., Wei, Y., et al. (2018). Prediction of radiotherapy response with a 5-microRNA signature-based nomogram in head and neck squamous cell carcinoma. *Cancer Med.* 7, 726–735. doi: 10.1002/cam4.1369

David, C. R. (1972). Regression models and life tables. *J. R. Stat. Soc. B.* 34, 187–220.

Eschrich, S., Zhang, H. L., Zhao, H. Y., Boulware, D., Lee, J. H., Bloom, G., et al. (2009). Systems biology modeling of the radiation sensitivity network: a biomarker discovery platform. *Int. J. Radiat. Oncol. Biol. Phys.* 75, 497–505. doi: 10.1016/j.ijrobp.2009.05.056

Goldman, M., Craft, B., Hastie, M., Repeèka, K., McDade, F., Kamath, A., et al. (2019). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv [Preprint]* Available online at: https://www.biorxiv.org/content/10.1101/326470v6 (accessed September 26, 2019).

Gorlia, T., van den Bent, M. J., Hegi, M. E., Mirimanoff, R. O., Weller, M., Cairncross, J. G., et al. (2008). Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *Lancet Oncol.* 9, 29–38. doi: 10.1016/S1470-2045(07)70384-70384

Gui, J., and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 21, 3001–3008. doi: 10.1093/bioinformatics/bti422

Han, P., Ji, X., Zhang, M., and Gao, L. J. E. R. M. P. S. (2018). Upregulation of lncRNA LINC00473 promotes radioresistance of HNSCC cells through activating Wnt/beta-catenin signaling pathway. *Eur. Rev. Med. Pharmacol.* 22, 7305–7313. doi: 10.26355/eurrev_201811_16267

Harrell, F. E. Jr., Harrell, M. F. E. Jr., and Hmisc, D. (2019). *Package 'rms'.* Nashville: Vanderbilt University, 229.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56, 337–344.

Heagerty, P. J., Saha-Chaudhuri, P., and Saha-Chaudhuri, M. P. (2013). *Package 'survivalROC'.*

Hu, X. H., Xie, W. B., Wu, C. C., and Xu, S. Z. (2019). A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnol. J.* 17, 2011–2020.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., et al. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 531, 518–522.

Lin, W., Feng, R., and Li, H. Z. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Am. Stat. Assoc.* 110, 270–288. doi: 10.1080/01621459.2014.908125

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-558

Lubsen, J., Pool, J., and Der Does, E. V. (1978). A practical device for the application of a diagnostic or prognostic function. *Methods Inf. Med.* 17, 127–129. doi: 10.1055/s-0038-1636613

Ma, J. B., Li, R., and Wang, J. (2019). Characterization of a prognostic four-gene methylation signature associated with radiotherapy for head and neck squamous cell carcinoma. *Mol. Med. Report.* 20, 622–632. doi: 10.3892/mmr.2019.10294

Scaife, J. E., Barnett, G. C., Noble, D. J., Jena, R., Thomas, S. J., West, C. M. L., et al. (2015). Exploiting biological and physical determinants of radiotherapy toxicity to individualize treatment. *Brit. J. Radiol.* 88:20150172. doi: 10.1259/bjr.20150172

Segal, M. R. (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics* 7, 268–285. doi: 10.1093/biostatistics/kxj006

Suh, Y. E., Raulf, N., Gaken, J., Lawler, K., Urbano, T. G., Bullenkamp, J., et al. (2015). MicroRNA-196a promotes an oncogenic effect in head and neck cancer cells by suppressing annexin A1 and enhancing radioresistance. *Int. J. Cancer* 137, 1021–1034. doi: 10.1002/ijc.29397

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380<3.0.co;2-3

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014

Tripepi, G., and Catalano, F. (2004). Kaplan-Meier analysis. *G. Ital. Nefrol.* 21, 540–546.

van der Heijden, M., Essers, P. B. M., de Jong, M. C., de Roest, R. H., Sanduleanu, S., Verhagen, C. V. M., et al. (2019). Biological determinants of chemo-radiotherapy response in HPV-negative head and neck cancer: a multicentric external validation. *Front. Oncol.* 9:1470. doi: 10.3389/fonc.2019.01470

Wang, H. C., Chan, L. P., and Cho, S. F. (2019). Targeting the immune microenvironment in the treatment of head and neck squamous cell carcinoma. *Front. Oncol.* 9:1084. doi: 10.3389/fonc.2019.01084

Wang, W., Zhou, Y., Cheng, M.-T., Wang, Y., Zheng, C.-H., Xiong, Y., et al. (2020). Potential pathogenic genes prioritization based on protein domain interaction network analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 99, 1–1.

Weng, J. H., Yu, C. C., Lee, Y. C., Lin, C. W., Chang, W. W., and Kuo, Y. L. (2016). miR-494-3p induces cellular senescence and enhances radiosensitivity in human oral squamous carcinoma cells. *Int. J. Mol. Sci.* 17:1092. doi: 10.3390/ijms17071092

Xi, J., Li, A., and Wang, M. (2018). HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 422–434.

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. J. B. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. 36, 1855–1863.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction

Kai Bian[1], Mengran Zhou[1,2], Feng Hu[1] and Wenhao Lai[1]*

[1] School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, China, [2] State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology, Huainan, China

Breast cancer is one of the most common cancer diseases in women. The rapid and accurate diagnosis of breast cancer is of great significance for the treatment of cancer. Artificial intelligence and machine learning algorithms are used to identify breast malignant tumors, which can effectively solve the problems of insufficient recognition accuracy and long time-consuming in traditional breast cancer diagnosis methods. To solve these problems, we proposed a method of attribute selection and feature extraction based on random forest (RF) combined with principal component analysis (PCA) for rapid and accurate diagnosis of breast cancer. Firstly, RF was used to reduce 30 attributes of breast cancer categorical data. According to the average importance of attributes and out of bag error, 21 relatively important attribute data were selected for feature extraction based on PCA. The seven features extracted from PCA were used to establish an extreme learning machine (ELM) classification model with different activation functions. By comparing the classification accuracy and training time of these different models, the activation function of the hidden layer was determined as the sigmoid function. When the number of neurons in the hidden layer was 27, the accuracy of the test set was 98.75%, the accuracy of the training set was 99.06%, and the training time was only 0.0022 s. Finally, in order to verify the superiority of this method in breast cancer diagnosis, we compared with the ELM model based on the original breast cancer data and other intelligent classification algorithm models. The algorithm used in this article has a faster recognition time and a higher recognition accuracy than other algorithms. We also used the breast cancer data of breast tissue reactance features to verify the reliability of this method, and ideal results were obtained. The experimental results show that RF-PCA combined with ELM can significantly reduce the time required for the diagnosis of breast cancer, which has the ability of rapid and accurate identification of breast cancer and provides a theoretical basis for the intelligent diagnosis of breast cancer.

Keywords: breast cancer, artificial intelligence, random forest, principal component analysis, extreme learning machine

OPEN ACCESS

Edited by:
Shu Tao,
UCLA Jonsson Comprehensive Cancer Center, United States

Reviewed by:
Kumardeep Chaudhary,
Icahn School of Medicine at Mount Sinai, United States
Wei Chen,
City of Hope National Medical Center, United States

****Correspondence:***
Mengran Zhou
mrzhou8521@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 May 2020*
Accepted: *19 August 2020*
Published: *09 September 2020*

Citation:
Bian K, Zhou M, Hu F and Lai W (2020) RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction. Front. Genet. 11:566057. doi: 10.3389/fgene.2020.566057

INTRODUCTION

Cancer is a disease that seriously threatens human health. The latest annual report on cancer incidence in the United States (Siegel et al., 2020) shows that it is estimated that in 2020, 1,806,590 new cancer cases will be found in the United States, which is equivalent to nearly 5,000 people suffering from cancer every day. There will be 606,520 cancer deaths, which is equivalent to more than 1,600 cancer deaths per day. Over the most recent 5—year period (2012–2016), the breast cancer incidence rate increased slightly by 0.3% per year (DeSantis et al., 2019). Cancer not only affects people's normal life but also brings a huge economic burden to people with high medical costs. Therefore, more and more researchers are committed to the research of cancer diagnosis and treatment methods (Gebauer et al., 2018). Among them, the incidence rate of breast cancer is only second after the lung cancer incidence rate in the world (Wang et al., 2018). Early detection and diagnosis of breast cancer are very helpful for treatment. If breast cancer is detected early, it can guide clinically targeted prevention and treatment measures, reduce the recurrence rate of breast cancer, improve the prognosis of patients, and prolong the life cycle of patients (Charaghvandi et al., 2017). How to quickly and accurately predict breast malignant tumors has become the key to the breast cancer diagnosis.

The traditional diagnosis method of breast cancer is mainly a fine-needle aspiration cell method (Dennison et al., 2015). The degree of canceration can be determined by observing the abnormal cell morphology of the collected tissue sections under the light microscope. This method needs the operation of experts with senior clinical experience, but it may cause the wrong diagnosis due to various uncertain subjective factors, which will also consume a lot of working time. In recent years, various prediction algorithms in machine learning can be well used in disease diagnosis, and more intelligent prediction results can be used to assist doctors, so as to speed up the time of diagnosis and improve the accuracy of diagnosis. For example, Cui et al. (2018) used neural network cascade (NNC) model identified numerous candidate miRNA biomarkers to detect breast cancer and obtained equivalent diagnostic performance. Wang et al. (2017) used a support vector machine (SVM)-based weighted AUC ensemble learning model to achieve a reliable and robust diagnosis of breast cancer. Noorul et al. (2019) proposed a transfer learning-based deep convolutional neural network (CNN) for segmentation to improve the detection rate of breast cancer for histopathological images. However, most of these machine learning algorithms analyze all the attributes of breast cancer data, which fails to take into account the influence of redundant information on the experimental results and the relationship between the attribute factors. Deep learning algorithm used to detect breast cancer needs to analyze the histopathological images of breast cancer, which not only requires a large number of samples, but also consumes a lot of time, and the prediction efficiency is low. Some artificial intelligence algorithms and classification models have been proposed to identify breast malignant tumor by using the Wisconsin Breast Cancer Database (WBCD). For example, Sewak et al. (2007) provide a resemble learning method based on SVMs to classify the breast malignant tumor and achieved with acceptable prediction accuracy. Nahato et al. (2015) combined rough set indiscernibility relation method with back propagation (BP) neural network for analysis of breast cancer dataset and the breast cancer dataset obtained its higher performance with a reduct of least number of attributes. Mert et al. (2014) used the independent component analysis and the discrete wavelet transform to reduce the dimension of data. A probabilistic neural network (PNN) classification model is established to increases the performance of breast cancer classification as benign and malignant and reduce the computational complexity. Jhajharia et al. (2016) used the principal component analysis (PCA) to preprocess the original breast cancer data, and then a decision tree (DT) prediction model was established to achieve the prognostic analysis of breast cancer data. Yang and Xu (2019) developed a feature extraction method by PCA and a differential evolution algorithm to optimize the parameter of SVM for the identification of breast tumors to present a superior classification performance.

Random forest (RF) is a supervised learning algorithm, which can select features according to the importance of attributes and reduce the complexity of the model (Odindi et al., 2014). Saraswat and Arya (2014) introduced a novel Gini importance-based binary random RF selection method to extract the relevant features of leukocytes and got a high classification accuracy. Zhou et al. (2017) proposed an iterative RF method to select candidate biomarkers and completed the classification of renal fibrosis. Wade et al. (2016) used the regularized RF to select the features of high dimensional shape data from subcortical brain surfaces. PCA is a kind of unsupervised learning feature extraction algorithm which maps high-dimensional data to low-dimensional space by linear projection and reduces the dimension of data sets (Simas Filho and Seixas, 2016). Skala et al. (2007) chose a method based on PCA to use the information inherent in the dose-volume histograms (DVH) to analyze after image-guided radiation therapy for prostate cancer. Fabris et al. (2014) employed sparse PCA to assess the glucose variability index of continuous glucose monitoring (CGM) time-series. Garbis et al. (2018) used PCA for proteomic quantitative analysis of primary cancer-associated fibroblasts in esophageal adenocarcinoma. Extreme learning machine (ELM) is an efficient and intelligent algorithm that can be used to solve classification or regression problems (Huang et al., 2010). Kavitha et al. (2015) combined the ELM with fractal feature analysis to assess glaucoma. Bueno-Crespo et al. (2017) put forward a method based on ELM to automatically design a multitask learning machine. At present, most of the researches are to use feature selection and feature extraction methods independently, but we combine feature extraction and feature selection to carry out the follow-up research work in this article.

The present work is concerned with the development of analytical method for rapid identification of breast cancer categorical data based on attribute selection and feature extraction. Firstly, the RF is used for characteristic attribute selection processing of original breast cancer data, and the samples are divided into a training set and test set. Then, feature extraction and dimensionality reduction of selected attribute data by the PCA. Finally, the extracted characteristic data are used as the input of the ELM to establish the identification model of breast malignant tumor. Brief conclusions and future work are summarized at the end of the article.

MATERIALS AND METHODS

Collection of Breast Cancer Data

The validity and feasibility of the methods described in this article were verified by the University of Wisconsin breast cancer data sets (Street et al., 1993). There are 569 cases of breast tumor data in this nuclear micrograph of breast tumor lesion tissue database, including 357 cases of benign tumors and 212 cases of malignant tumors. To facilitate the proportional division of samples, 400 cases were randomly selected as the study objects, including 200 cases of benign tumors and 200 cases of malignant tumors. The quantitative real-valued features of the nuclear micrograph of breast tumor lesion tissue include radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter (sum of the distances between consecutive boundary points), area (perimeter to compensate for digitization error), smoothness (local variation in radius lengths), compactness (perimeter^2/area − 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry (relative difference in length between pairs of line segments perpendicular to the major axis), and fractal dimension ("coastline approximation" − 1). A set of data for each case includes 30 attributes, including the average value, standard deviation, and worst value (the average value of the three largest data of each feature) of the 10 characteristic quantities of each nucleus in the sampled tissue. The 30 attributes were already present from the data sets. Each sample data is composed of 32 fields. The first field is case number, the second field is diagnosis result, B is benign, M is malignant. The other fields are all the attributes of 10 quantitative features, and the first to the tenth attributes are the average value of 10 quantitative features. The 11th to 20th attributes are the standard deviation of 10 quantitative features. The 21st to 30th attributes are the worst value (average value of the three largest data of each feature) of 10 quantitative features. These characteristics can reflect the nature of the breast tumor.

The hardware conditions of the computer used in the experiment include an Intel Core i7 processor, an NVIDIA RTX 2070 graphics card, and a 16G Kingston memory module, etc. The algorithm simulation is run in MATLAB R2016b (MathWorks, United States) environment.

Random Forest for Attribute Selection

The feature selection method is to select features from the original attribute data and get a new feature subset composed of the original features, so as to reduce the number of attributes in the attribute set. It is an inclusive relationship and does not change the original feature space (Guyon and Elisseeff, 2003). RF is a supervised learning algorithm that uses multiple DT to train samples. This algorithm was proposed by Breiman (2001), which can be used to solve classification and regression problems. The RF feature selection method will give the importance score of each variable (Genuer et al., 2010), evaluate the role of each variable in the classification problem, and delete the attribute with lower importance. If a feature is randomly added with noise, the accuracy of out of bag data changes significantly, which shows that this feature has a greater impact on the predictive results of samples. Furthermore, it shows that its importance is high, so it is necessary to select and delete the attributes with low importance. The out of bag error (Mitchell, 2011) is usually used to evaluate the importance of features by RF.

The steps for attribute selection of RF algorithm are as follows:

Step 1: calculate the importance of each attribute and arrange it in descending order of importance

Attribute importance I_m:

$$I_m = \frac{1}{N}\sum(errOOB2 - errOOB1) \tag{1}$$

Where, N is the tree in the RF, $errOOB2$ represents the out of bag error of data with noise interference, and $errOOB1$ denotes the out of bag error of original data;

Step 2: Set the threshold value, delete the attributes whose importance is lower than the threshold value from the current attributes, and the remaining attributes will form a new attribute set again;

Step 3: A new RF is established by using the new attribute set, the importance of each attribute in the attribute sets are calculated and arranged in descending order;

Step 4: Repeat *step 2* and *step 3* until all the attribute importance values are greater than the threshold value;

Step 5: Each attribute set corresponds to a RF, and the corresponding out of bag error rate is calculated;

Step 6: Take the attribute set with the lowest out of bag error rate as the last selected attribute set.

Normalization of Data

Standardization refers to the pre-processing of data so that the values fall into a unified range of values. In the process of modeling, the difference of each feature amount is reduced (He et al., 2010). Different data often have different dimension units and do not belong to the same order of magnitude. The data with a too-large difference will eventually affect the evaluation results. To eliminate the influence of too large dimensional difference between indicators, before using PCA for feature selection, data need to be standardized to solve the error caused by the difference between data indicators (Sun et al., 2016). The common standardization methods are Min − Max normalization (Snelick et al., 2005) and Z-score normalization (Ribaric and Fratric, 2006). Min–max normalization can normalize data to interval [0, 1] and interval [−1, 1] respectively.

[0, 1] normalization:

$$X_{[0,1]} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{2}$$

[−1, 1] normalization:

$$X_{[-1,1]} = \frac{X - \mu}{X_{Max} - X_{Min}} \tag{3}$$

Z-score normalization:

$$X_Z = \frac{X - \mu}{\sigma} \tag{4}$$

Where X is the original sample data, X_{Max} is the maximum value of the original sample data, X_{Min} is the minimum value of the original sample data, μ denotes the average value of the original sample data, and σ represents the standard deviation of the original sample data.

Principal Component Analysis for Feature Extraction

The method of feature extraction is mainly to transform the feature space through the relationship between attributes, map the original feature space to the low-dimensional feature space, so as to complete the purpose of dimension reduction (Wang and Paliwal, 2003). As an unsupervised learning dimensionality reduction method, PCA reduces the data dimension through the correlation between multidimensional data groups. On the premise of minimizing the information loss, it can simplify the data structure, make the data set easier to use, completely without parameter limitation, and reduce the calculation cost of the algorithm (Hess and Hess, 2018).

The steps of the PCA algorithm for feature extraction are as follows:

Step 1: Input the original sample data matrix X:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{5}$$

Step 2: Set each column as a feature and average each feature. Subtract the average value from the original data to the new centralized data;

Step 3: Calculate the covariance matrix:

$$D(X) = \frac{1}{n} XX^T \tag{6}$$

Step 4: Solve eigenvalue λ and eigenvector q of covariance matrix by the eigenvalue decomposition method;

Step 5: Sort the eigenvalues from large to small, and select the largest k of them. Then the corresponding k eigenvectors are used as row vectors to form eigenvector matrix Q;

Step 6: Multiply the data set m^*n by the eigenvector of n dimensional eigenvector, and obtain the data matrix $Y = QX$ of the last dimension reduction.

As the basis of selecting the number k of principal components, the cumulative contribution rate of principal components is generally required to be more than 85%.

Extreme Learning Machine for Classification

The ELM is a simple and efficient learning algorithm proposed by professor Huang (Huang et al., 2006) of the Nanyang Polytechnic, it can be used to solve the problem of classification and regression in pattern recognition. This algorithm only needs to set the number of hidden layer neurons of the network, it does not need to adjust the input weight of the network and the bias of hidden layer neurons in the process of implementation, and produces a unique optimal solution, so the learning speed is fast and the generalization performance is good (Zhang and Ding, 2017). ELM is a single-layer feedforward network that can train training set quickly. There are only three layers in the network, namely the input layer, the hidden layer, and the output layer. The network structure of ELM is shown in **Figure 1**. From left to right, there are input layer neurons, hidden layer neurons, and output layer neurons.

There are S different training samples s, where $x_i = [x_1, x_2, x_3, \cdots, x_m]^T$, $x_i \in R^m . p_i = \left[p_1, p_2, p_3, \cdots, p_n\right]$, $t_i \in R^n$. Set the activation function $g(x)$, with K hidden layer nodes output as follows:

$$p_i = \sum_{i=1}^{K} \beta_i g(\omega_i \cdot x_j + b_i) = \sum_{i=1}^{K} \beta_i F(\omega_i, b_i, x_j) \tag{7}$$

Where $j = 1, 2, \cdots, N$, $\omega_i = [\omega_1, \omega_2, \cdots, \omega_m]^T$ is the input weight of the hidden layer neuron, b_i is the hidden layer neuron bias, and $\beta_i = [\beta_1, \beta_2, \cdots, \beta_n]^T$ is the output weight of the output neuron.

The steps of the ELM algorithm are as follows:

Step 1: Select (ω_i, b_i) randomly and map the samples to the feature space according to $h(x) = \left[F(\omega_1, b_1, x), \cdots F(\omega_K, b_K, x)\right]^T$. If the feature mapping $h(x)$ forms the hidden layer matrix H, then it exists

$$H\beta = P \tag{8}$$

Where $H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_S) \end{bmatrix} = \begin{bmatrix} F(\omega_1, b_1, x_1) & \cdots & F(\omega_K, b_K, x_1) \\ \vdots & & \vdots \\ F(\omega_1, b_1, x_S) & \cdots & F(\omega_K, b_K, x_S) \end{bmatrix}_{S \times K}$,

$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_K^T \end{bmatrix}_{K \times n}$ and $P = \begin{bmatrix} p_1^T \\ \vdots \\ p_S^T \end{bmatrix}_{S \times \, n}$.

Sin function, Hardlim function, and Sigmoid function can be selected as the activation function of hidden layer neurons (Song et al., 2015).*

Step 2: In the new feature space, the optimal output weight β^* is obtained from Eq. (8) by using the least square method, where H^+ is the Moore-Penrose generalized inverse of H, $\beta^* = \mathrm{H}^+ P$.

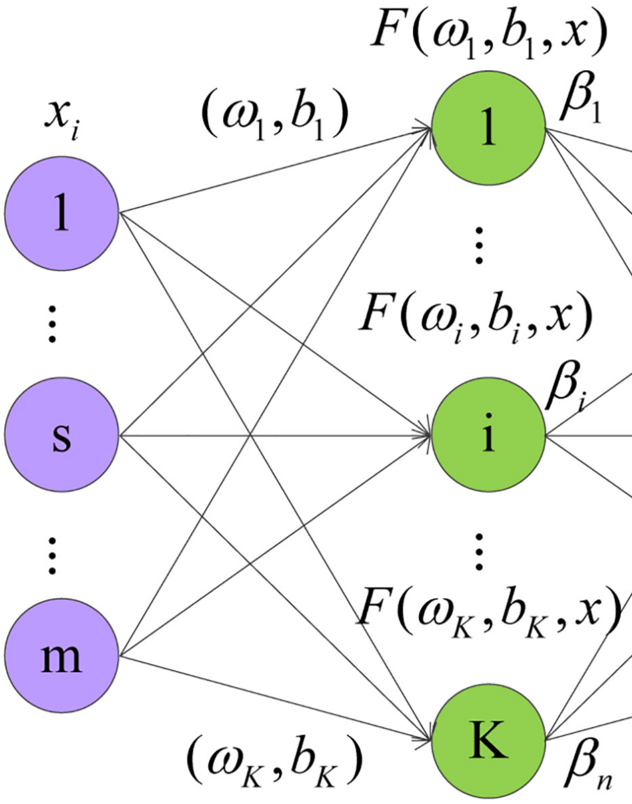


FIGURE 1 | Network structure diagram of ELM.

Evaluation Index of Classifier Performance

In order to better evaluate the performance of classifier, we introduce the confusion matrix. In the field of machine learning, confusion matrix is a visual tool to evaluate the performance of classification models. Among them, each column of the matrix represents the situation of predictive samples and each row of the matrix represents the situation of actual samples (Deng et al., 2016). The confusion matrix consists of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The accuracy, precision, sensitivity, specificity, F1-score and MCC (Azar and El-Said, 2012; Zheng et al., 2018) can be obtained from the confusion matrix and all of them are used as evaluation indexes of performance. In general,

Accuracy is the ratio of the correctly classified examples to the total sample size.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

Precision is the percentage of samples are correctly classified as true positive.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

Sensitivity is the percentage of samples are correctly classified as true positive in total positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

Specificity is the percentage of samples are correctly classified as true negative in total negative samples.

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

F1-score is an index used to measure the accuracy of a binary classification model.

$$F1 - score = \frac{2 * Precision * Sensitivity}{TN + FP} \tag{13}$$

MCC is essentially a balanced index that describes the correlation coefficient between the actual classification and the predicted classification, which is used to measure the classification performance of binary classification. The value range of MCC is [−1,1]. The closer the MCC value is to 1, the better the classifier performance.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

RESULTS AND DISCUSSION

Attribute Selection Based on Random Forest

There are 30 attributes in the original breast cancer data, each of which contains the corresponding information of breast tumor lesion tissue. Different attributes play different roles in the analysis of breast cancer data. Redundant and less important attributes will affect the establishment of breast cancer of a predictive model, which cannot achieve high prediction accuracy, but also increase the complexity of the model and reduce the efficiency of breast cancer prediction. Attribute selection based on RF of the method is used to select more important attributes to improve the efficiency of modeling and prediction ability. Before RF is used, we set the number of trees to 200, the number of leaf node samples to 1, and the number of fboot to 1.

The importance ranking of the first selected attribute is shown in **Figure 2**. From the top to the bottom, the importance of attributes is sorted according to the order of importance from the largest to the smallest. We can find that there are significant differences in the importance of each attribute. The 28th attribute is the most important, with a value of 0.96. The 20th attribute is the least important, which is the standard deviation of the quantitative features, with a value of only 0.05. The areas with high importance are mainly concentrated in the 21st to 24th attribute range and the 28th attribute, which are the worst values of the quantitative features, all of which are above 0.8. This shows that the worst value of the quantitative characteristics of nuclear micrograph covers a large amount of important information about data. However, the importance of the 16th, 10th, and 20th attributes is less than 0.1, which indicates that the importance of these three attributes is very low and the influence on the predictive results of breast cancer is very small, which belongs to redundancy attribute information.

The threshold value of attribute selection based on RF is set to 0.1, the attributes whose importance is lower than the threshold value are deleted, and the remaining 27 attributes are selected as the result of RF initial attribute selection. 27 attributes of the first reduction are continued to be selected by RF. We delete the redundant attributes whose importance is lower than the threshold value, calculate the importance of the remaining attribute sets and each attribute in it, and arrange them in descending order of importance.

The ranking of attribute importance for four iterations is shown in **Figure 3**. Because of the randomness of RF, it can be seen from **Figures 3A,B** that there are differences in the ranking of the importance of the first two attributes. The

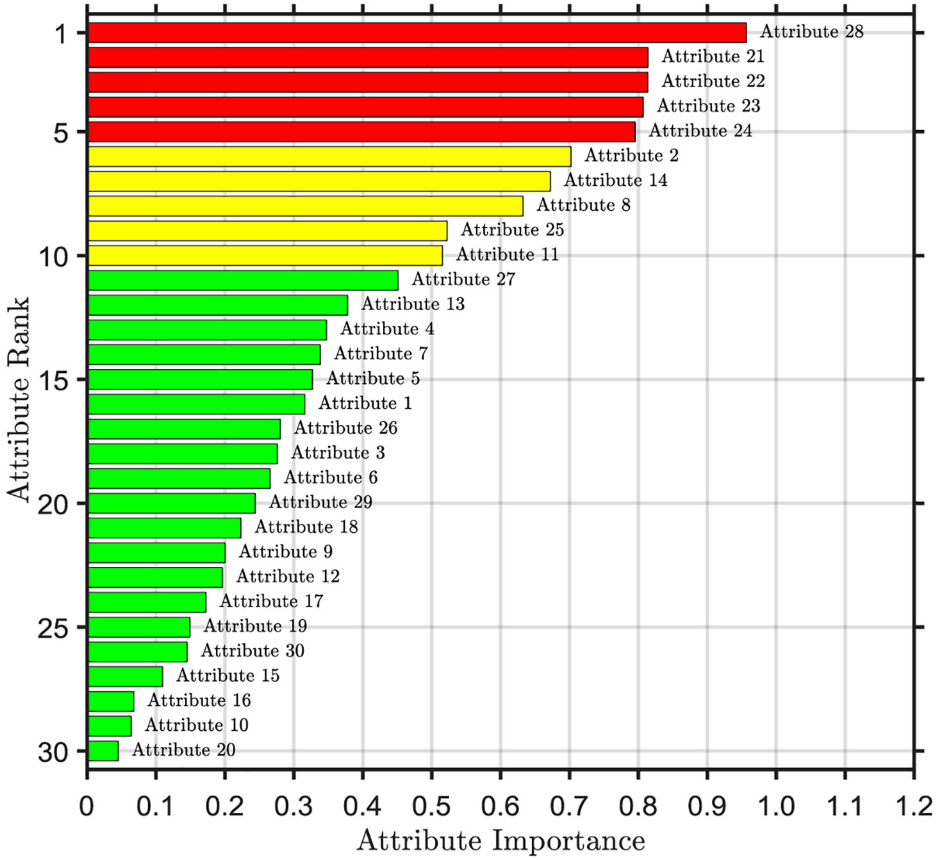


FIGURE 2 | Ranking of attribute importance for RF initial selection.

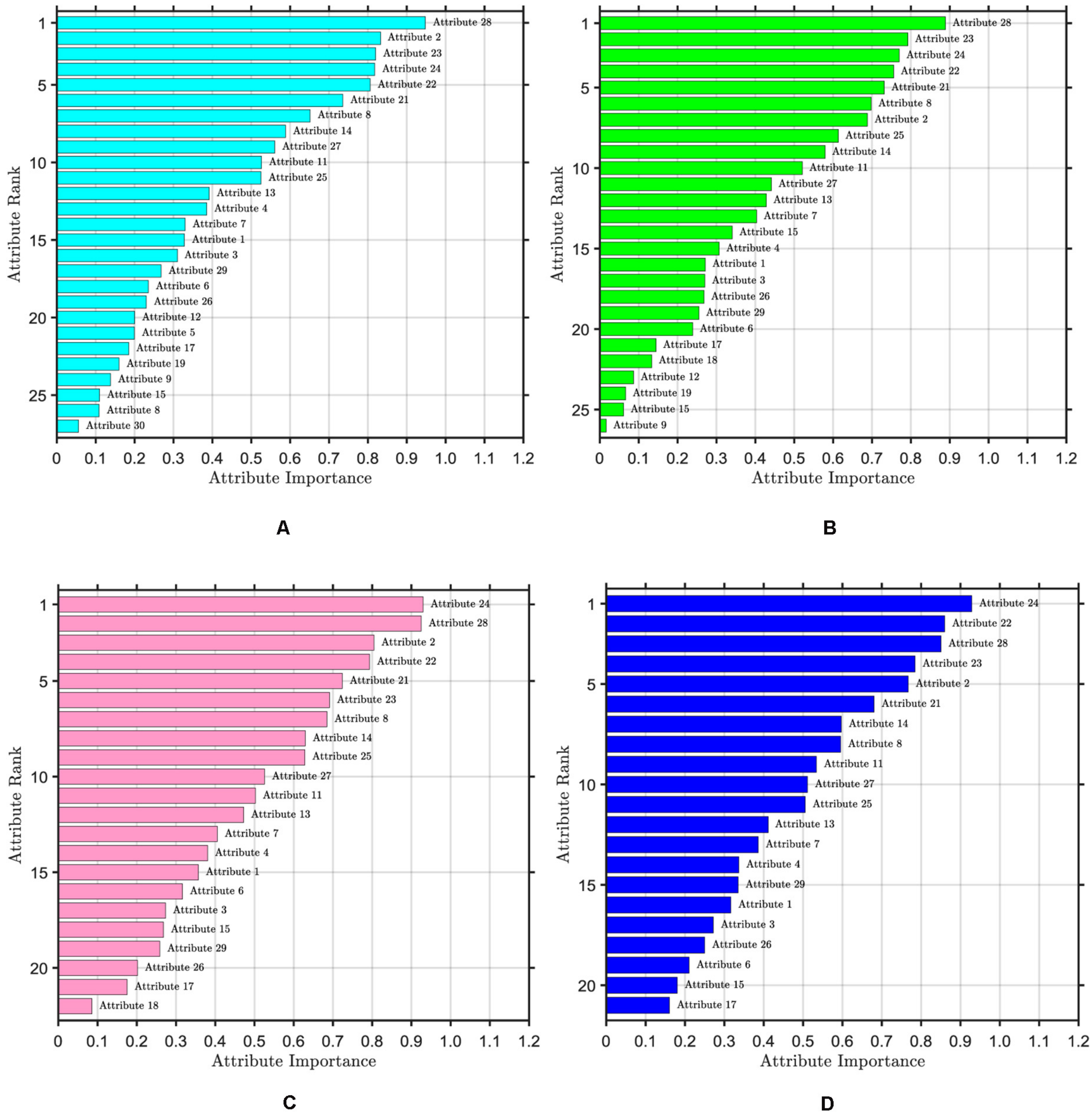


FIGURE 3 | Ranking of attribute importance. The threshold value of attribute selection based on RF is set to 0.1. **(A)** Ranking of attribute importance after one iteration, including 27 attributes. **(B)** Ranking of attribute importance after two iteration, including 26 attributes. **(C)** Ranking of attribute importance after three iteration, including 22 attributes. **(D)** Ranking of attribute importance after four iteration, including 21 attributes.

maximum value of attribute importance for both iterations is obtained at the 28th attribute. After the first iteration, only the 30th attribute is below the threshold, while the second is the 12th, 19th, 15th, and 9th attributes. As can be seen from **Figures 3C,D**, compared with the previous two iterations, the attribute of the maximum importance has changed, which is the 24th attribute. After the third iteration, only the importance of the 18th attribute is below the threshold, and after the fourth iteration, the importance of all attributes is greater than the threshold. We take the average attribute importance and out of bag error as the evaluation indexes of attribute selection based on RF. The larger the average attribute importance of attributes and the smaller the out of bag error, the more useful information these attributes contain, the less redundant information they have. The evaluation indexes of five iterations are presented in **Table 1**. From the table, we can see that with the increasing number of iterations, redundant attributes are gradually eliminated, and the corresponding evaluation indicators are also changing. When the number of iterations is 4, the average attribute importance reaches a maximum of 0.5214, the out of bag error reaches a minimum of 0.0318, and the number of attributes selected by RF is 21. In the fifth iteration, the importance of each attribute is still greater than the threshold, the number of attributes selected by RF remains unchanged, and each attribute retains the relatively important and effective information of breast cancer data. Finally, 21 attributes selected by four iterations are the result of the attribute selection of the RF algorithm.

Feature Extraction Based on Principal Component Analysis

After RF selection, the number of attributes is reduced by 9 compared with the original data, and there is a lot of redundant information in these 9 attributes. In order to achieve the requirement of accurate prediction of breast cancer, PCA needs to be used to further simplify the data attributes. When PCA is used to extract features, to prevent PCA from over capturing some features with large values, which results in the loss of a large amount of information and the impact of features with large values on the results, we will standardize each feature first, so that their sizes are within the same range. PCA is employed to extract the 21 attributes of breast cancer data after attribute selection, and the cumulative contribution rate is 95%.

The 160 samples of each group are selected, and a total of 320 samples of breast cancer data are used as the training set. The remaining 40 samples of each group are selected, and a total of 80 samples of breast cancer data are used as the test set. [0, 1], [−1, 1], and *Z*-score normalization methods are used to normalize the breast cancer data after feature selection. The training set is used to establish the predictive model of breast cancer based on ELM, and the test set is used to test the prediction ability of the model. Under different standardized methods, we input the data of feature extraction into the predictive model of ELM, and compare their prediction accuracy of the training set and test set, then select the best normalization method.

The predictive results of different normalization methods are shown in **Table 2**. It can be seen that the main component scores of [0, 1] and [−1, 1] normalization methods are only two, and the accuracy of the training set is relatively low. The prediction accuracy of the *Z*-score of the training set and test set is significantly higher than the other two methods, which fully shows that the proper selection of the data standardization method plays a key role. *Z*-score is selected as the normalization method of data.

From the variance contribution rate of the principal components in **Figure 4**, we can see that the first principal component bears 56.43% of the difference. The variance contribution rate of the first principal component is the largest, and the variance contribution rate of the other principal components is decreasing in turn, then seven principal components can be obtained. The cumulative contribution rate of principal components is shown in **Table 3**. The cumulative contribution rate of the first seven principal components is 95.99%, which achieves the goal of 95%. Therefore, the first seven principal components are selected as the feature of PCA extraction. Finally, the dimension of breast cancer data is reduced to 7 dimensions, which is more conducive to the subsequent recognition and prediction of breast cancer.

TABLE 1 | Evaluation indexes of five iterations.

Iterative number	Attributes	Average attribute importance	Out of bag error
1	27	0.4315	0.0335
2	26	0.4381	0.0320
3	22	0.4792	0.0337
4	21	0.5214	0.0318
5	21	0.4987	0.0322

TABLE 2 | Predictive results of different normalization methods.

Normalization method	Principal components	Predictive accuracy/%	
		Training set	Test set
[0,1]	2	90.94 (291/320)	96.25 (77/80)
[−1,1]	2	90.63(290/320)	95 (76/80)
Z-score	7	99.06 (317/320)	98.75 (79/80)

Predictive Models for Breast Cancer

The prediction performance of the ELM model is affected by the type of activation function. By comparing and analyzing the predictive results of breast cancer under three different activation functions of sin, hardlim and sigmoid, the activation function with the best prediction effect was selected. Seven feature data are used to establish the predictive model of ELM under different activation functions, and the predictive results are shown in **Table 4**. When the Sigmoid function is used as the ELM activation function, both the training set and the test set have higher prediction accuracy.

In the predictive model of ELM, the number of input layer neurons, hidden layer neurons, and output layer neurons and network structure should be determined. The number of extracted features is 7, so the number of input layer neurons is 7. Because two types of breast tumors are predicted, the number of output neurons is 2. The number of hidden layer neurons is the key parameter that affects the prediction ability and generalization performance of ELM. The initial number of neurons in the hidden layer is set to 1. It is necessary to analyze the prediction of breast cancer by the ELM model corresponding to the number of different hidden layers. In order to reduce the training time of the model, the number of hidden layer neurons is set within 200.

As shown in **Figure 5**, when there is only one neuron in the hidden layer, the prediction accuracy of the test set is only 50%. When the number of hidden layer neurons is 2, the prediction accuracy increases to 91.25%. The number of neurons increases from 3 to 5, the prediction accuracy gradually increases to a higher value of 92.5%, and then began to fluctuate in the range of 81∼99%. The overall trend is relatively stable, and the average accuracy is about 92%. However, it is not that the more the number of hidden layer neurons, the better the prediction effect of the model. After the number of hidden layer neurons reaches 120, the accuracy of the test set fluctuates greatly, ranging from 81 to 96%, and the average accuracy is about 90%. When the number of hidden layer neurons is 27, the ELM model has the best prediction effect on the test set, and the prediction accuracy reaches 98.75%. **Figure 6** shows the relationship between the number of hidden layer neurons and the training time. We

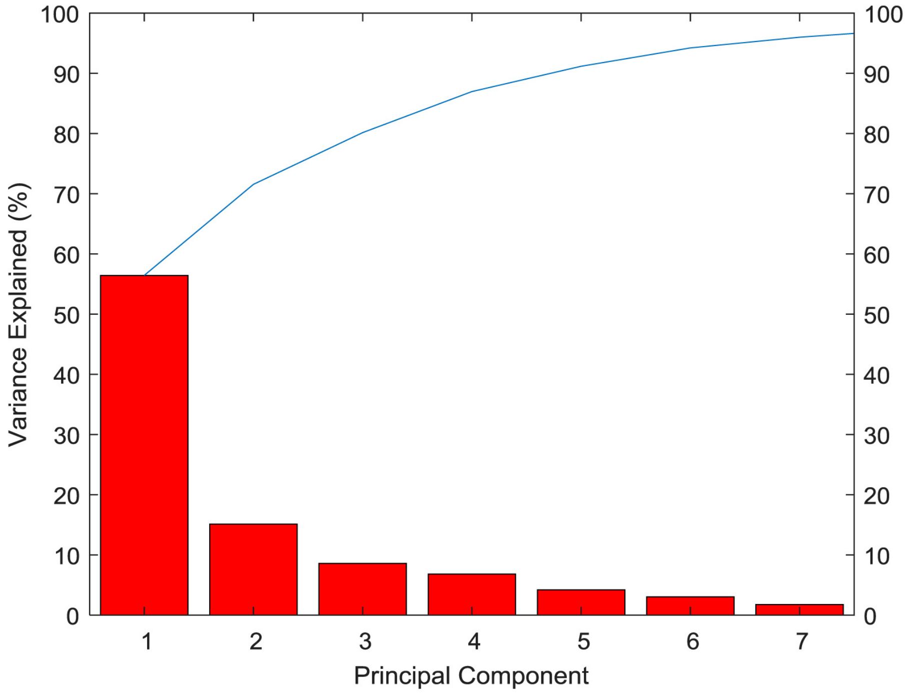


FIGURE 4 | Variance contribution rate of the principal components.

can find that with the increase in the number of hidden layer neurons, the overall training time is on the rise. Compared with **Figure 5**, when the prediction accuracy reaches the maximum, the number of hidden layer neurons is 27, and the training time is only 0.0022 s.

In order to prove the reliability of attribute selection and feature extraction algorithm for breast cancer data modeling, the predictive results of the original data, the data after attribute selection, and the data after feature extraction are compared and analyzed, and the results are shown in **Table 5**. It can be seen that the accuracy of the training set and test set after dimension reduction is higher than that of original data modeling, which shows that attribute selection and feature extraction methods improve the predictive learning ability of model training and test samples. The number of features obtained by single RF and PCA dimensionality reduction methods is less than that of the original data, and the number of features is reduced to 70 and 33% of the original data, respectively. RF combined with PCA (RF-PCA) process the original data to get the least number of features and the number of features is only 23% of the original data. The accuracy of the training set and test set is not only higher than that of original data modeling but also higher than that of single RF and single PCA modeling. Because the classifiers used are ELM, so there is little difference in training time, only about 0.002 s, and the number of hidden layer neurons corresponding to the optimal accuracy is different.

In order to verify the superiority of the predictive model based on breast cancer data after RF-PCA dimensionality reduction, we also compared and analyzed the prediction performance of several different modeling methods based on the data after dimension reduction, such as a PNN, SVM, BP neural network, and DT. The optimal parameter *spread* of PNN is set to 0.87. The radial basis function (RBF) is used as a kernel function of SVM. SVM uses a fivefold cross-validation method to find the best penalty coefficient C and kernel function parameter g in the range of $[2^{-10}, 2^{10}]$, at which point, $C = 2.2974$, $g = 0.0625$. BP adopts the same network structure as ELM, in which the number of hidden layer neurons is 27, the learning step of BP is set to 0.3, the minimum mean square error is set to 10^{-8}, and the minimum gradient is set to 10^{-20}.

The predictive results of different modeling methods are shown in **Table 6**. According to the accuracy (Acc), precision

TABLE 3 | Cumulative contribution of principal components.

Principal component	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7
Cumulative contribution rate/%	56.43	71.56	80.15	86.97	91.17	94.22	95.99

TABLE 4 | Predictive results of different activation functions.

Activation function	Time/s	Predictive accuracy/%		Hidden layer neurons
	Training samples	Training set	Test set	
Sin	0.0067	97.81 (313/320)	95 (76/80)	104
Hardlim	0.0029	98.13 (314/320)	98.75 (79/80)	53
Sigmoid	0.0022	99.06 (317/320)	98.75 (79/80)	27

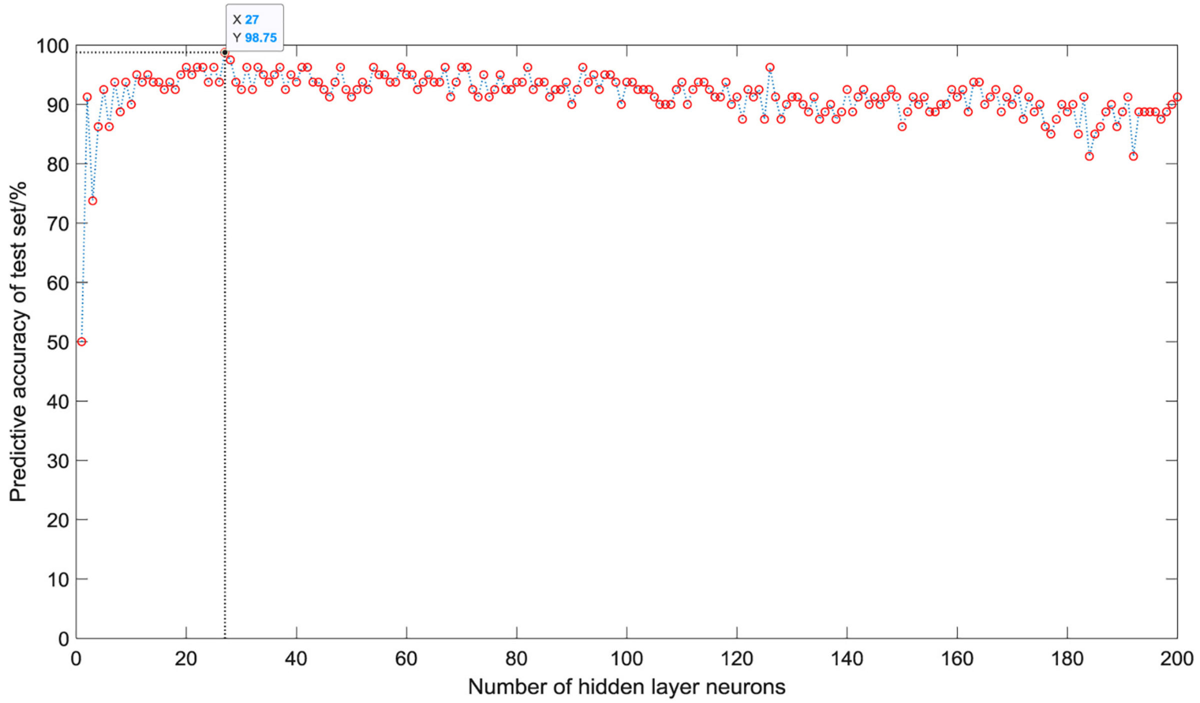


FIGURE 5 | Predictive accuracy of different hidden layer neurons.

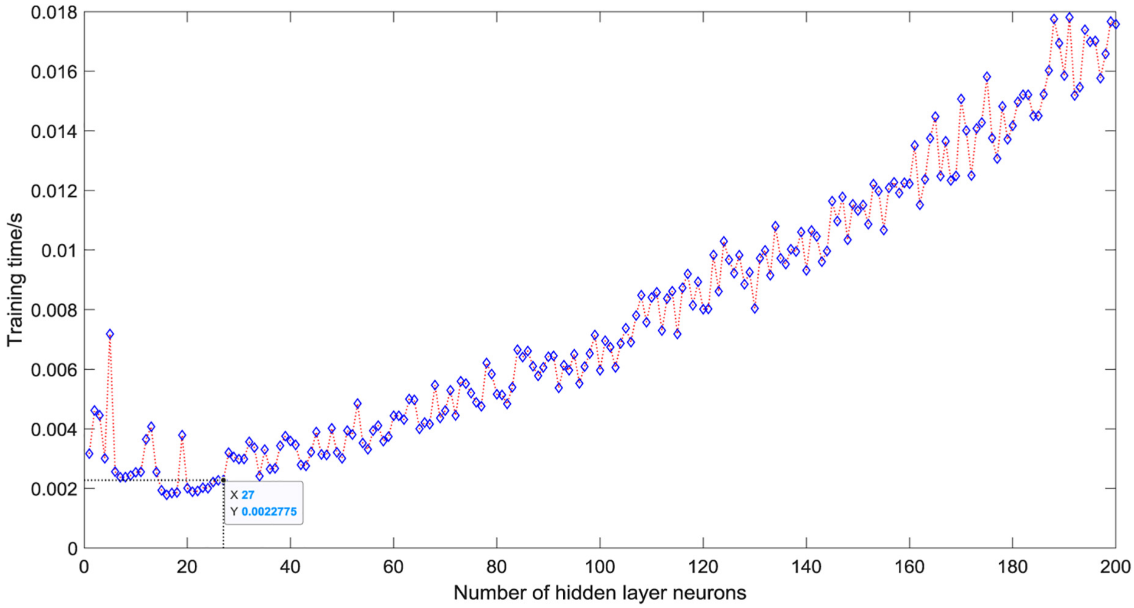


FIGURE 6 | Training time of different hidden layer neurons.

(Pr), sensitivity (Se), specificity (Sp), F1-score (F1) and MCC, we find that although the accuracy and other evaluation indexes of the BP training set is as high as 100% and higher than that of other models. The accuracy of the test set are the lowest and other evaluation indexes are relatively low, which indicates that BP based on gradient descent method has slight over-fitting. The training time of BP is 9.6259 s, and the prediction speed is obviously slower than other methods. The accuracy of the test set of PNN, SVM, and DT is 95%, and their MCC are all 0.9, which shows that they have similar prediction performance, and the difference is mainly reflected in the evaluation index of the training set and training time.

TABLE 5 | Predictive results of different dimensionality reduction methods.

Dimension reduction method	Features	Predictive accuracy/%		Time/s	Hidden layer neurons
		Training set	Test set	Training samples	
ELM	30	95.31 (305/320)	95 (76/80)	0.0020	14
RF + ELM	21	97.5 (312/320)	96.25 (77/80)	0.0023	24
PCA + ELM	10	97.19 (311/320)	97.5 (78/80)	0.0028	13
RF-PCA + ELM	7	99.06 (317/320)	98.75 (79/80)	0.0022	27

TABLE 6 | Predictive results of different modeling methods.

Modeling method	Time/s	Training set						Test set					
	Training	Acc	Pr	Se	Sp	F1	MCC	Acc	Pr	Se	Sp	F1	MCC
PNN	0.0339	99.69%	99.38%	100%	99.38%	99.69%	0.99	95%	95%	95%	95%	95%	0.9
SVM	1.4601	99.06%	98.16%	100%	98.13%	99.07%	0.98	95%	97.37%	92.5%	97.5%	94.87%	0.9
BP	9.6259	100%	100%	100%	100%	100%	1	93.75%	92.68%	95%	92.5%	93.83%	0.88
DT	0.1669	98.13%	98.13%	98.13%	98.13%	98.13%	0.96	95%	95%	95%	95%	95%	0.9
ELM	0.0022	99.06%	98.16%	100%	98.13%	99.07%	0.98	98.75%	97.56%	100%	97.5%	98.76%	0.98

TABLE 7 | Predictive results of dimensionality reduction by RF-PCA.

Modeling method	Time/s	Training set						Test set					
	Training	Acc	Pr	Se	Sp	F1	MCC	Acc	Pr	Se	Sp	F1	MCC
Raw + ELM	0.0096	95%	85.71%	100%	85.71%	92.31%	0.86	92.31%	92.86%	97.5%	92.5%	95.12%	0.9
ELM	0.0011	100%	100%	100%	100%	100%	1	96.15%	92.31%	100%	92.86%	96%	0.93
PNN	0.0314	91.25%	94.59%	87.5%	95%	90.91%	0.83	88.46%	80%	100%	78.57%	88.89%	0.79
SVM	0.1592	97.5%	95.24%	100%	95%	97.56%	0.95	96.15%	100%	91.67%	100%	95.65%	0.93
BP	1.3080	100%	100%	100%	100%	100%	1	84.62%	78.57%	91.67%	78.57%	84.62	0.7
DT	0.0551	96.25%	93.02%	100%	92.5%	96.39%	0.93	92.31%	91.67%	91.67%	92.86%	91.67%	0.85

In the comparison of these three methods, the training set of PNN has the highest prediction performance and the training speed of PNN is the fastest. Finally, by comprehensively comparing the evaluation indexes of training time, Acc, Pr, Se, Sp, F1, and MCC, we can clearly see that the training time of ELM is much faster than other models, and the evaluation index of predictive performance is better than other models, which fully verifies the superiority of RF-PCA combined with ELM, and meets the requirements of real-time breast cancer auxiliary diagnosis.

The same algorithm can be applied to different data sets to ensure the reliability of the algorithm. If the algorithm proposed in this article can achieve good prediction results for different data sets, it can show that the algorithm has strong adaptability and generalization performance. The generalization performance of the algorithm is verified by the data (Jossinet, 1996) in UCI database. The data was obtained by jossinet's team using electrical impedance tomography to measure the impedance of 106 pathological breast tissue from 64 women. The sample were divided into pathological tissue and normal tissue, according to the pathology and morphology of the breast. Among them, pathological tissue includes mastopathy: benignant and non-inflammatory disease of the breast (MA), fibro-adenoma (FA) and carcinoma (CA), while normal tissue includes mammary gland (MG), connective tissue (CT) and adipose subcutaneous fatty tissue (AT). A total of 80 samples were randomly divided into training sets and the remaining 26 samples were used as test sets. Firstly, RF is used for attribute selection, and then PCA is used for feature extraction. Finally, the dimension of data is reduced from 9 to 4 dimensions. The reduced dimension data of RF-PCA is fed into ELM and a predictive model is established.

We also compare common methods of classifiers used in the literature about breast cancer recognition for the new data. The reduced dimension data is fed into other classifiers and a predictive model is established. When the number of neurons in the hidden layer was 97, the ELM model has a best prediction performance. The optimal parameter *spread*of PNN is set to 0.68. The optimal C of SVM is 42.2243 and g is 9.1896. **Table 7** is the comparison between the predictive results of the ELM model of the data after dimensionality reduction by RF-PCA and the raw data. It can be seen that the prediction performance of the training set and the test set of the ELM model established by the dimensionality reduction method of RF-PCA is higher than that of the ELM model established by the raw data. All the samples of the training set are predicted correctly, and only one sample of the test set is predicted incorrectly. The number of features of the data is almost reduced to half of the raw data, and the training time is only about 0.0011 s. In the training set, we find that the prediction performance of PNN is the worst. BP has the same prediction performance as ELM, but the training time is the longest. In the test set, SVM has a similar prediction performance as ELM, and a faster training speed. BP has the worst prediction performance. Comparing all kinds of evaluation indexes of the model, it can be seen that ELM and SVM have a good prediction effect and the fast training time in the electrical impedance data, but the performance of the model established by the method of BP is poor.

All of these shows that the method proposed in this article can still achieve a better prediction performance and faster speed when applied to the new dataset to predict new samples. To a certain extent, the proposed method can exclude the possibility of overfitting of the models.

CONCLUSION

In this article, we put forward a new solution based on attribute selection and feature extraction for rapid diagnosis of breast cancer, which is called RF-PCA. Firstly, we used the attribute selection based on RF of algorithm to select the useful attributes of quantitative feature data of breast tumor cell images and then used the feature extraction algorithm based on PCA to reduce the dimension of data after attribute selection. Finally, the ELM model was established to test the prediction effect of breast cancer. In order to verify the reliability of this algorithm, we compared the prediction accuracy of ELM model after using RF or PCA alone. To verify the superiority of this algorithm, we also compared the prediction performance of different models and used the impedance data of the breast tissue to verify the adaptability of the algorithm.

The results show that (1) The feature selection based on RF or feature extraction based on PCA of a method can not only reduce the complexity of the training model but also improve the prediction accuracy of the model to a certain extent; (2) Combining feature selection with feature selection, we use the advantages of the two methods to reduce the dimension of data. Compared with the single dimension reduction method, it can reflect the effective information of the original data with fewer features, make the model simple, and improve the efficiency and reliability of modeling; (3) ELM model has high prediction accuracy and short training time, which effectively avoids over-fitting and has a certain generalization ability; (4) RF-PCA combined with ELM model can significantly reduce the training time of the network, and more adapt to the requirements of a rapid and accurate breast cancer aided diagnosis.

Despite the achievement of some research results, there are some limitations in this study. When the proposed algorithm in this article is used in breast cancer diagnosis, the training time is reduced and the prediction accuracy is better. However, these advantages mainly focus on the fast prediction speed and does not reach the optimal accuracy of all samples. Therefore, in future work, it will be necessary to study some optimization algorithms to improve the performance of the model and achieve the highest prediction accuracy on the basis of ensuring faster prediction speed.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29, https://archive.ics.uci.edu/ml/datasets/Breast+Tissue. We have uploaded the code by using Matlab to GitHub. With this URL (https://github.com/bkfly/test.git), you can easily download the code of this article. In addition, you can visit https://help.github.com/en#dotcom for more instructions on the use of GitHub.

AUTHOR CONTRIBUTIONS

KB conceived the study. MZ developed the method and supervised the study. KB and WL implemented the algorithms. KB and FH analyzed the data. KB wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This study was supported by the Major Science and Technology Program of Anhui Province (No. 201903a07020013), the National Key Research and Development Program of China (No. 2018YFC0604503), and the New Generation of Information Technology Innovation Project (No. 2019ITA01010).

ACKNOWLEDGMENTS

We would like to thank the Departments of Computer Science, Surgery, and Human Oncology at the University of Wisconsin, Madison, United States to make the database available.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.566057/full#supplementary-material

REFERENCES

Azar, A. T., and El-Said, S. A. (2012). Probabilistic neural network for breast cancer classification. *Neural Comput. Appl.* 23, 1737–1751. doi: 10.1007/s00521-012-1134-8

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/a:1010933404324

Bueno-Crespo, A., Menchón-Lara, R.-M., Martínez-España, R., and Sancho-Gómez, J.-L. (2017). Bioinspired architecture selection for multitask learning. *Front. Neuroinform.* 11:39. doi: 10.3389/fninf.2017.00039

Charaghvandi, R. K., Van Asselen, B., Philippens, M. E. P., Verkooijen, H. M., Van Gils, C. H., Van Diest, P. J., et al. (2017). Redefining radiotherapy for early-stage breast cancer with single dose ablative treatment: a study protocol. *BMC Cancer* 17:181. doi: 10.1186/s12885-017-3144-5

Cui, X., Li, Z., Zhao, Y., Song, A., and Zhu, W. (2018). Breast cancer identification via modeling of peripherally circulating mirnas. *PeerJ* 6:e4551.

Deng, X., Liu, Q., Deng, Y., and Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* 34, 250–261. doi: 10.1016/j.ins.2016.01.033

Dennison, G., Anand, R., Makar, S. H., and Pain, J. A. (2015). A prospective study of the use of fine-needle aspiration cytology and core biopsy in the diagnosis of breast cancer. *Breast J.* 9, 491–493. doi: 10.1046/j.1524-4741.2003.09611.x

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., et al. (2019). Breast cancer statistics, 2019. *CA A Cancer J. Clin.* 69, 438–451. doi: 10.3322/caac.21583

Fabris, C., Facchinetti, A., Sparacino, G., Zanon, M., and Cobelli, C. (2014). Glucose variability indices in type 1 diabetes: parsimonious set of indices revealed by sparse principal component analysis. *Diabetes Technol. Ther.* 16, 644–652. doi: 10.1089/dia.2013.0252

Garbis, S. D., Manousopoulou, A., Underwood, T. J., Hayden, A. L., and White, C. H. (2018). Quantitative proteomic profiling of primary cancer-associated fibroblasts in oesophageal adenocarcinoma. *Br. J. Cancer* 118, 1200–1207. doi: 10.1038/s41416-018-0042-9

Gebauer, J., Higham, C., Langer, T., Denzer, C., and Brabant, G. (2018). Long-term endocrine and metabolic consequences of cancer treatment: a systematic review. *Endocr. Rev.* 40, 711–767. doi: 10.1210/er.2018-00092

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognit. Lett.* 31, 2225–2236. doi: 10.1016/j.patrec.2010.03.014

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182. doi: 10.1162/153244303322753616

He, M., Horng, S.-J., Fan, P., Run, R.-S., Chen, R.-J., Lai, J.-L., et al. (2010). Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognit.* 43, 1789–1800. doi: 10.1016/j.patcog.2009.11.018

Hess, A. S., and Hess, J. R. (2018). Principal component analysis. *Transfusion* 58, 1580–1582.

Huang, G.-B., Ding, X., and Zhou, H. (2010). Optimization method based extreme learning machine for classification. *Neurocomputing* 74, 155–163. doi: 10.1016/j.neucom.2010.02.019

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Jhajharia, S., Verma, S., and Kumar, R. (2016). "A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data," in *Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT)* (Coimbatore: IEEE), doi: 10.1109/inventive.2016.7830107

Jossinet, J. (1996). Variability of impedivity in normal and pathological breast tissue. *Med. Biol. Eng. Comput.* 34, 346–350. doi: 10.1007/bf02520002

Kavitha, S., Duraiswamy, K., and Karthikeyan, S. (2015). Assessment of glaucoma using extreme learning machine and fractal feature analysis. *Int. J. Ophthalmol.* 8, 1255–1257. doi: 10.3980/j.issn.2222-3959.2015.06.33

Mert, A., Kılıç, N., and Akan, A. (2014). An improved hybrid feature reduction for increased breast cancer diagnostic performance. *Biomed. Eng. Lett.* 4, 285–291. doi: 10.1007/s13534-014-0148-9

Mitchell, M. W. (2011). Bias of the random forest out-of-bag (oob) error for certain input parameters. *Open J. Stats.* 01, 205–211. doi: 10.4236/ojs.2011.13024

Nahato, K. B., Harichandran, K. N., and Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. *Comput. Math. Methods Med.* 2015, 1–13. doi: 10.1155/2015/460189

Noorul, W., Asifullah, K., and Soo, L. Y. (2019). Transfer learning based deep cnn for segmentation and detection of mitoses in breast cancer histopathological images. *Microscopy* 68, 216–233. doi: 10.1093/jmicro/dfz002

Odindi, J., Adam, E., Ngubane, Z., Mutanga, O., and Slotow, R. (2014). Comparison between WorldView-2 and SPOT-5 images in mapping the bracken fern using the random forest algorithm. *J. Appl. Remote Sens.* 8:083527. doi: 10.1117/1.jrs.8.083527

Ribaric, S., and Fratric, I. (2006). "Experimental evaluation of matching-score normalization techniques on different multimodal biometric systems," in *Proceedings of the IEEE Mediterranean Electrotechnical Conference 2006* (Piscataway, NJ: IEEE)498–501. doi: 10.1109/MELCON.2006.1653147

Saraswat, M., and Arya, K. V. (2014). Feature selection and classification of leukocytes using random forest. *Med. Biol. Eng. Comput.* 52, 1041–1052. doi: 10.1007/s11517-014-1200-8

Sewak, M., Vaidya, P., Chan, C.-C., and Duan, Z.-H. (2007). "SVM Approach to Breast Cancer Classification," in *Proceedings of the 2nd International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)* (Iowa City, IA: IEEE), doi: 10.1109/imsccs.2007.46

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA A Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Simas Filho, E. F., and Seixas, J. M. (2016). Unsupervised statistical learning applied to experimental high-energy physics and related areas. *Int. J. Mod. Phys. C* 27:1630002. doi: 10.1142/s0129183116300025

Skala, M., Rosewall, T., Dawson, L., Divanbeigi, L., Lockwood, G., Thomas, C., et al. (2007). Patient-assessed late toxicity rates and principal component analysis after image-guided radiation therapy for prostate cancer. *Int. J. Radiat. Oncologybiol.* 68, 690–698. doi: 10.1016/j.ijrobp.2006.12.064

Snelick, R., Uludag, U., Mink, A., Indovina, M., and Jain, A. (2005). Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 450–455. doi: 10.1109/tpami.2005.57

Song, Z., Jiang, A., and Jiang, Z. (2015). Back analysis of geomechanical parameters using hybrid algorithm based on difference evolution and extreme learning machine. *Math. Probl. Eng.* 2015, 1–11. doi: 10.1155/2015/821534

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). "Nuclear feature extraction for breast tumor diagnosis," In *Proceedings of the SPIE 1905, Biomedical Image Processing and Biomedical Visualization* (Washington, DC: SPIE)861–870. doi: 10.1117/12.148698

Sun, W., Jin, J. H., Reed, M. P., Gayzik, F. S., Danelson, K. A., Bass, C. R., et al. (2016). A method for developing biomechanical response corridors based on principal component analysis. *J. Biomech.* 49, 3208–3215. doi: 10.1016/j.jbiomech.2016.07.034

Wade, B. S. C., Joshi, S. H., Gutman, B. A., and Thompson, P. M. (2016). Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods. *Pattern Recognit.* 63, 36–43. doi: 10.1016/j.patcog.2016.09.034

Wang, H., Zheng, B., Yoon, S. W., and Ko, H. S. (2017). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Operat. Res.* 267, 687–699. doi: 10.1016/j.ejor.2017.12.001

Wang, R., Yin, Z., Liu, L., Gao, W., Li, W., Shu, Y., et al. (2018). Second primary lung cancer after breast cancer: a population-based study of 6,269 women. *Front. Oncol.* 8:427. doi: 10.3389/fonc.2018.00427

Wang, X., and Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognit.* 36, 2429–2439. doi: 10.1016/s0031-3203(03)00044-x

Yang, L., and Xu, Z. (2019). Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. *Int. J. Mach. Learn. Cybern.* 10, 591–601. doi: 10.1007/s13042-017-0741-1

Zhang, J., and Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong. *Int J. Environ. Res. Public Health* 14:114. doi: 10.3390/ijerph14020114

Zheng, S., Jiang, M., Zhao, C., Zhu, R., Hu, Z., Xu, Y., et al. (2018). e-Bitter: bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* 6:82. doi: 10.3389/fchem.2018.00082

Zhou, L.-T., Cao, Y.-H., Lv, L.-L., Ma, K.-L., Chen, P.-S., Ni, H.-F., et al. (2017). Feature selection and classification of urinary mrna microarray data by iterative random forest to diagnose renal fibrosis: a two-stage study. *Sci. Rep.* 7:39832. doi: 10.1038/srep39832

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 29 September 2020
doi: 10.3389/fgene.2020.570530

Identification of Signatures of Prognosis Prediction for Melanoma Using a Hypoxia Score

Yanhong Shou[1]†, Lu Yang[1]†, Yongsheng Yang[1]*, Xiaohua Zhu[1], Feng Li[1] and Jinhua Xu[1,2]*

[1] Department of Dermatology, Huashan Hospital, Fudan University, Shanghai, China, [2] Institute of Dermatology, Shanghai, China

Melanoma is one of the most aggressive cancers. Hypoxic microenvironment affects multiple cellular pathways and contributes to tumor progression. The purpose of the research was to investigate the association between hypoxia and melanoma, and identify the prognostic value of hypoxia-related genes. Based on the GSVA algorithm, gene expression profile collected from The Cancer Genome Atlas (TCGA) was used for calculating the hypoxia score. The Kaplan–Meier plot suggested that a high hypoxia score was correlated with the inferior survival of melanoma patients. Using differential gene expression analysis and WGCNA, a total of 337 overlapping genes associated with hypoxia were determined. Protein-protein interaction network and functional enrichment analysis were conducted, and Lasso Cox regression was performed to establish the prognostic gene signature. Lasso regression showed that seven genes displayed the best features. A novel seven-gene signature (including ABCA12, PTK6, FERMT1, GSDMC, KRT2, CSTA, and SPRR2F) was constructed for prognosis prediction. The ROC curve inferred good performance in both the TCGA cohort and validation cohorts. Therefore, our study determined the prognostic implication of the hypoxia score in melanoma and showed a novel seven-gene signature to predict prognosis, which may provide insights into the prognosis evaluation and clinical decision making.

Keywords: melanoma, hypoxia score, prognosis, gene signature, prediction model

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Mohit Kumar Jolly,
Indian Institute of Science (IISc), India
Suman Ghosal,
National Institutes of Health (NIH), United States

****Correspondence:***
Yongsheng Yang
yangyongsheng73512@126.com
Jinhua Xu
jinhuaxu@fudan.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 June 2020*
Accepted: *08 September 2020*
Published: *29 September 2020*

Citation:
Shou Y, Yang L, Yang Y, Zhu X, Li F and Xu J (2020) Identification of Signatures of Prognosis Prediction for Melanoma Using a Hypoxia Score. Front. Genet. 11:570530. doi: 10.3389/fgene.2020.570530

INTRODUCTION

Melanoma is one of the highly malignant cutaneous neoplasms with a rising incidence around the world (Hallberg and Johansson, 2013; Domingues et al., 2018), characterized by its strong metastasis rate and poor prognosis (Nakamura and Fujisawa, 2018). Although surgery, chemotherapy, immunotherapy, and radiation have been performed to treat malignant melanoma, the efficacy of therapies remains limited (Domingues et al., 2018). Therefore, investigating the underlying biological mechanism and identifying new therapeutic targets are demanded.

Tumor microenvironment (TME) refers to the biological environment where tumors initiate, locate, and progress (Brandner and Haass, 2013; Roma-Rodrigues et al., 2019). The interaction between tumor and its TME influence the survival, migration, and invasion of tumor cells (Whiteside, 2008). Hypoxia is one of the essential features in the TME, which originates from the proliferation of tumor cells and increased oxygen consumption (Manoochehri Khoshinani et al., 2016). Tumor Hypoxia results in the activation of hypoxia-inducible factor (HIF), which mediates the expression of genes regulating metabolic pathways, pH regulation, DNA replication, and protein synthesis (Al Tameemi et al., 2019). Thus, tumor hypoxia contributes to heterogeneous changes, genetic instability, angiogenesis, and resistance to treatments, which has become an adverse prognostic factor of tumor assessment (Walsh et al., 2014; Jing et al., 2019). Many studies have suggested that hypoxia is related to poor prognosis in solid tumors (Winter et al., 2007; Ward et al., 2013). Likewise, hypoxia is a critical molecular program in melanoma, promoting tumor growth, invasion, treatment resistance, and relapse through the stabilization of HIF and the regulation of hypoxia-related responses (Widmer et al., 2013; Qin et al., 2016). In light of the essential role of hypoxia in melanoma, the detection and assessment of tumor hypoxia plays a critical role in clinical practice.

Assessment of the oxygen concentration, report of physiologic processes involving oxygen markers, and evaluation of endogenous molecules expression are considered as three major groups to detect tumor hypoxia status (Walsh et al., 2014). Deeply understanding the gene characteristics to estimate the degree of hypoxia would help the prognostic evaluation and treatment options. Immunohistochemistry (IHC) and plasma protein assays were developed for determining hypoxia (Russell et al., 2009; Khan et al., 2013). Recently, bioinformatics has been utilized to determine broader signatures. Based on the 26-gene hypoxia signature (Eustace et al., 2013), hypoxia status classifier was administrated in head and neck cancer (Brooks et al., 2019), and hypoxia score was implemented in lung adenocarcinoma (Liu Z. et al., 2020). Up till now, the hypoxia score in melanoma has not been investigated in detail.

Here, we calculated the hypoxia score for the analysis of gene expression profiles of melanoma which were collected from The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov). The correlation between hypoxia and prognosis was investigated, and hypoxia-associated molecules were determined. A seven-gene signature was further conducted using the profiles from TCGA and verified in the GSE54467, GSE53118, and GSE22153 dataset, providing novel insights for the assessment, treatment, and prognosis of melanoma. The workflow presenting the design of the present research was shown in **Figure 1**.

MATERIALS AND METHODS

Data Collection

The clinical information and RNA-sequencing data of skin cutaneous melanoma (SKCM) were downloaded from the TCGA database[1].

Calculation of the Hypoxia Score

Hypoxia score was calculated based on the 26-gene hypoxia signature (Eustace et al., 2013) and a gene set variation analysis (GSVA) (Eustace et al., 2013; Hänzelmann et al., 2013). GSVA is a GSE method which estimates variation of pathway activity over a sample population in an unsupervised manner (Hänzelmann et al., 2013). Hence, we used the 26-gene hypoxia signature and evaluated the GSVA score of each sample using the GSVA algorithm. The GSVA score was recognized as the hypoxia score, which represented the hypoxia status of each sample. The cut-off value was identified according to the method of best separation in R package survminer, and patients were divided into high- and low-hypoxia score groups. Such grouping aims to minimize the P value of the survival curve. Additionally, T-test was used to judge the differences of clinical indexes between groups.

Definition of Differentially Expressed Genes (DEGs)

EdgeR package was used to identify DEGs between high- and low-hypoxia score groups. The fold change ($|\text{fold change}| \geq 1.5$) and adj.p < 0.05 were considered significant. Pheatmap package was used to generate the heatmap.

Identification of Hypoxia-Associated Genes by the Weighted Gene Co-expression Network Analysis (WGCNA)

The top 9829 genes, based on standard deviation, were used for further investigation. Co-expression networks were performed by using the R package WGCNA (Langfelder and Horvath, 2008). Among all the soft threshold values, we chose the β that showed the highest mean connectivity (β = 3). As the module Eigengenes (ME) was recognized to define the interpretation of gene expression profile, we associated the ME with the hypoxia feature, which showed high and low hypoxia score. Module with the highest correlation was selected, and genes of which were named hypoxia-related genes.

The Protein-Protein Interaction (PPI) Network and Functional Annotation

The overlapping genes between DEGs and hypoxia-related genes were depicted by the online Venn diagram analysis[2]. We used the STRING (version 11.0, Search Tool for the Retrieval of Interacting Genes) and Cytoscape software (version 3.7.0) to construct the PPI network (Shannon et al., 2003; Szklarczyk et al., 2015). Molecular Complex Detection (MCODE) was utilized to determine the interaction clusters. The R package clusterprofile was used to perform functional enrichment analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis (Yu et al., 2012).

Survival Analysis and Construction of the Hypoxia-Related Signature for Melanoma

For survival analysis, we utilized Kaplan–Meier survival. Survival-related genes in the multivariate Cox regression analysis were inferred using the least absolutes shrinkage and selection operator (LASSO) by the R package glmnet. Risk scores

[1] https://tcga-data.nci.nih.gov/

[2] http://bioinformatics.psb.ugent.be/webtools/Venn/

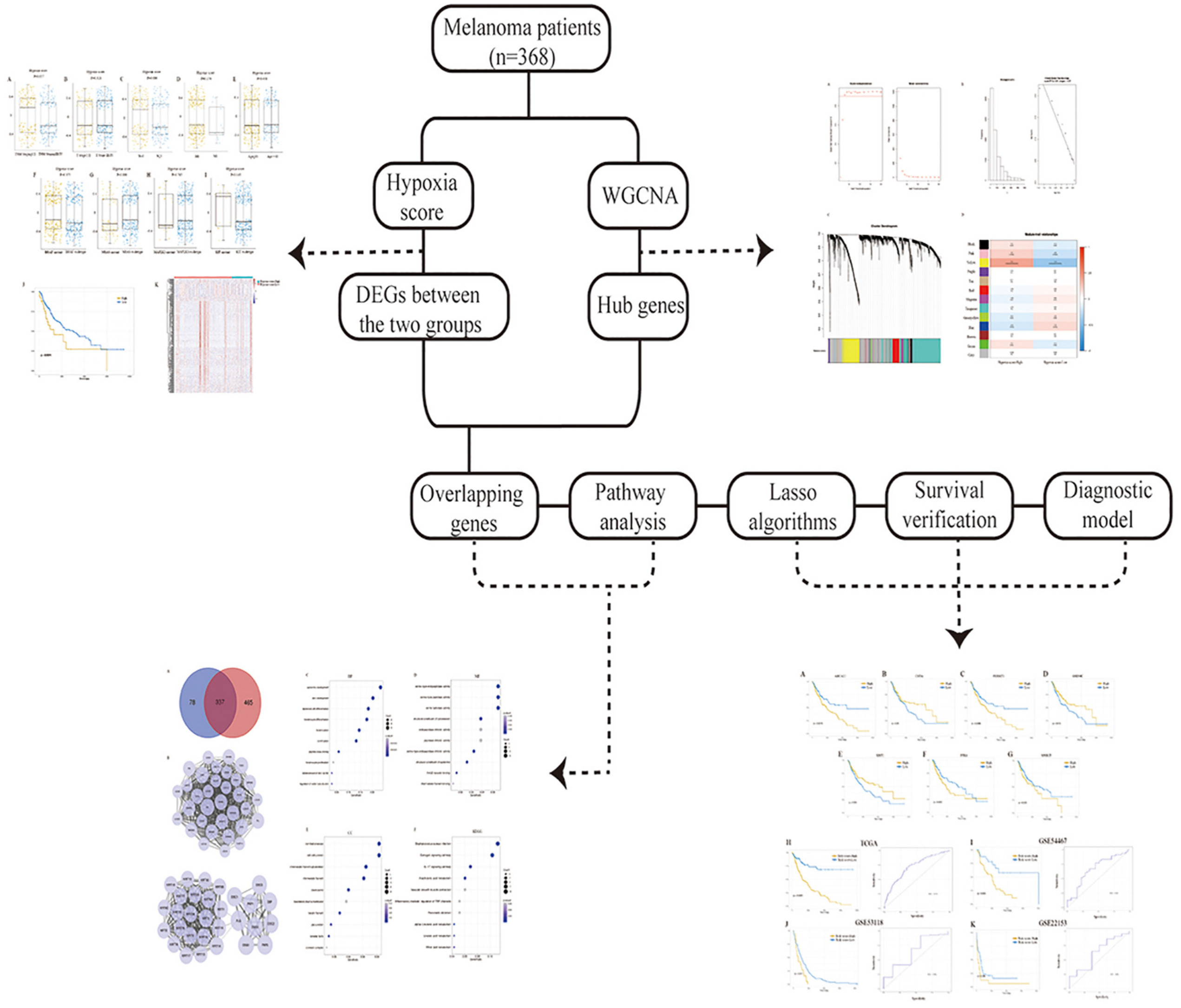


FIGURE 1 | The workflow of the research.

were obtained according to genes expression multiplied by a linear combination of regression coefficient acquired from the multivariate Cox regression, and patients were divided into a high-risk group and low-risk group based on the optimal cut-off point of risk score using the R package survminer. The Kaplan–Meier analysis and the receiver operating characteristic (ROC) curve were carried out using the R package ROCR.

Interaction Network Between the 7-Gene Signature and the 26-Gene Hypoxia Signature

To investigate the association between the 7-gene signature and the 26-gene list, genes from these two gene lists were input to the Gene-Cloud of Biotechnology Information (GCBI) analysis platform[3] for data analysis.

External Validation of the Hypoxia-Related Signature Model

The signature model was validated using the GSE54467, GSE53118, and GSE22153 dataset derived from the Gene Expression Omnibus (GEO) database[4]. Risk scores were calculated using the same formula, and Kaplan–Meier and ROC curve analyses were implemented.

RESULTS

Evaluation of the Degree of Hypoxia

Hypoxia scores were distributed between −0.699 to 0.659. A total of 368 patients were divided into high- and low-score groups based on the optional cut-off point of hypoxia

[3] https://www.gcbi.com.cn

[4] https://www.ncbi.nlm.nih.gov/geo

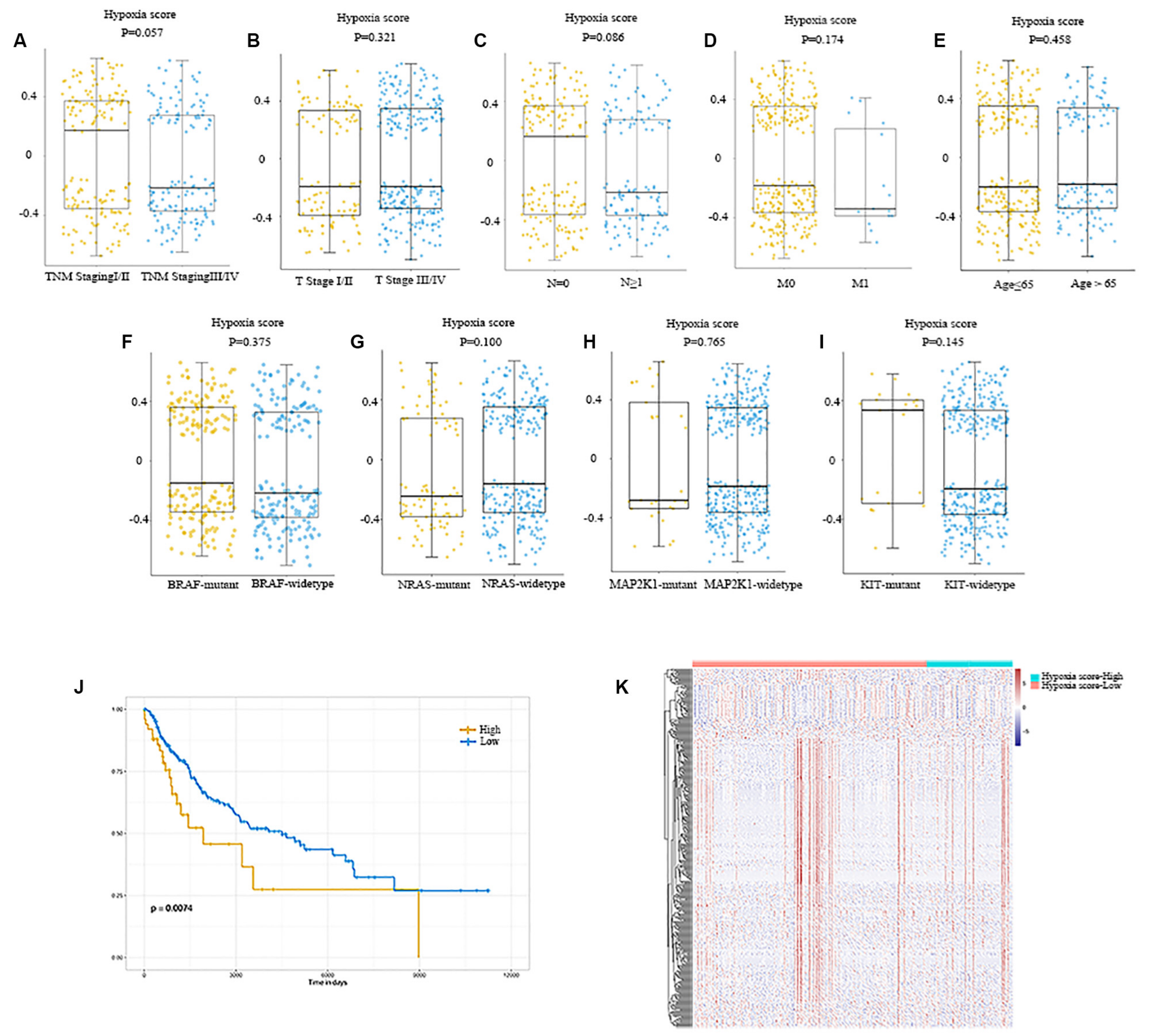


FIGURE 2 | Distribution of hypoxia score in melanoma. **(A)** Distribution of hypoxia score of patients with different TNM staging. **(B)** Distribution of hypoxia score of patients with different T stage. **(C)** Distribution of hypoxia score of patients with or without lymph node metastasis. **(D)** Distribution of hypoxia score of patients with or without distant metastasis. **(E)** Distribution of hypoxia score of patients younger than 65 and those older than 65 years of age. **(F–I)** Distribution of hypoxia score of patients with BRAF mutant and BRAF wildtype, patients with NRAS mutant and NRAS wildtype, patients with MAP2K1 mutant and MAP2K1 wildtype, and patients with KIT mutant and KIT wildtype, respectively. **(J)** Patients were divided into high- and low-hypoxia score groups based on the cut-off value. Patients with a high hypoxia score showed a better prognosis compared to patients with a low score (*(P)* = 0.007). **(K)** Heatmap of the DEGs of high-hypoxia score group vs. low-hypoxia score group. $p < 0.05$, |fold change| ≥ 1.5. DEGs, differentially expressed genes.

score (0.43, **Supplementary Figure S1**). As shown in **Figures 2A–E**, no obvious differences in hypoxia scores were detected in patients with different clinical features. Additionally, mutations were common in melanoma, including BRAF (50%), NRAS (30%), MAP2K1 (6%), and KIT (2.6%). So, we also plotted the distribution of hypoxia scores to the status of driver mutations and found they were not significant (P = 0.375, P = 0.100, P = 0.765, P = 0.145, **Figures 2F–I**).

The effects of hypoxia on prognosis were analyzed. The Kaplan–Meier plot suggested that patients with high hypoxia scores had a poor prognosis (P = 0.007, **Figure 2J**). To further determine the correlation of gene expression with hypoxia scores, we did differential gene expression analysis between high and low hypoxia scores. Of the 415 differential expression genes (DEGs), 365 genes were upregulated, while 50 genes were downregulated. Heatmaps in **Figure 2K** inferred distinct gene expression profiles of cases belong to high- vs. low-hypoxia scores.

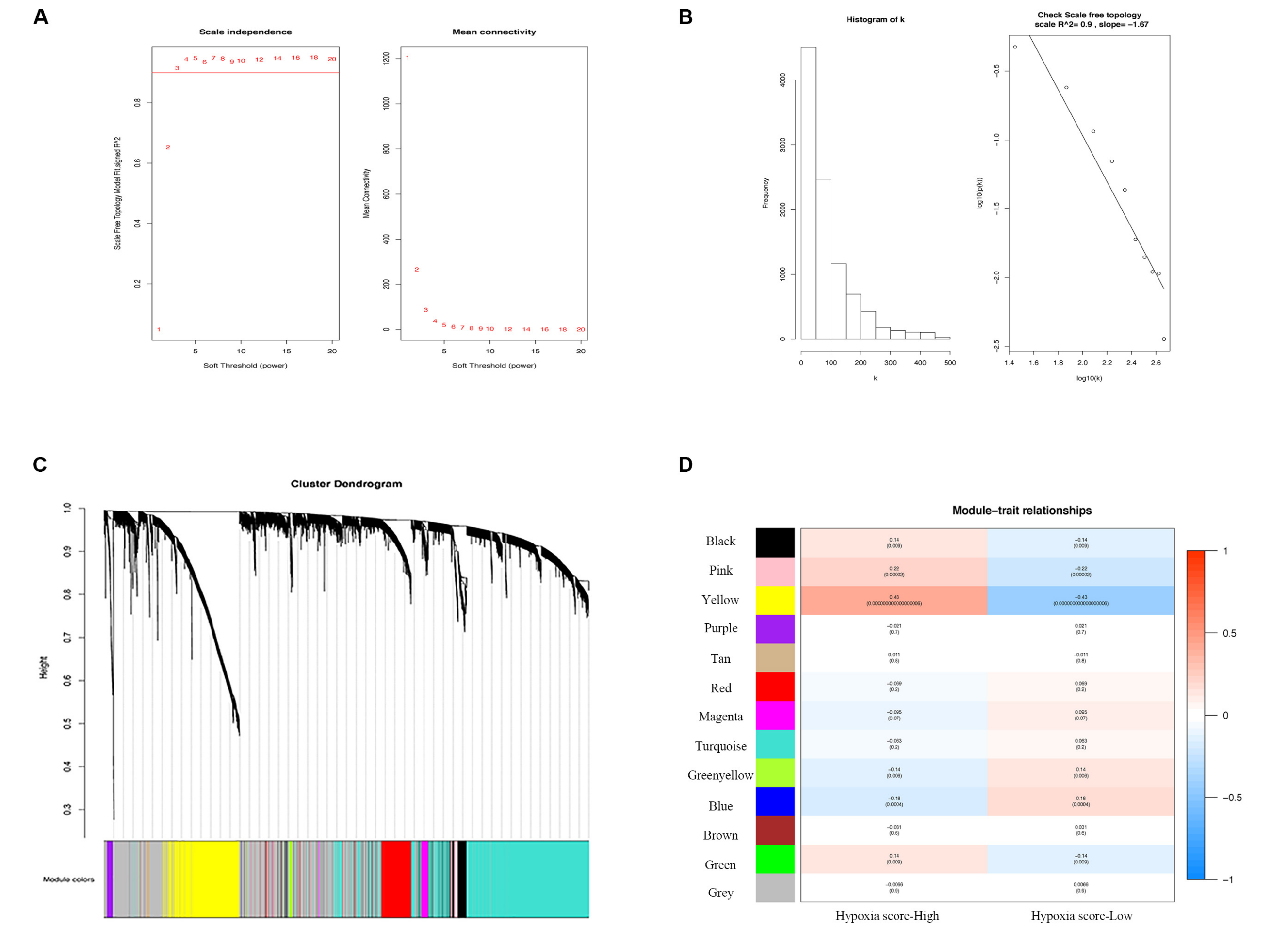


FIGURE 3 | Determination of modules correlated with the hypoxia of melanoma in the WGCNA. **(A)** Analysis of the scale-free fit index and the mean connectivity for various soft-thresholding powers. **(B)** Checking the scale free topology when β = 3. Correlation coefficient = 0.9, which showed scale-free topology. **(C)** Dendrogram of genes clustered according to a dissimilarity measure (1-TOM). **(D)** Heatmap of the correlation between module Eigengenes and hypoxia. WGCNA, the weighted gene co-expression network analysis.

Identification of the Most Relevant Module Genes for Hypoxia in Melanoma

We selected the top 9829 (of 19658) after sorting by the standard deviation (**Figures 3A–C**). The co-expression network was constructed, and 13 modules were determined. Correlation analysis between the module eigengenes and hypoxia scores showed that the yellow module (**Figure 3D**, Module–trait relationships = 0.43, *P* = 0.000) had the highest association with the degree of hypoxia. Then, 802 genes in the module were considered to be hub hypoxia-related genes for further investigation.

Protein-Protein Interactions and Functional Enrichment Analysis

A total of 337 genes were overlapped between DEGs and hypoxia-related genes (**Figure 4A**). To explore the interplay among 337 overlapping genes, the STRING tool with confidence > 0.7 was used to construct a PPI network. There were 10 modules in the network, including 195 nodes and 1173 edges. Modules with 10 or more nodes were selected for further analysis (**Figure 4B**). Based on the connection degree, we named these modules IVL, and FLG modules, respectively. In the IVL module, 528 edges involving 33 nodes were formed in the network. IVL, TGM1, LOR, SPRR1B, and PPL were the remarkable nodes, as they had the most connections with others. In the FLG module, FLG, DSG1, DSG3, PKP3, PKP1, KRT14, and DSC1 occupied the center of the module.

To better understand the biological significance, we conducted enrichment analysis of the 337 overlapping genes. As shown in **Figure 4**, a total of 27 terms of biological process (BP), 8 terms of cellular component (CC), and 14 terms of molecular function (MF) were enriched (*P* < 0.05). Top GO terms comprised epidermis development, skin development and epidermal cell differentiation (**Figure 4C**), serine type endopeptidase activity, serine type peptidase activity, and serine hydrolase activity (**Figure 4D**), and cornified envelope and cell-cell junction (**Figure 4E**). Besides, KEGG analysis

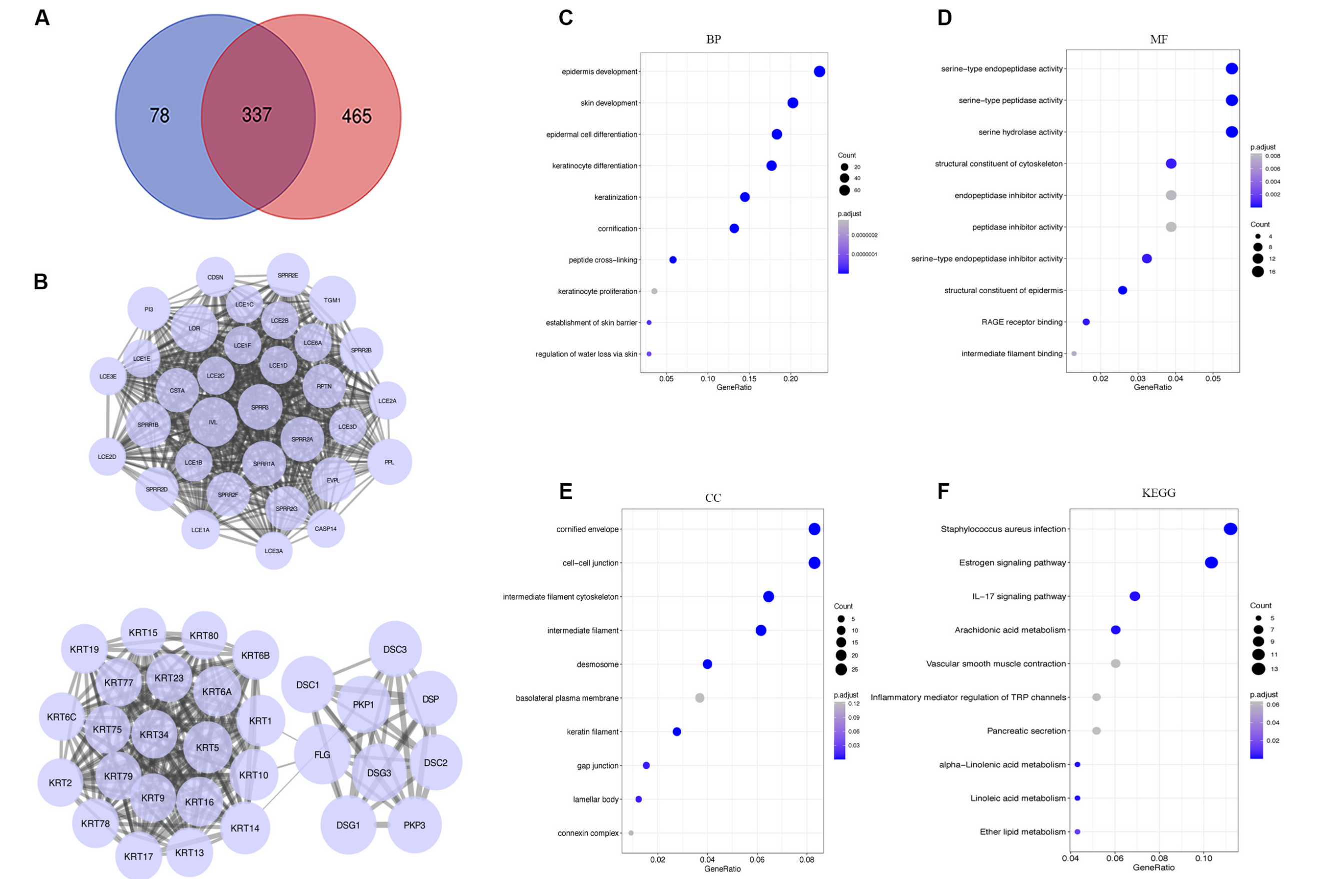


FIGURE 4 | Analysis of DEGs. **(A)** Venn diagrams showing the number of commonly genes in DEGs and yellow module. **(B)** PPI networks of overlapping genes. A large node represented a higher degree. **(C–E)** Go enrichment analysis of biological process (BP), molecular function (MF), and cellular component (CC). **(F)** KEGG pathway enrichment analysis of the overlapping genes. DEGs, differentially expressed genes; PPI, the protein-protein interaction; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

suggested overlapping genes were enriched in *Staphylococcus aureus* infection, estrogen signaling pathway, and IL-17 signaling pathway (**Figure 4F**).

Determination of Prognostic Molecules and a Prognostic Risk Model

We generated Kaplan–Meier survival curves to explore the independent prognostic impact of 337 overlapping genes and found that 29 genes were associated with prognosis in the log-rank test ($P < 0.05$). A total of 7 genes identified with the LASSO algorithms included ATP-binding cassette subfamily a member 12 (ABCA12), protein tyrosine kinase 6 (PTK6), fermitin family member 1 (FERMT1), gasdermin C (GSDMC), keratin 2 (KRT2), cystatin A (CSTA), and small proline rich protein 2F (SPRR2F), and constructed as a seven-gene signature model (**Table 1**). The risk score = 0.26084 ∗ Expression (ABCA12) + 0.05797 ∗ Expression (PTK6) + 0.14404 ∗ Expression (FERMT1) + (−0.44473) ∗ Expression (GSDMC) + (−0.09102) ∗ Expression (KRT2) + (−0.02677) ∗ Expression (CSTA) + 0.11245 ∗ Expression (SPRR2F). The roles of these 7 genes in melanoma and hypoxia responses were described in **Table 2**. Also, we explored the relationships among genes from the 7-gene signature and 26-gene list. Although genes from the 7-gene signature were different from those of the 26-gene one, there were common regulators associated with hypoxic responses, including EGFR, ERBB2, and miR-125a (**Figures 5A,B**, **Table 3**).

Kaplan–Meier curve and ROC were utilized to assess the prognostic capacity of the seven-gene signature model, and similar procedures were performed in the external data. The

TABLE 1 | The results of Univariate Cox regression analysis.

	HR	Z	P
ABCA12	0.564	−3.171	0.002
PTK6	0.617	−2.687	0.007
FERMT1	0.626	−2.607	0.009
GSDMC	1.547	2.419	0.02
KRT2	1.460	2.105	0.04
CSTA	1.442	2.042	0.04
SPRR2F	0.682	−1.955	0.04

HR: hazard ratio.

TABLE 2 | The roles of 7 genes in melanoma and hypoxia response.

Gene	Function	Role in melanoma	Role in hypoxia response
ABCA12	Membrane transport	Associated with skin malignancies including melanoma	Not reported
PTK6	Protein phosphorylation	Identified as a prognostic biomarker for metastatic skin cancers including malignant melanoma	Up-regulated by HIF-1α and HIF-2α
FERMT1	Keratinocyte proliferation	Not reported	Down-regulated in the condition of hypoxia
GSDMC	Pyroptosis	Present in malignant melanoma and associated with the metastasis	Not reported
KRT2	Keratinization	Not reported	Not reported
CSTA	Keratinocyte differentiation	Not reported	Up-regulated in hypoxic cells
SPRR2F	Epidermis development	Not reported	Not reported

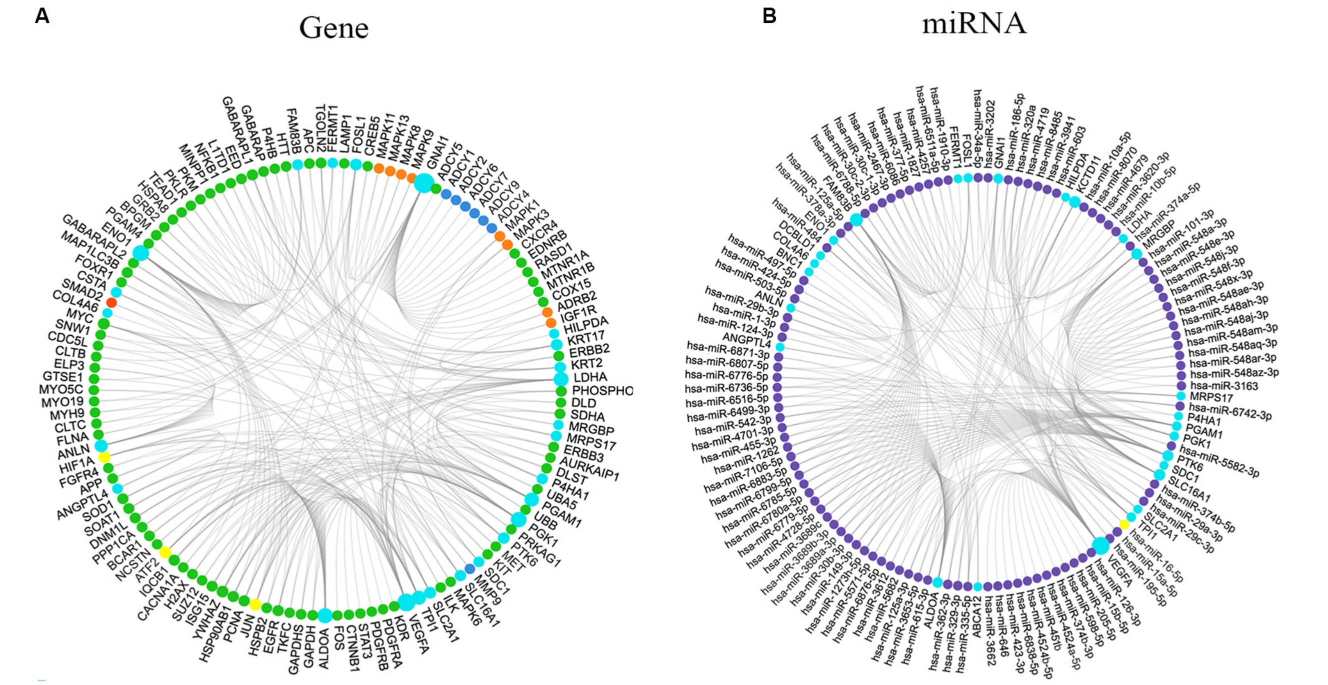


FIGURE 5 | Interaction network of molecules associated with the genes from the 7-gene signature and the 26-gene list. **(A)** Interaction network of genes associated with the 7-gene signature and the 26-gene list. Colors indicated types of genes: light blue, input genes; orange, activated genes; red, expressed genes; green, associated genes; dark blue, inhibited genes; yellow, the largest connection counts. Node size was adjusted according to the number of associated genes. **(B)** Interaction network of miRNAs correlated with the 7-gene signature and the 26-gene list. Colors indicated types of molecules: light blue, input genes; purple, targeted miRNAs; yellow, the largest connection counts. Node size was adjusted according to the number of associated miRNAs.

TABLE 3 | Common regulators and downstream effectors.

	Targeted-genes in the 7-gene signature	Targeted-genes in the 26-gene list
EGFR	PTK6	ALDOA, TPI1
ERBB2	PTK6	KRT17
MiR-125a	FERMT1	VEGFA, ENO1, TPI1

results showed that genes in the signature model performed well-predicting prognosis within the TCGA cohort (**Figures 6A–G**). **Figures 6H–K** suggested that patients with low-risk scores had significantly longer overall survival than those with high-risk scores in TCGA, GSE54467, GSE53118, and GSE22153 dataset ($P < 0.001$, $P = 0.004$, $P = 0.017$, $P = 0.048$). The AUCs were 0.716 (95% CI: 0.661–0.771), 0.667 (95% CI: 0.541–0.792), 0.648 (95% CI: 0.419–0.878), and 0.628 (95% CI: 0.406–0.849), respectively (**Figures 6H–K**).

DISCUSSION

Hypoxia, one of the hallmarks of TME, is a biological condition present in most tumors (Jing et al., 2019). Tumor

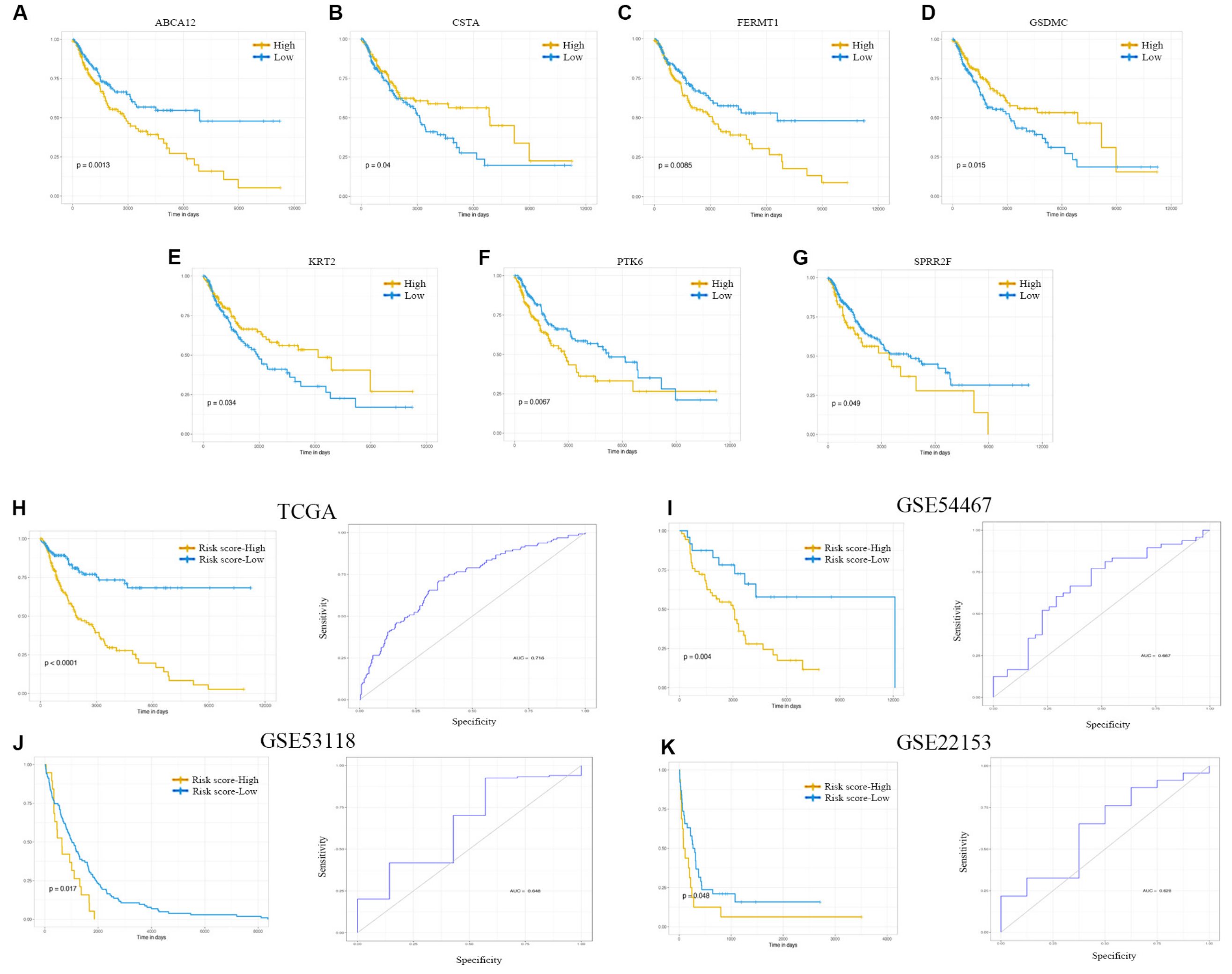


FIGURE 6 | Kaplan–Meier analysis, risk score analysis, and ROC analysis for the seven-gene signature. **(A–G)** Kaplan–Meier curves for overall survival of ABCA12, CSTA, FERMT1, GSDMC, KRT2, PTK6, and SPRR2F. **(H–K)** Kaplan–Meier curves for overall survival of risk score and ROC analysis for the seven-gene signature in TCGA cohort, GSE54467, GSE53118, and GSE22153. ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas.

hypoxia exacerbates progression and metastasis through both physiological and genomic mechanisms (Akanji et al., 2019). Investigating crucial features of tumor hypoxia environment may facilitate clinical decision-making.

Previous studies identified several genes, long non-coding RNAs and miRNAs as promising therapeutic biomarkers in melanoma (Zhang et al., 2017; Wei et al., 2019; Xu et al., 2019). However, the differentially expressed signatures were explored between the normal and tumor samples, or between the primary and metastatic tissues, and molecules associated with the progression of cancer were not taken into consideration. Notably, the focus of our study was to estimate the degree of hypoxia according to the evidential basis for 26-gene hypoxia signature (Eustace et al., 2013), and high hypoxia score was demonstrated as a strong predictor of poor clinical outcome. Subsequently, we identified the promising hypoxia-related genes associated with prognosis.

Based on bioinformatics methods and databases, hypoxia score was calculated, and patients were divided into high- and low-score groups. DEGs were collected using differential gene expression analysis. At the same time, WGCNA analysis was performed to select the modules with the strongest relationship between genes in the modules and the module traits. The overlapping 337 genes of the above two clusters were determined as the hypoxia strongly associated genes related to melanoma. Functional analysis showed these 337 genes to be closely related to the development of melanoma, like via cell-cell junction. Cell junction was reported to be relevant for the metastatic process (Knights et al., 2012). Also, the process of epidermis development and epidermal cell differentiation were enriched. Previous studies showed the hyperplastic epidermal region was accompanied by aberrant expression of keratin 14, and melanoma cells were able to increase expression of keratins 8, 19 (Kodet et al., 2015). keratin 8, 14,19 were also observed in the FLG module from PPI network. These results showed that epidermis surrounding melanoma performed hyperplastic features, and indicated the possible interaction between melanoma cells and keratinocytes. KEGG analysis highlighted the estrogen signaling pathway and the IL-17 signaling pathway. Several studies pointed out that the estrogen signaling pathway relied on the balance between estrogen receptor (ER) α and ERβ expression, and the levels of ERβ regulated the capacity of melanoma invasion (Marzagalli et al., 2016; Rajabi et al., 2017). Additionally, IL-17/IL-17RA pathway stimulated cell proliferation of mouse B16F10 and human A375 and A2058 cell lines (Chen et al., 2019). IL-17 and IL-23 immunohistochemistry expression were increased in the melanoma tissues, possibly enhancing VEGF expression and angiogenesis (Ganzetti et al., 2015). Therefore, these analyses supported the hypothesis of the importance of hypoxia microenvironment in the regulation of the biological behavior of tumor cells and surrounding non-tumor cells.

Based on the log-rank test identifying the genes associated with prognosis, LASSO was performed, and seven characteristic variables were extracted. ABCA12 was upregulated in ovarian carcinoma and colorectal cancer, which was recognized as a promising candidate marker (Hlavata et al., 2012; Elsnerova et al., 2016). Mutations in ABCA12 were related to malignant melanoma (Natsuga et al., 2007). PTK6, a non-receptor type tyrosine kinase, was involved in breast, pancreatic cancer and metastatic skin cancer. It was recognized that PTK6 regulated proliferation and migration (Gotoh et al., 2014; Ito et al., 2017; Liu G. et al., 2020). FERMT1, encoding Kindlin-1, was correlated with metastasis and poor prognosis in several solid tumors (Liu et al., 2016; Sarvi et al., 2018). GSDMC functioned as an oncogene, enhancing cell proliferation and tumorigenesis in lung adenocarcinoma and colorectal carcinogenesis (Miguchi et al., 2016; Wei et al., 2020). It presented high in malignant melanoma but undetectable in normal epithelial cells, which might be associated with the metastasis of cells (Xia et al., 2019). CSTA, one of the tumor suppressors, had the anti-apoptotic effect and maintaining cell-cell adhesion. It was upregulated in several epithelial-derived malignancies, including squamous cell carcinoma (Gupta et al., 2015; Ma et al., 2018). KRT2 was found to form a mechanically resilient cytoskeleton and contribute to the skin homeostasis (Fischer et al., 2016). SPRR2F, a cross-linked envelope protein of keratinocytes, providing the protective barrier function (Cabral et al., 2001). Although there was no report of KRT2 and SPRR2F as a prognostic molecule of tumors, KRT2 and SPRR2F might function as promising biomarkers in melanoma. The consistency of our findings regarding ABCA12, PTK6, FERMT1, GSDMC and CSTA with previous studies suggested our method to be reliable, and thus supported the reliability of these potential prognostic and therapeutic targets to a certain extent.

Previous studies inferred that the expression of PTK6 were up-regulated, and FERMT1 were down-regulated in response to the hypoxia condition (Hlavata et al., 2012; Regan Anderson et al., 2013; Lin and Liu, 2019). PTK6 expression depended on both HIF-1α and HIF-2α, which were reported to have a direct regulation of PTK6 transcription. In the analytic process of investigating the effect of hypoxia on the vhl-deficient cells, HIF-regulated genes were obtained. FERMT1 was one of the 214 downregulated DEGs. Additionally, the increased expression of CSTA was detected in hypoxic A431 cells (Park et al., 2010). Although there was no common gene between the 7- and 26-gene signatures, a total of 3 genes, including epidermal growth factor receptor (EGFR), erb-b2 receptor tyrosine kinase 2 (ERBB2), and miR-125a, were identified as common regulators and effectors in these two gene lists in a context-dependent manner. PTK6 was reported to enhance EGFR signaling by direct phosphorylation of EGFR and inhibition of its degradation (Li et al., 2012), and EGFR might promote the cellular response to hypoxia by increasing HIF-1α expression (Swinson and O'Byrne, 2006). Through the split ubiquitin (Ub)-based membrane yeast two-hybrid assay, EGFR was reported to be physically associated with aldolase (ALDOA) and triosephosphate isomerase 1 (TPI1), respectively (Deribe et al., 2009). However, the potential functions of ALDOA and TPI1 need to be further explored. Furthermore, ERBB2, also known as HER2, was recognized as a regulator of HIF-2α and a driver of hypoxic responses (Jarman et al., 2019). PTK6 was coamplified with ERBB2 to promote cell proliferation (Xiang et al., 2008). Additionally, ERBB2 and keratin 17 (KRT17) were found to locate in the same chromosome region, which might have the following tumor associations (Zhang et al., 2013). Apart from regulating the expression of genes, hypoxia-regulated microRNAs (miRNAs) were identified. MiR-125a was a direct target of HIF-1α and drove the reduction of vascular endothelial growth factor A (VEGFA) (Dai et al., 2015; Pan et al., 2018). Based on the map of human miRNA interactome, enolase 1 (ENO1), FERMT1, and TPI1 were observed in the interaction sites of miR-125a and further examinations were demanded (Helwak et al., 2013).

Saxena and Jolly summarized different extents of hypoxia (Saxena and Jolly, 2019). Under acute hypoxia, HIF-1α levels stayed high to regulate acute response, while HIF-2α levels were stabilized later and played a crucial role during chronic hypoxia. Besides, cyclic hypoxia enhanced the expression of HIF-1α instead of HIF-2α. Several factors implicated in these hypoxia conditions were determined, including HSP-70, HAF, H3, H4, REST, and miR-429. Although genes identified in our study have been reported to function in hypoxic responses, there was no report of them to make a distinction of conditions of hypoxia, and further experimental verification is required.

Considering the accuracy of these prognostic genes, a seven-signature model was established based on the combination of genes. Cases in the low-risk group inferred obviously better survival than patients in the high-risk group. The prognosis predictive performance of the model was relatively good not only in the TCGA melanoma cohort but also in the GSE54467, GSE53118, and GSE22153 cohort. Additionally, we investigated whether the clinical features were correlated with the degree of hypoxia, and the results showed that no apparent differences in hypoxia score were observed. BRAF mutation was found to increase HIF-1α expression and influenced survival in previous studies (Kumar et al., 2007; Zerilli et al., 2010). KIT mutant was reported to require HIF-1α to transform melanocytes into melanoma cells (Monsel et al., 2010). In our cohort, the hypoxia score in the BRAF-mutant or KIT-mutant group was slightly higher than that of the wildtype group, but it was not statistically significant. It could be because of an inevitable limitation, the sample size. There were two other limitations to our study. Firstly, data were collected from TCGA, where the potential for selection bias could not be excluded, but we validated the results in the GEO database and demonstrated the reliability to some extent. Secondly, analysis in our study was descriptive, further research *in vitro* and *in vivo* could enhance our understanding of the critical genes.

In conclusion, we applied the hypoxia score to determine the degree of hypoxia in TME and identified the prognostic role of hypoxia score. Furthermore, using bioinformatics and machine learning methods, we determined the seven-gene prognostic signature as a potential prognostic predictor and therapeutic targets for melanoma.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/, GSE53118; https://www.ncbi.nlm.nih.gov/geo/, GSE54467; https://www.ncbi.nlm.nih.gov/geo/, GSE22153.

AUTHOR CONTRIBUTIONS

XZ, FL, YY, and JX contributed to the design of this study. YS, LY, and YY contributed to the analysis of this study. YS contributed to drafting the text and preparing the tables and figures. All authors participated in the data collection, critical review, revision of this manuscript, contributed to the article, and approved the submitted version.

FUNDING

This research is supported by a grant from The National Natural Science Foundation of China (81673060).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.570530/full#supplementary-material

Supplementary Figure 1 | The process of finding optimal cut-off value to divide the patients into high- and low-hypoxia score groups.

REFERENCES

Akanji, M. A., Rotimi, D., and Adeyemi, O. S. (2019). Hypoxia-inducible factors as an alternative source of treatment strategy for cancer. *Oxid. Med. Cell Longev.* 2019:8547846. doi: 10.1155/2019/8547846

Al Tameemi, W., Dale, T. P., Al-Jumaily, R. M. K., and Forsyth, N. R. (2019). Hypoxia-modified cancer cell metabolism. *Front. Cell Dev. Biol.* 7:4. doi: 10.3389/fcell.2019.00004

Brandner, J. M., and Haass, N. K. (2013). Melanoma's connections to the tumour microenvironment. *Pathology* 45, 443–452. doi: 10.1097/PAT.0b013e328363b3bd

Brooks, J. M., Menezes, A. N., Ibrahim, M., Archer, L., Lal, N., Bagnall, C. J., et al. (2019). Development and Validation of a Combined Hypoxia and Immune Prognostic Classifier for Head and Neck Cancer. *Clin. Cancer Res.* 25, 5315–5328. doi: 10.1158/1078-0432.Ccr-18-3314

Cabral, A., Voskamp, P., Cleton-Jansen, A.-M., South, A., Nizetic, D., and Backendorf, C. (2001). Structural Organization and Regulation of the Small Proline-rich Family of Cornified Envelope Precursors Suggest a Role in Adaptive Barrier Function. *J. Biol. Chem.* 276, 19231–19237. doi: 10.1074/jbc.M100336200

Chen, Y.-S., Huang, T.-H., Liu, C.-L., Chen, H.-S., Lee, M.-H., Chen, H.-W., et al. (2019). Locally Targeting the IL-17/IL-17RA Axis Reduced Tumor Growth in a Murine B16F10 Melanoma Model. *Hum. Gene Ther.* 30, 273–285. doi: 10.1089/hum.2018.104

Dai, J., Wang, J., Yang, L., Xiao, Y., and Ruan, Q. (2015). miR-125a regulates angiogenesis of gastric cancer by targeting vascular endothelial growth factor A. *Int. J. Oncol.* 47, 1801–1810. doi: 10.3892/ijo.2015.3171

Deribe, Y. L., Wild, P., Chandrashaker, A., Curak, J., Schmidt, M. H. H., Kalaidzidis, Y., et al. (2009). Regulation of epidermal growth factor receptor trafficking by lysine deacetylase HDAC6. *Sci. Signal* 2:ra84. doi: 10.1126/scisignal.2000576

Domingues, B., Lopes, J. M., Soares, P., and Populo, H. (2018). Melanoma treatment in review. *Immunotargets Ther.* 7, 35–49. doi: 10.2147/ITT.S134842

Elsnerova, K., Mohelnikova-Duchonova, B., Cerovska, E., Ehrlichova, M., Gut, I., Rob, L., et al. (2016). Gene expression of membrane transporters: importance for prognosis and progression of ovarian carcinoma. *Oncol Rep.* 35, 2159–2170. doi: 10.3892/or.2016.4599

Eustace, A., Mani, N., Span, P. N., Irlam, J. J., Taylor, J., Betts, G. N. J., et al. (2013). A 26-Gene Hypoxia Signature Predicts Benefit from Hypoxia-Modifying Therapy in Laryngeal Cancer but Not Bladder Cancer. *Clin. Cancer Res.* 19, 4879–4888. doi: 10.1158/1078-0432.Ccr-13-0542

Fischer, H., Langbein, L., Reichelt, J., Buchberger, M., Tschachler, E., and Eckhart, L. (2016). Keratins K2 and K10 are essential for the epidermal integrity of plantar skin. *J. Dermatol. Sci.* 81, 10–16. doi: 10.1016/j.jdermsci.2015.10.008

Ganzetti, G., Rubini, C., Campanati, A., Zizzi, A., Molinelli, E., Rosa, L., et al. (2015). IL-17, IL-23, and p73 expression in cutaneous melanoma: a pilot study. *Melanoma Res.* 25, 232–238. doi: 10.1097/CMR.0000000000000151

Gotoh, N., Ono, H., Basson, M. D., and Ito, H. (2014). PTK6 promotes cancer migration and invasion in pancreatic cancer cells dependent on ERK signaling. *PLoS One* 9:e96060. doi: 10.1371/journal.pone.0096060

Gupta, A., Nitoiu, D., Brennan-Crispi, D., Addya, S., Riobo, N. A., Kelsell, D. P., et al. (2015). Cell cycle- and cancer-associated gene networks activated by Dsg2: evidence of cystatin A deregulation and a potential role in cell-cell adhesion. *PLoS One* 10:e0120091. doi: 10.1371/journal.pone.0120091

Hallberg, Ö, and Johansson, O. (2013). Increasing Melanoma—Too Many Skin Cell Damages or Too Few Repairs? *Cancers* 5, 184–204. doi: 10.3390/cancers5010184

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-17

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the Human miRNA Interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi: 10.1016/j.cell.2013.03.043

Hlavata, I., Mohelnikova-Duchonova, B., Vaclavikova, R., Liska, V., Pitule, P., Novak, P., et al. (2012). The role of ABC transporters in progression and clinical outcome of colorectal cancer. *Mutagenesis* 27, 187–196. doi: 10.1093/mutage/ger075

Ito, K., Park, S. H., Katsyv, I., Zhang, W., De Angelis, C., Schiff, R., et al. (2017). PTK6 regulates growth and survival of endocrine therapy-resistant ER+ breast cancer cells. *npj Breast Cancer* 3, 1–7. doi: 10.1038/s41523-017-0047-41

Jarman, E. J., Ward, C., Turnbull, A. K., Martinez-Perez, C., Meehan, J., Xintaropoulou, C., et al. (2019). HER2 regulates HIF-2α and drives an increased hypoxic response in breast cancer. *Breast Cancer Res.* 21:10. doi: 10.1186/s13058-019-1097-1090

Jing, X., Yang, F., Shao, C., Wei, K., Xie, M., Shen, H., et al. (2019). Role of hypoxia in cancer therapy by regulating the tumor microenvironment. *Mol. Cancer* 18:157. doi: 10.1186/s12943-019-1089-1089

Khan, R. H., Ahmad, Y., Sharma, N. K., Garg, I., Ahmad, M. F., Sharma, M., et al. (2013). An Insight into the Changes in Human Plasma Proteome on Adaptation to Hypobaric Hypoxia. *PLoS One* 8:e67548. doi: 10.1371/journal.pone.0067548

Knights, A. J., Funnell, A. P., Crossley, M., and Pearson, R. C. (2012). Holding tight: cell junctions and cancer spread. *Trends Cancer Res* 8, 61–69.

Kodet, O., Lacina, L., Krejčí, E., Dvořánková, B., Grim, M., Štork, J., et al. (2015). Melanoma cells influence the differentiation pattern of human epidermal keratinocytes. *Mol. Cancer* 14:1. doi: 10.1186/1476-4598-14-11

Kumar, S. M., Yu, H., Edwards, R., Chen, L., Kazianis, S., Brafford, P., et al. (2007). Mutant V600E BRAF Increases Hypoxia Inducible Factor-1α Expression in Melanoma. *Cancer Res.* 67, 3177–3184. doi: 10.1158/0008-5472.Can-06-3312

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, X., Lu, Y., Liang, K., Hsu, J. M., Albarracin, C., Mills, G. B., et al. (2012). Brk/PTK6 sustains activated EGFR signaling through inhibiting EGFR degradation and transactivating EGFR. *Oncogene* 31, 4372–4383. doi: 10.1038/onc.2011.608

Lin, H., and Liu, H. (2019). Hypoxia on vhl-deficient cells to obtain hif related genes through bioinformatics analysis. *bioRxiv [Preprint]* doi: 10.1101/863662

Liu, C. C., Cai, D. L., Sun, F., Wu, Z. H., Yue, B., Zhao, S. L., et al. (2016). FERMT1 mediates epithelial–mesenchymal transition to promote colon cancer metastasis via modulation of β-catenin transcriptional activity. *Oncogene* 36, 1779–1792. doi: 10.1038/onc.2016.339

Liu, G., Li, C., Zhen, H., Zhang, Z., and Sha, Y. (2020). Identification of prognostic gene biomarkers for metastatic skin cancer using data mining. *Biomed. Rep.* 13, 22–30. doi: 10.3892/br.2020.1307

Liu, Z., Tao, H., Tong, L., and Li, H. (2020). Genome-wide analysis of the hypoxia-related DNA methylation-driven genes in lung adenocarcinoma progression. *Biosci. Rep.* 40:BSR20194200. doi: 10.1042/bsr20194200

Ma, Y., Chen, Y., Li, Y., Grün, K., Berndt, A., Zhou, Z., et al. (2018). Cystatin A suppresses tumor cell growth through inhibiting epithelial to mesenchymal transition in human lung cancer. *Oncotarget* 9, 14084–14098. doi: 10.18632/oncotarget.23505

Manoochehri Khoshinani, H., Afshar, S., and Najafi, R. (2016). Hypoxia: a double-edged sword in cancer therapy. *Cancer Invest.* 34, 536–545. doi: 10.1080/07357907.2016.1245317

Marzagalli, M., Montagnani Marelli, M., Casati, L., Fontana, F., Moretti, R. M., and Limonta, P. (2016). Estrogen receptor beta in melanoma: from molecular insights to potential clinical utility. *Front. Endocrinol.* 7:140. doi: 10.3389/fendo.2016.00140

Miguchi, M., Hinoi, T., Shimomura, M., Adachi, T., Saito, Y., Niitsu, H., et al. (2016). Gasdermin C Is Upregulated by Inactivation of Transforming Growth Factor beta Receptor Type II in the Presence of Mutated Apc, Promoting Colorectal Cancer Proliferation. *PLoS One* 11:e0166422. doi: 10.1371/journal.pone.0166422

Monsel, G., Ortonne, N., Bagot, M., Bensussan, A., and Dumaz, N. (2010). c-Kit mutants require hypoxia-inducible factor 1alpha to transform melanocytes. *Oncogene* 29, 227–236. doi: 10.1038/onc.2009.320

Nakamura, Y., and Fujisawa, Y. (2018). Diagnosis and management of acral lentiginous melanoma. *Curr. Treatment Options Oncol.* 19:42. doi: 10.1007/s11864-018-0560-y

Natsuga, K., Akiyama, M., Kato, N., Sakai, K., Sugiyama-Nakagiri, Y., Nishimura, M., et al. (2007). Novel ABCA12 Mutations Identified in Two Cases of Non-Bullous Congenital Ichthyosiform Erythroderma Associated with Multiple Skin Malignant Neoplasia. *J. Invest. Dermatol.* 127, 2669–2673. doi: 10.1038/sj.jid.5700885

Pan, L., Zhou, L., Yin, W., Bai, J., and Liu, R. (2018). miR-125a induces apoptosis, metabolism disorder and migrationimpairment in pancreatic cancer cells by targeting Mfn2-related mitochondrial fission. *Int. J. Oncol.* 53, 124–136. doi: 10.3892/ijo.2018.4380

Park, J. E., Tan, H. S., Datta, A., Lai, R. C., Zhang, H., Meng, W., et al. (2010). Hypoxic Tumor Cell Modulates Its Microenvironment to Enhance Angiogenic and Metastatic Potential by Secretion of Proteins and Exosomes. *Mol. Cell. Proteom.* 9, 1085–1099. doi: 10.1074/mcp.M900381-MCP200

Qin, Y., Roszik, J., Chattopadhyay, C., Hashimoto, Y., Liu, C., Cooper, Z. A., et al. (2016). Hypoxia-Driven Mechanism of Vemurafenib Resistance in Melanoma. *Mol. Cancer Ther.* 15, 2442–2454. doi: 10.1158/1535-7163.Mct-15-0963

Rajabi, P., Bagheri, M., and Hani, M. (2017). Expression of Estrogen Receptor Alpha in Malignant Melanoma. *Adv. Biomed. Res.* 6:14. doi: 10.4103/2277-9175.200789

Regan Anderson, T. M., Peacock, D. L., Daniel, A. R., Hubbard, G. K., Lofgren, K. A., Girard, B. J., et al. (2013). Breast tumor kinase (Brk/PTK6) is a mediator of hypoxia-associated breast cancer progression. *Cancer Res.* 73, 5810–5820. doi: 10.1158/0008-5472.CAN-13-0523

Roma-Rodrigues, C., Mendes, R., Baptista, P., and Fernandes, A. (2019). Targeting Tumor Microenvironment for Cancer Therapy. *Int. J. Mol. Sci.* 20:840. doi: 10.3390/ijms20040840

Russell, J., Carlin, S., Burke, S. A., Wen, B., Yang, K. M., and Ling, C. C. (2009). Immunohistochemical detection of changes in tumor hypoxia. *Int. J. Radiat. Oncol. Biol. Phys.* 73, 1177–1186. doi: 10.1016/j.ijrobp.2008.12.004

Sarvi, S., Patel, H., Li, J., Dodd, G. L., Creedon, H., Muir, M., et al. (2018). Kindlin-1 Promotes Pulmonary Breast Cancer Metastasis. *Cancer Res.* 78, 1484–1496. doi: 10.1158/0008-5472.Can-17-1518

Saxena, K., and Jolly, M. K. (2019). Acute vs. chronic vs. cyclic hypoxia: their differential dynamics, molecular mechanisms, and effects on tumor progression. *Biomolecules* 9:339. doi: 10.3390/biom9080339

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Swinson, D. E., and O'Byrne, K. J. (2006). Interactions between hypoxia and epidermal growth factor receptor in non-small-cell lung cancer. *Clin. Lung Cancer* 7, 250–256. doi: 10.3816/CLC.2006.n.002

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Walsh, J. C., Lebedev, A., Aten, E., Madsen, K., Marciano, L., and Kolb, H. C. (2014). The clinical importance of assessing tumor hypoxia: relationship of tumor hypoxia to prognosis and therapeutic opportunities. *Antioxid. Redox Signal.* 21, 1516–1554. doi: 10.1089/ars.2013.5378

Ward, C., Langdon, S. P., Mullen, P., Harris, A. L., Harrison, D. J., Supuran, C. T., et al. (2013). New strategies for targeting the hypoxic tumour microenvironment in breast cancer. *Cancer Treatment Rev.* 39, 171–179. doi: 10.1016/j.ctrv.2012.08.004

Wei, C. Y., Zhu, M. X., Lu, N. H., Peng, R., Yang, X., Zhang, P. F., et al. (2019). Bioinformatics-based analysis reveals elevated MFSD12 as a key promoter of cell proliferation and a potential therapeutic target in melanoma. *Oncogene* 38, 1876–1891. doi: 10.1038/s41388-018-0531-536

Wei, J., Xu, Z., Chen, X., Wang, X., Zeng, S., Qian, L., et al. (2020). Overexpression of GSDMC is a prognostic factor for predicting a poor outcome in lung adenocarcinoma. *Mol. Med. Rep.* 21, 360–370. doi: 10.3892/mmr.2019.10837

Whiteside, T. L. (2008). The tumor microenvironment and its role in promoting tumor growth. *Oncogene* 27, 5904–5912. doi: 10.1038/onc.2008.271

Widmer, D. S., Hoek, K. S., Cheng, P. F., Eichhoff, O. M., Biedermann, T., Raaijmakers, M. I. G., et al. (2013). Hypoxia contributes to melanoma heterogeneity by triggering HIF1alpha-dependent phenotype switching. *J. Invest. Dermatol.* 133, 2436–2443. doi: 10.1038/jid.2013.115

Winter, S. C., Buffa, F. M., Silva, P., Miller, C., Valentine, H. R., Turley, H., et al. (2007). Relation of a Hypoxia Metagene Derived from Head and Neck Cancer to Prognosis of Multiple Cancers. *Cancer Res.* 67, 3441–3449. doi: 10.1158/0008-5472.Can-06-3322

Xia, X., Wang, X., Cheng, Z., Qin, W., Lei, L., Jiang, J., et al. (2019). The role of pyroptosis in cancer: pro-cancer or pro-"host"? *Cell Death Dis.* 10:650. doi: 10.1038/s41419-019-1883-1888

Xiang, B., Chatti, K., Qiu, H., Lakshmi, B., Krasnitz, A., Hicks, J., et al. (2008). Brk is coamplified with ErbB2 to promote proliferation in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12463–12468. doi: 10.1073/pnas.0805009105

Xu, Y., Han, W., Xu, W. H., Wang, Y., Yang, X. L., Nie, H. L., et al. (2019). Identification of differentially expressed genes and functional annotations associated with metastases of the uveal melanoma. *J. Cell Biochem.* 120, 19202–19214. doi: 10.1002/jcb.29250

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zerilli, M., Zito, G., Martorana, A., Pitrone, M., Cabibi, D., Cappello, F., et al. (2010). BRAF(V600E) mutation influences hypoxia-inducible factor-1alpha expression levels in papillary thyroid cancer. *Mod. Pathol.* 23, 1052–1060. doi: 10.1038/modpathol.2010.86

Zhang, E. Y., Cristofanilli, M., Robertson, F., Reuben, J. M., Mu, Z., Beavis, R. C., et al. (2013). Genome Wide Proteomics of ERBB2 and EGFR and Other Oncogenic Pathways in Inflammatory Breast Cancer. *J. Proteome Res.* 12, 2805–2817. doi: 10.1021/pr4001527

Zhang, Q., Wang, Y., Liang, J., Tian, Y., Zhang, Y., and Tao, K. (2017). Bioinformatics analysis to identify the critical genes, microRNAs and long noncoding RNAs in melanoma. *Medicine* 96:e7497. doi: 10.1097/MD.0000000000007497

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 October 2020
doi: 10.3389/fbioe.2020.559387

Prediction of Proximal Junctional Kyphosis After Posterior Scoliosis Surgery With Machine Learning in the Lenke 5 Adolescent Idiopathic Scoliosis Patient

Li Peng[1]†, Lan Lan[1]†, Peng Xiu[2], Guangming Zhang[1], Bowen Hu[2], Xi Yang[2], Yueming Song[2], Xiaoyan Yang[1], Yonghong Gu[1], Rui Yang[3] and Xiaobo Zhou[4]*

[1] West China Biomedical Big Data Center, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, China, [2] Department of Orthopedic Surgery, West China Hospital, Sichuan University, Chengdu, China, [3] Department of Ultrasound, West China Hospital, Sichuan University, Chengdu, China, [4] Center for Computational Systems Medicine, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Zhibin Lv,
University of Electronic Science and Technology of China, China
Mengran Zhou,
Anhui University of Science and Technology, China

***Correspondence:**
Li Peng
2018224020165@stu.scu.edu.cn;
pengli_bonne@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 May 2020*
Accepted: *08 September 2020*
Published: *06 October 2020*

Citation:
Peng L, Lan L, Xiu P, Zhang G, Hu B, Yang X, Song Y, Yang X, Gu Y, Yang R and Zhou X (2020) Prediction of Proximal Junctional Kyphosis After Posterior Scoliosis Surgery With Machine Learning in the Lenke 5 Adolescent Idiopathic Scoliosis Patient. Front. Bioeng. Biotechnol. 8:559387. doi: 10.3389/fbioe.2020.559387

Objective: To build a model for proximal junctional kyphosis (PJK) prognostication in Lenke 5 adolescent idiopathic scoliosis (AIS) patients undergoing long posterior instrumentation and fusion surgery by machine learning and analyze the risk factors for PJK.

Materials and Methods: In total, 44 AIS patients (female/male: 34/10; PJK/non-PJK: 34/10) who met the inclusion criteria between January 2013 and December 2018 were retrospectively recruited from West China Hospital. Thirty-seven clinical and radiological features were acquired by two independent investigators. Univariate analyses between PJK and non-PJK groups were carried out. Twelve models were built by using four types of machine learning algorithms in conjunction with two oversampling methods [the synthetic minority technique (SMOTE) and random oversampling]. Area under the receiver operating characteristic curve (AUC) was used for model discrimination, and the clinical utility was evaluated by using F1 score and accuracy. The risk factors were simultaneously analyzed by a Cox regression and machine learning.

Results: Statistical differences between PJK and non-PJK groups were as follows: gender (p = 0.001), preoperative factors [thoracic kyphosis (p = 0.03), T1 slope angle (T1S, p = 0.078)], and postoperative factors [T1S (p = 0.097), proximal junctional angle (p = 0.003), upper instrumented vertebra (UIV) – UIV + 1 (p = 0.001)]. Random forest using SMOTE achieved the best prediction performance with AUC = 0.944, accuracy = 0.909, and F1 score = 0.667 on independent testing dataset. Cox model revealed that male gender and larger preoperative T1S were independent prognostic factors of PJK (odds ratio = 10.701 and 57.074, respectively). Gender was also at the first place in the importance ranking of the model with best performance.

Conclusion: The random forest using SMOTE model has the great value for predicting the individual risk of developing PJK after long instrumentation and fusion surgery in Lenke 5 AIS patients. Moreover, the combination of the outcomes of a Cox model and the feature ranking extracted by machine learning is more valuable than any one alone, especially in the interpretation of risk factors.

Keywords: spinal deformity, proximal junctional kyphosis, sagittal malalignment, machine learning, prediction model

INTRODUCTION

For adolescent idiopathic scoliosis (AIS) patients, orthopedic operations are employed to reconstruct the coronal and sagittal alignment in an attempt to maintain the stability of the spine (Mimura et al., 2017). Long posterior instrumentation and fusion surgery is the preferred treatment strategy for improving the management of progressive scoliotic spines (Suk et al., 1995). Although all the efforts have been made to design a suitable operative procedure, the prognosis is not always satisfactory (Humke et al., 1995; Bridwell, 1997). Proximal junctional kyphosis (PJK), a multifactorial proximal adjacent segment disease following fusion treatment, has drawn the attention of many spine surgeons (Watanabe et al., 2010; Kim et al., 2013). It affects around 28% of the adolescent idiopathic scoliosis (AIS) population, with regional pain and poor life quality in some severe cases (Kim et al., 2007; O'Shaughnessy et al., 2012; Passias et al., 2018; Sebaaly et al., 2018). The most commonly adopted definition of PJK is accepted in this study: the Cobb angle between the upper instrumented vertebra (UIV) and the two supra-adjacent vertebrae is superior to 10° and at least 10° greater than its preoperative value (Glattes et al., 2005).

Currently, most researchers are devoted to extracting proper prognostic information by using statistical methods to have an insight into the characteristics with high risks (Kim et al., 2008; Scheer et al., 2016). Previous studies also showed the potential of binary logistic regression in risk factors identification, such as old age, gender, fusion levels, type of instrumentation at the UIV, and various sagittal spinopelvic radiographic parameters (Sebaaly et al., 2018; Zhao et al., 2018). To our knowledge, no reported studies analyzed cervical balance parameters in conjunction with well-known clinical prognostic factors to confirm that it is an independent risk factor for AIS patients. In addition, logistic regression models depend heavily on the linear separability of samples, which is vulnerable to the degree of multicollinearity between variables and may result in a model with underfitting and low accuracy to provide unreliable outcome prediction for a personalized surgical planning. Therefore, it seems unreasonable to make use of linear models for accurate preoperative prediction in the era of personalization of medicine. Non-linear machine learning methods (e.g., random forest) have a distinct advantage over the linear approach because they distinctly provide inherent data pattern recognition and map non-linear relationships between high-dimensional variables to estimate the clinical outcome for each individual (Karhade et al., 2019). Scheer et al. (2016) have constructed a decision tree model (accuracy = 0.860) on 510 adult spinal deformity patients by commercially available software. Nonetheless, in the study, just 13 variables were considered for the highly heterogeneous study population.

The purpose of this study was to establish preoperative risk models for Lenke 5 AIS patients undergoing long posterior instrumentation and fusion surgery. We also explored and compared the outcomes of machine learning and a commonly used model in clinic (Cox regression) at risk factor identification for PJK.

MATERIALS AND METHODS

Patient Population

The institutional review boards approved this retrospective study and waived the requirement to obtain written informed consent. Between January 2013 and December 2018, 293 AIS patients were admitted to West China Hospital. Inclusion criteria were as follows: (1) Lenke 5 curves (2) long posterior instrumentation and fusion surgery with > 6 instrumented motion segments, (3) at least 1 year follow-up; (4) adequate preoperative, immediate postoperative (3–7 days after surgery), and final follow-up anteroposterior and lateral standing long-cassette radiographs; (5) radiographs with good quality. Finally, a total of 44 Lenke 5 patients with posterior instrumentation (34 without PJK and 10 with PJK) were recruited on the basis of the eligibility criteria (**Figure 1**).

Parameters Collection

Patient demographics and surgical factors including amount of correction, upper instrumented vertebra (UIV) level, lower instrumented vertebra (LIV) level, and the number of instrumented vertebras were recorded from the electronic medical records.

Two coronal and 28 sagittal parameters were collected according to the results of previous researches on PJK (Glattes et al., 2005; Kim et al., 2005, 2013, 2014; Yagi et al., 2011; Hostin et al., 2013; Ghailane et al., 2017; Sebaaly et al., 2018; Zhao et al., 2018; Alzakri et al., 2019). Specifically, coronal parameters included the following: coronal vertical axis (CVA, offset of C7 plumb-line relative to the center sacral vertical line) and the main scoliosis curve Cobb angle (CAMSC); sagittal parameters included the following: the sagittal vertical axis (SVA, offset of C7 plumb-line relative to S1 on the sagittal plane), pelvic tilt (PT), pelvic incidence (PI), PI-LL mismatch, sacral slope (SS), upper segmental lumbar lordosis from L1 to L4 (ULL), lower segmental

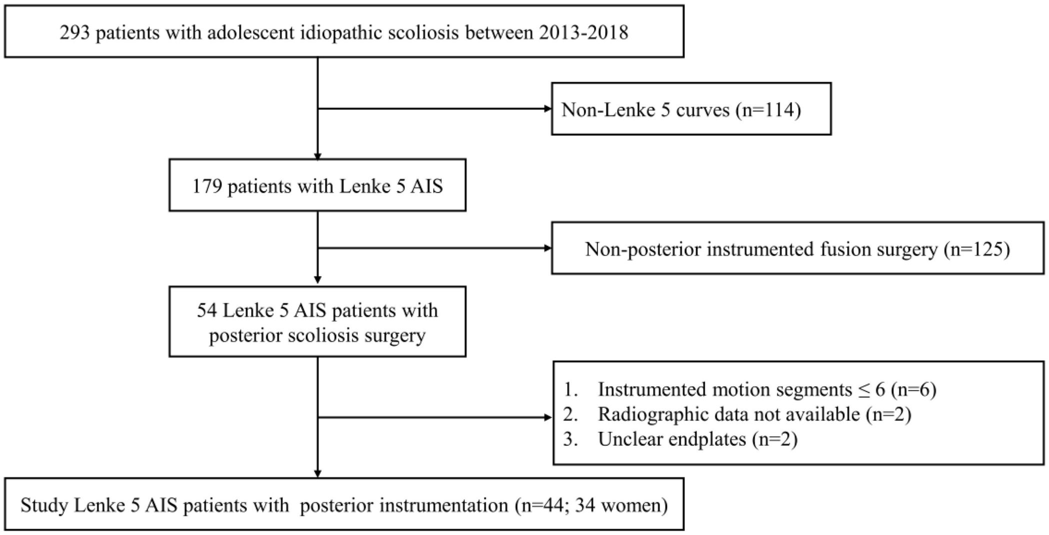

293 patients with adolescent idiopathic scoliosis between 2013-2018
Non-Lenke 5 curves (n=114)
179 patients with Lenke 5 AIS
Non-posterior instrumented fusion surgery (n=125)
54 Lenke 5 AIS patients with posterior scoliosis surgery
1. Instrumented motion segments ≤ 6 (n=6)
2. Radiographic data not available (n=2)
3. Unclear endplates (n=2)
Study Lenke 5 AIS patients with posterior instrumentation (n=44; 34 women)

FIGURE 1 | Flow diagram of patient inclusion and exclusion. AIS, adolescent idiopathic scoliosis.

lumbar lordosis from L4 to S1 (LLL), lumbar lordosis (LL, Cobb angle between superior endplate of L1 and superior endplate of S1), thoracic kyphosis (TK, Cobb angle between superior endplate of T4 and inferior endplate of T12), rod contour angle (RCA, angle between the superior plate of UIV and the inferior plate of one vertebra caudal to the UIV), UIV − UIV + 1 (angle between the inferior endplate of UIV and the superior endplate of one cephalad vertebrae), proximal junctional angle (PJA, angle between the inferior endplate of UIV and the superior endplate of two cephalad vertebrae), T1 slope (T1S, Cobb angle between a horizontal line and the upper endplate of T1), and T1SpinoPelvic inclination (T1SPI, the angle between the vertical plumb-line and the line drawn from vertebral body center of T1 and the center of the bicoxofemoral axis).

It is worthy of note that the value of PI was constant before and after surgery; thus, we only demanded the preoperative PI. Moreover, RCA was defined as a postoperative variable as stated by Kim et al. (2007) and Lonner et al. (2017). The specific measurement methods are presented in **Figures 2**, **3**.

Univariate Analyses

Continuous and categorical data were shown as mean ± standard deviation and numbers with percentages in parentheses, respectively. Shapiro-Wilk test was performed to test the normality of data distribution. Two-sided Student *t*-test (for normally distributed data) and Mann–Whitney–Wilcoxon test (for non-parametric data) were used to determine the statistical differences in continuous data between PJK and non-PJK groups, whereas chi-square test was performed for categorical variables. $p < 0.1$ was indicative of a statistically significant difference.

Machine Learning Model Construction

Thirty-seven variables were normalized to reduce the effect of data scale while maintaining the distributions of original data. Data were split into training and testing sets at a random stratified ratio of 3:1 by preserving the percentage of samples for each class, and the testing set was held out for examining the generalization ability of the models. To address the class imbalance problem which could lead to a severely imbalanced degree of accuracy with the majority class having nearly 100% accuracy while the minority one having worse accuracy of 0–10%, two oversampling methods, the synthetic minority technique (SMOTE) (Chawla et al., 2002) and random oversampling (ROS) (He and Garcia, 2009), were applied for model training (Mendoza-Lattes et al., 2011; Lei et al., 2016, 2017; Lan, 2017; Sebaaly et al., 2018).

We established four kinds of popular supervised machine learning models [random forest (RF), support vector machine (SVM), k neighbors classifier (KNN), and linear regression (LR)] for risk prediction, which had superior advantages in solving the small-sample size problem. The parameters of the model were optimized by cross-validated grid search over a parameter grid, such as the number of estimators and criterion and the minimum number of samples required to split for RF; kernel, regularization parameter, and gamma for SVM; and number of neighboring samples, power parameter for the Minkowski metric, and weight function for KNN (Swami and Jain, 2012; Peng et al., 2016; Zhao et al., 2019). Leave-one-out cross-validation was implemented to evaluate the performance of models in training stage. More specifically, one patient from all patients was used for model testing while the rest for training, and these procedures were repeated until each patient had been used once as a testing sample. Final evaluation was be done with the independent test set as the model training was fulfilled to reflect the ability of a model to unknown sample.

Model discrimination was measured by area under the receiver operating characteristic curve (AUC). Accuracy was used to assess the difference between the predicted clinical results (PJK) and ground truth derived from follow-up study. The clinical utility of the model was also evaluated with F1 score,

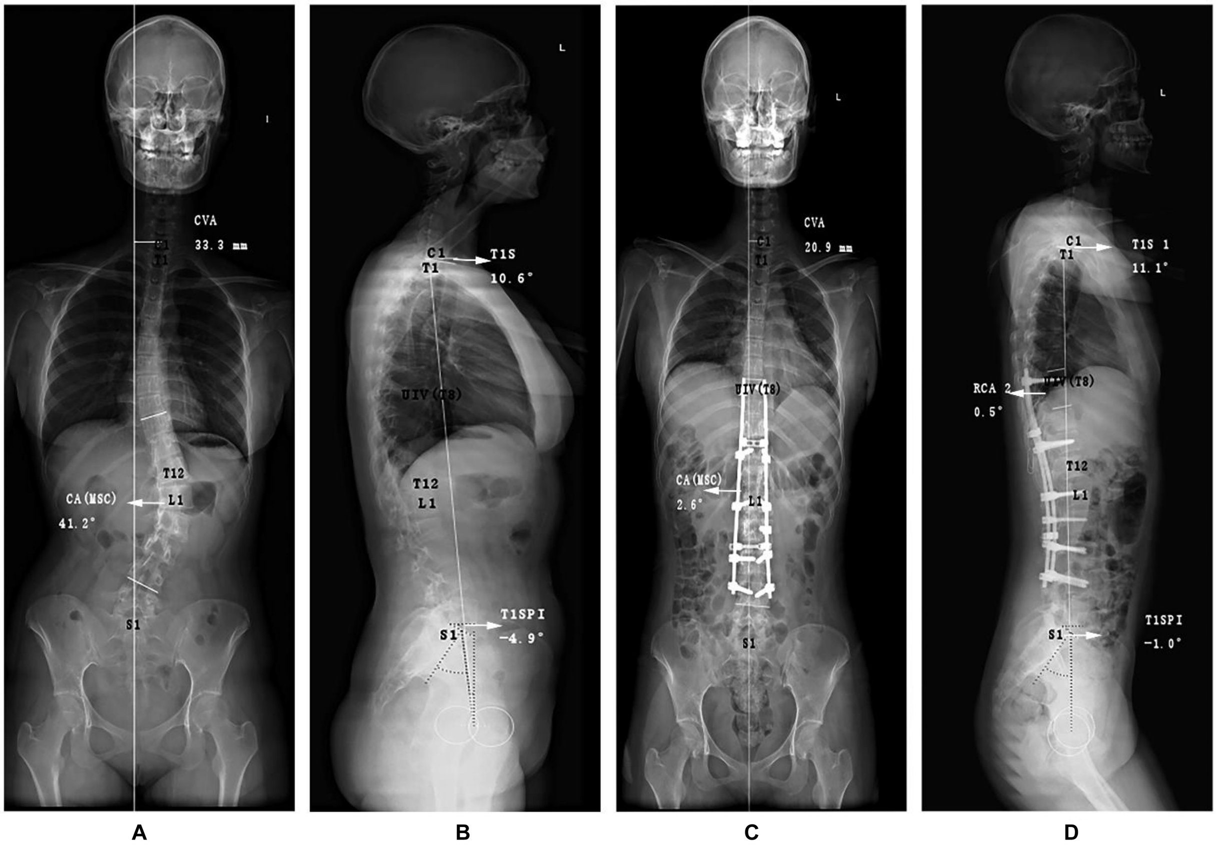


FIGURE 2 | Graphic representations of special angles of an adolescent idiopathic scoliosis patient with PJK postoperatively. Different from the conventional measurements, the **(A)** anteroposterior and **(B)** lateral preoperative radiographs purposely included the following measurements for demonstrating coronal and sagittal malalignment: coronal vertical axis (CVA), the main curve coronal angle (CAMSC), T1 Slope (T1S), and T1SpinoPelvic inclination (T1SPI). At immediate postoperative X-ray films **(C,D)**, rod curve angle (RCA) was also measured. PJK, proximal junctional kyphosis.

which is a necessary synthesized indicator by conveying the balance between the precision and the recall in imbalanced dataset (Chawla et al., 2002). At last, the model with the best prognostic performance was considered as the final prediction model to obtain the feature importance in PJK occurrence by ranking factor influences (Ji et al., 2015). Python version 3.5 (Python Software Foundation, Wilmington, DE, United States) was used for modeling.

Cox Proportional Hazards Regression

A Cox proportional hazards regression model was also applied to select PJK-related features. Event-free survival was defined as the time from the date of surgery to the date of PJK occurrence. Follow-up time for patients without complications were censored at the last visit, and PJK patients contributed follow-up time until the outcomes were first recorded. The predictors of PJK with statistical significance in the univariable analysis were included in the multivariable Cox model. The final model was selected by forward Wald method. And the proportional hazards assumption of models was verified by examining the scaled Schoenfeld residual plots. The results were compared with the feature importance information acquired by machine learning model for exploring the interpretability and predictive value of variables. Statistical analysis was performed using SPSS 25.0 (IBM Corp., Armonk, NY).

RESULTS

Clinical Characteristics

Tables 1, **2** show detailed baseline and clinical-radiologic characteristics of all patients. A total of 44 patients (female/male: 34/10) were recruited for this study. The average age at surgery, follow-up time, and instrumented vertebras were 18.27 ± 3.61 years, 3.15 ± 2.67 years, 6.80 ± 1.37 vertebras, respectively. At final follow-up, there were 10 (22.7%) patients with PJK, while 34 patients demonstrated no significant PJK by follow-up investigation.

Between PJK and non-PJK groups, significant differences ($p < 0.1$) were observed in the following variables: gender distribution ($p = 0.001$), preoperative TK ($p = 0.03$), preoperative T1S ($p = 0.078$), postoperative T1S ($p = 0.097$), PJA ($p = 0.003$), and postoperative UIV − UIV + 1 ($p = 0.001$). However, there were no differences in age at surgery, body mass

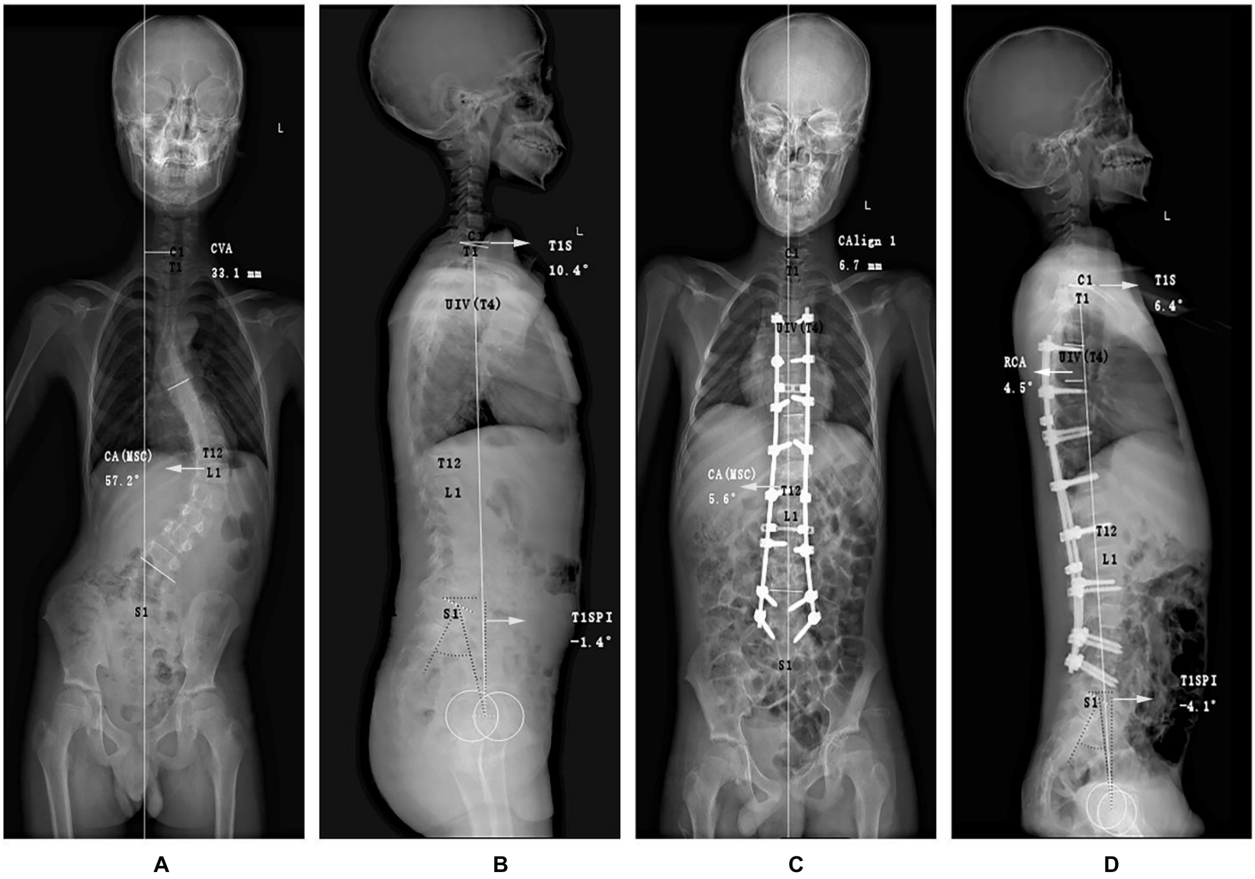


FIGURE 3 | Graphic representations of special angles of an adolescent idiopathic scoliosis patient without PJK postoperatively. Different from the conventional measurements, the **(A)** anteroposterior and **(B)** lateral preoperative radiographs purposely included the following measurements for demonstrating coronal and sagittal malalignment: coronal vertical axis (CVA), the main curve coronal angle (CAMSC), T1 Slope (T1S), and T1SpinoPelvic inclination (T1SPI). At immediate postoperative X-ray films **(C,D)**, rod curve angle (RCA) was also measured. PJK, proximal junctional kyphosis.

TABLE 1 | Demographic and clinical variables.

Variable	**None (*n* = 34)**	**PJK (*n* = 10)**	***p*-value**
Age at surgery, mean ± SD	18.50 ± 3.71	17.50 ± 3.31	0.669
Gender Female, ***n*** (%) Male, ***n*** (%)	30 (68.2%) 4 (9.1%)	4 (9.1%) 6 (13.6%)	**0.001* (χ ^{2} = 10.23)**
BMI	18.84 ± 2.39	21.29 ± 6.31	0.216
Amount of correction	80.4% ± 14.5%	81.6% ± 12.6%	0.901
Follow-up time (years)	2.88 ± 1.32	4.22 ± 4.54	0.648
UIV levels T1–T5 T6–T9 T10–T12	6 (13.6%) 20 (45.5%) 8 (18.2%)	2 (4.5%) 8 (18.2%) 0	0.232 (c^2 = 2.92)
LIV levels L3 L4 L5	6 (13.6%) 18 (40.9%) 10 (22.7%)	5 (11.4%) 3 (6.9%) 2 (4.5%)	0.114 (χ^2 = 4.338)
Number of instrumented vertebrae, mean ± SD	6.85 ± 1.31	6.60 ± 1.075	0.344

*Bold and * values both represent a statistically significant difference between the PJK and None groups. SD, mean ± standard deviations; BMI, body mass index; UIV, upper instrumented vertebra; LIV, lower instrumented vertebra.*

index (BMI), amount of correction, and UIV and LIV levels (**Table 1**). Additionally, no significant differences were observed in preoperative data including CAMSC, CVA, LL, ULL, LLL, SVA, PT, PI, SS, PI-LL mismatch, T1SPI, PJA, and UIV – UIV + 1, and immediate postoperative parameters including CAMSC, CVA, TK, LL, ULL, LLL, SVA, PT, PI, SS, PI-LL mismatch, T1SPI, and RCA (**Table 2**).

Machine Learning Results

The average accuracies of machine learning models without oversampling for predicting PJK occurrence in the train and test sets were 0.728 and 0.783, whereas, models trained with ROS were 0.80 and 0.73, and models with SMOTE were 0.82 and 0.78, respectively. The average AUC for models without oversampling, with ROS, and with SMOTE were 0.64,

TABLE 2 | Radiographic variables.

Abbreviation	Parameter	Type	None ($n = 34$)	PJK ($n = 10$)	*p*-value
Coronal parameters					
CAMSC	The coronal main scoliosis curve Cobb angle (°)	Pre Post	43.96 ± 10.60 8.56 ± 6.31	39.74 ± 5.91 7.01 ± 4.81	0.237 0.478
CVA	Coronal vertical axis (mm)	Pre Post	14.01 ± 14.42 5.89 ± 18.86	16.69 ± 18.00 8.57 ± 12.18	0.628 0.675
Sagittal parameters					
TK	Thoracic kyphosis (°)	Pre Post	21.45 ± 9.42 19.35 ± 8.66	29.51 ± 11.68 23.50 ± 10.07	**0.030*** 0.207
LL	Lumber lordosis (°)	Pre Post	47.67 ± 11.43 48.54 ± 9.11	52.64 ± 10.20 47.20 ± 7.61	0.772 0.673
ULL	Upper segmental lordosis from L1 to L4 (°)	Pre Post	19.13 ± 9.19 20.55 ± 5.84	20.60 ± 934 19.20 ± 7.16	0.660 0.544
LLL	Lower segmental lordosis from L4 to S1 (°)	Pre Post	34.88 ± 8.84 31.23 ± 8.31	39.11 ± 10.97 30.11 ± 7.34	0.215 0.704
SVA	Sagittal vertical axis (°)	Pre Post	−8.77 ± 27.36 3.13 ± 30.83	1.39 ± 15.81 10.85 ± 25.78	0.271 0.476
PT	Pelvic tilt (°)	Pre Post	7.23 ± 7.84 3.68 ± 8.43	4.96 ± 10.03 3.31 ± 9.14	0.452 0.905
PI	Pelvic incidence (°)	Pre	45.12 ± 11.48	42.30 ± 13.61	0.514
SS	Sacral slope (°)	Pre Post	37.89 ± 8.73 39.70 ± 8.23	37.34 ± 5.87 37.72 ± 5.80	1.000 0.483
PI-LL mismatch	Pelvic incidence-lumbar lordosis mismatch (°)	Pre Post	−3.92 ± 12.69 −2.88 ± 9.30	10.35 ± 14.77 −6.10 ± 13.89	0.196 0.397
T1S	T1 slope (°)	Pre Post	14.43 ± 7.50 12.33 ± 7.93	19.45 ± 8.47 16.93 ± 5.82	**0.078* 0.097***
T1SPI	T1SpinoPelvic inclination (°)	Pre Post	−4.07 ± 3.37 −1.87 ± 4.19	−3.05 ± 2.57 −1.03 ± 3.47	0.383 0.569
PJA	(°)	Pre Post	8.23 ± 5.45 7.74 ± 5.23	8.89 ± 3.64 13.35 ± 4.28	0.745 **0.003***
UIV – UIV + 1	(°)	Pre Post	4.93 ± 3.47 4.45 ± 3.12	6.63 ± 2.97 8.13 ± 3.38	0.115 **0.001***
RCA	Rod contour angle (°)	Post	4.03 ± 2.60	5.71 ± 4.56	0.464

*All values are shown as mean ± SD. *Values represent a statistically significant difference ($p < 0.1$) between the PJK and None groups. Pre, preoperative; Post, immediate postoperative (3–7 days after surgery); UIV, upper instrumented vertebrae.*

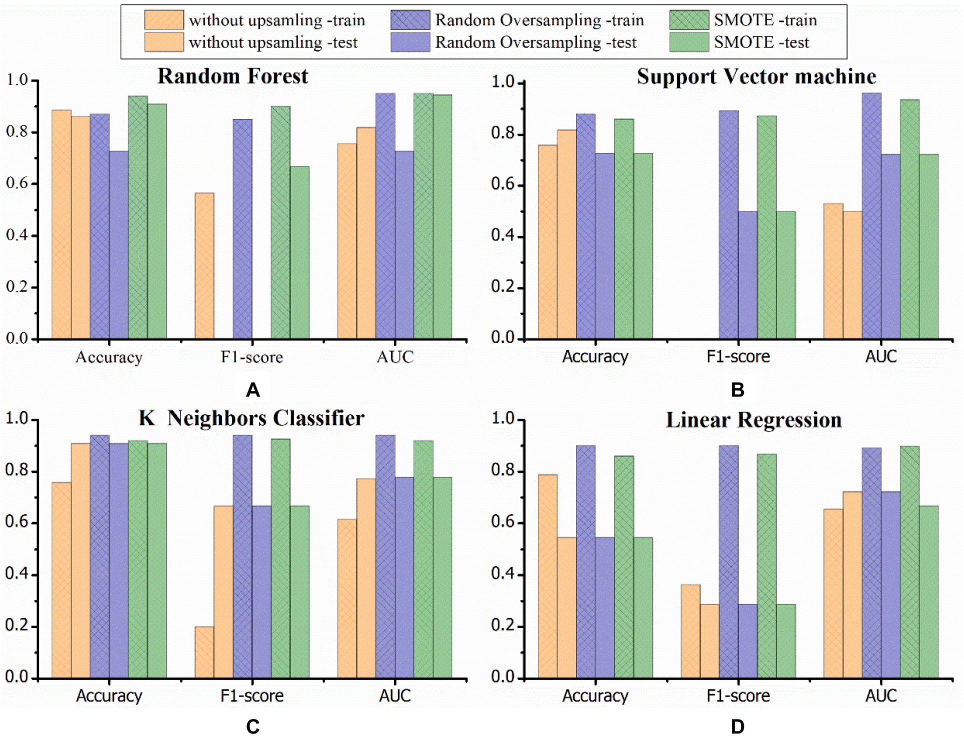


FIGURE 4 | Graphs show the performances for PJK risk prediction obtained by established models in the training and testing sets. Three colors demonstrate different data processing methods (orange, without data processing; blue, random oversampling; green, SMOTE). Random forest combined with SMOTE provided an excellent prediction performance compared with rival models. SMOTE, the synthetic minority technique; AUC, area under the receiver operating characteristic curve; PJK, proximal junctional kyphosis. **(A–D)** Respectively represent the model performance of random forest, support vector machine, K neighbors classifier, linear regression.

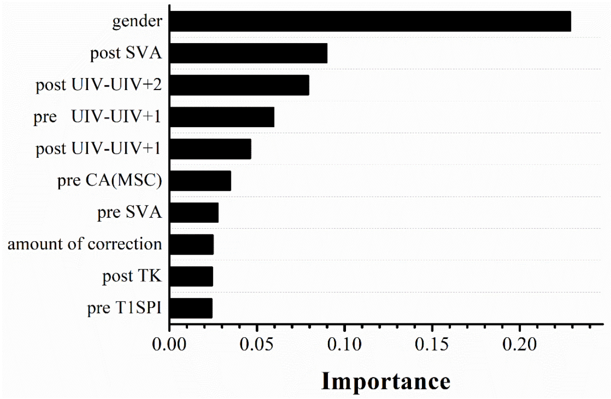


FIGURE 5 | Importance order of top 10 predictors (importance = 64%) ranked by random forest using SMOTE. SMOTE, the synthetic minority technique; pre, pre-operation; post, immediate postoperative (3–7 days after surgery); SVA, sagittal vertical axis; UIV, upper instrumented vertebrae; CA(MSC), the main scoliosis curve coronal angle; TK, thoracic kyphosis; T1SPI, T1SpinoPelvic inclination.

0.86, and 0.82 in the train set, respectively, and 0.70, 0.74, and 0.78 in the test set. The F1 score performances of the models that trained with oversampling were superior to that of without oversampling in both sets (train: 0.70 vs. 0.38, 0.70 vs. 0.24; test: 0.37 vs. 0.24, 0.53 vs. 0.24). The general tendency was that models with data oversampling had better robustness than the ones without preprocessing, and models that integrated SMOTE in the training stage yielded the best prognostic performance.

Discriminatory performance and prediction accuracy of all models in leave-one-out cross-validation and test set are shown in **Figure 4**. Random forest using SMOTE provided better prognostic ability (AUC = 0.944), better clinical usefulness compared with rival models (accuracy = 0.909, F1 score = 0.667), and low operation time (4 ms for each sample) in independent test set, whereas, linear regression had the worst performance (AUC = 0.545, F1 score = 0.228, accuracy = 0.704), suggesting non-linear machine learning models had more precise prognostication. The detailed prediction outputs of this model were nine true negative, one false negative, one true positive, and zero false positive on test data set, demonstrating a lower misdiagnosis rate. In addition, the model presented feature selection based on data attributes importance ranking, and the top 10 prognostic indicators were gender, four preoperative features (UIV – UIV + 1, CAMSC, SVA, and T1SPI), and five modifiable surgical features (SVA, PJA, UIV – UIV + 1, TK, and amount of correction) (**Figure 5**).

Multivariable Proportional Hazards Regression Model

To compare the feature selection results with the risk factors of PJK obtained by a model widely used for clinical research, a Cox proportional hazards regression model was also used. There were no significant violations of the proportional hazards assumption assessed by Schoenfeld residuals against time for all six statistically significant variables at univariable analysis. Multivariable Cox model based on aforementioned parameters demonstrated that male gender and larger preoperative T1S were the independent risk factors [odds ratio (OR) = 10.701 and 57.074, respectively] in **Table 3**. Gender was at the first place on the importance ranking in RF model, which accounted for 22.9%, compared with 1.2% of preoperative T1S.

DISCUSSION

The aim of our study was to develop prognostic models in Lenke 5 AIS patients undergoing long posterior instrumentation and fusion surgery and simultaneously explore the predictive value of clinical factors for PJK. We concluded that random forest that trained with SMOTE exhibited better performance in PJK prediction compared with other models. Specifically, in independent test set, the model provided better prognostic ability (AUC = 0.944, accuracy = 0.909, F1 score = 0.667) compared with other rival models, suggesting the reproducibility and reliability of the proposed model. In addition, a multivariable Cox model revealed that male gender and larger preoperative T1S were the independent prognostic factors for PJK (OR of male gender, 10.701 and OR of preoperative T1S, 57.074), and gender also ranked the first place with the prognostic importance of 22.9% in our prediction model.

For AIS patients, PJK was a complication after corrective surgery with unknown causation, and 22.7% of the patients in our study developed PJK (Hollenbeck et al., 2008; Zhao et al., 2018). The occurrence of PJK is multifactorial, including clinical, surgical, and radiographic factors. Linear regressions, such as binary logistic regression, may be simple and transparent for data analysis, however, they are not able to meet the needs of distinguishing high-dimensional and linear inseparable input data. Conversely, the power and potential of machine learning are increasingly recognized in the field of scoliosis correction (Group et al., 2015). In our study, we established four classes of models for PJK prediction. Models trained with oversampling methods

TABLE 3 | Cox proportional hazards regression model (forward Wald method) for risk factors of PJK.

Variable	B	SE	Wald	df	Sig.	Exp(B)	95%CI for Exp(B)	
							Lower	Upper
Male gender	2.370	0.719	8.804	1	0.002	10.701	2.062	34.510
Preoperative T1S	4.044	0.753	9.909	1	0.022	57.074	2.446	46.813

T1S, T1 slope.

showed relatively higher discrimination ability than that without using oversampling, suggesting rebalancing the class distribution for an imbalanced dataset was favorable to the construction of classifiers. In fact, SMOTE oversamples minority class by creating "synthetic" examples to build larger decision regions that contain nearby minority class points, rather than by oversampling with replacement, which actually diminishes and specifies the decision region for the minority class (Chawla et al., 2002; Ji et al., 2016). Our results also showed that random forest using SMOTE would be a useful approach that could effectively evaluate the risk of PJK postoperatively for patients with scoliosis in real time. In addition, the models may facilitate individualized surveillance policy. Specifically, low-risk patients may receive a less intensive surveillance regimen, even within the first year after surgery.

We carefully considered the potential risk factors for PJK. Several disputable factors were controlled in our study, including age, gender, TK, postoperative PJA, and UIV location. For example, UIV located in the lower thoracic region is a risk factor for PJK in Zhao et al. (2018), however, Zhao et al. recruited more PJK patients corrected by selected fusion with UIV stopping at lower thoracic levels, whereas, UIV always tended to stop at the upper thoracic regions (upper/lower: 36/8) in our study, which decreased the risk of PJK. In addition, we also included cervical alignment parameters in the analysis. T1S and male gender were independent risk factors in the multivariable Cox model when adjusting for other clinical prognostic factors. In fact, researchers have found that if middle or upper thoracic segments were fused, the postoperative compensation of cervical curvature would occur during the follow-up period (Sebaaly et al., 2018; Alzakri et al., 2019; Buell et al., 2019). We inferred that the proximal kyphosis might aggravate in PJK group to balance the cervical curvature for maintaining the global balance. Controversy exists on whether gender has an effect on the incidence of PJK or not. In accordance with Kim et al. (2007), which retrospectively assessed 410 patients and demonstrated that male gender had higher prevalence than female gender, our findings also suggested that male gender correlated significantly with PJK, although the underlying reasons were unclear.

Even though there were no differences in other sagittal spinopelvic parameters in Cox regression analysis, their importance in compensating for the misalignment of the spine in the long-term follow-up could not be ignored. In fact, the random forest model demonstrated that the top 10 prognostic indicators were gender, four preoperative features (UIV – UIV + 1, CAMSC, SVA, T1SPI), and five modifiable surgical parameters (SVA, PJA, UIV – UIV + 1, TK, amount of correction). Accordingly, the common points and differences between the results of the Cox model and the feature ranking extracted by the random forest model certified the significance of combined use of machine learning and statistical analysis. Five modifiable parameters of the prediction model may further supply a detailed assistant decision-making for preoperative surgical plan. We believe that our prediction models would affect operational design by individualizing management according to the risk profiles for PJK occurrence.

Our study had limitations. First, we developed our model for the Lenke 5 AIS patients, the most common Lenke type (Yang, 2003). However, further validation studies are warranted for other scoliosis types. Second, it was a retrospective analysis that suffers from inherent biases, although an independent data set was conducted to improve the reliability. Third, the sample size of this study was relatively small; our results require further validation with other institutions to check for the generalizability.

CONCLUSION

In conclusion, the random forest using SMOTE model has great value for predicting the individual risk of developing PJK after long instrumentation and fusion surgery in Lenke 5 AIS patients. The model may facilitate clinical decision making in the era of precision medicine for spinal orthopedics. The combination of the results of a Cox model and the feature ranking extracted by machine learning is a promising approach to identify prognostic factors and has great significance in the medical field. Further studies are required to explore the generalized utility of our model and translate the results into clinical practice.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because, the datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. Requests to access the datasets should be directed to LP, pengli_bonne@163.com.

AUTHOR CONTRIBUTIONS

LP, XZ, and GZ conceived and launched this study. XiaoY and YG designed the medical and statistical analysis. YS, PX, BH, and XiY collected cases and clinical diagnosis. LP and RY took the angle measurements in X-rays. LP and LL analyzed the data, carried out statistical experiments, and wrote the first draft of this manuscript. LL and XZ revised and edited the final version. All authors reviewed and approved the manuscript.

FUNDING

This work was supported by 1.3.5 project for disciplines of excellence-Clinical Research Incubation Project, West China Hospital, Sichuan University (Grant 2019HXFH022), 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (Grant ZYJC18010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.559387/full#supplementary-material

REFERENCES

Alzakri, A., Vergari, C., Van den Abbeele, M., Gille, O., Skalli, W., and Obeid, I. (2019). Global sagittal alignment and proximal junctional kyphosis in adolescent idiopathic scoliosis. *Spine Deformity* 7, 236–244. doi: 10.1016/j.jspd.2018.06.014

Bridwell, K. H. (1997). Spinal Instrumentation in the management of adolescent scoliosis. *Clin. Orthop. Relat. Res.* 335, 64–72. doi: 10.1097/00003086-199702000-00007

Buell, T. J., Chen, C.-J., Quinn, J. C., Buchholz, A. L., Mazur, M. D., Mullin, J. P., et al. (2019). Alignment risk factors for proximal junctional kyphosis and the effect of lower thoracic junctional tethers for adult spinal deformity. *World Neurosurg.* 121, e96–e103. doi: 10.1016/j.wneu.2018.08.242

Chawla, N., Bowyer, K., O Hall, L., and Philip Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique. *arXiv* [Preprint]. doi: 10.1613/jair.953

Ghailane, S., Pesenti, S., Peltier, E., Choufani, E., Blondel, B., and Jouve, J. L. (2017). Posterior elements disruption with hybrid constructs in AIS patients: is there an impact on proximal junctional kyphosis? *Arch. Orthop. Trauma Surg.* 137:631. doi: 10.1007/s00402-017-2684-0

Glattes, R. C., Bridwell, K. H., Lenke, L. G., Kim, Y. J., and Rinella, A. (2005). Proximal junctional kyphosis in adult spinal deformity following long instrumented posterior spinal fusion: incidence, outcomes, and risk factor analysis. *Spine* 30, 1643–1649. doi: 10.1097/01.brs.0000169451.76359.49

Group, I. S. S., Scheer, J. K., Smith, J. S., Schwab, F. J., Lafage, V., Shaffrey, C. I., et al. (2015). Development of validated computer based preoperative predictive model for proximal junction failure or clinically significant proximal junction kyphosis with 86% accuracy based on 510 adult spinal deformity patients with two-year follow-up. *Spine J.* 15, S137–S138.

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Hollenbeck, S. M., Glattes, R. C., Asher, M. A., Lai, S. M., and Burton, D. C. (2008). The prevalence of increased proximal junctional flexion following posterior instrumentation and arthrodesis for adolescent idiopathic scoliosis. *Spine* 33, 1675–1681. doi: 10.1097/BRS.0b013e31817b5bea

Hostin, R., McCarthy, I., O'Brien, M., Bess, S., Line, B., Boachie-Adjei, O., et al. (2013). Incidence, mode, and location of acute proximal junctional failures after surgical treatment of adult spinal deformity. *Spine* 38, 1008–1015. doi: 10.1097/brs.0b013e318271319c

Humke, T., Grob, D., Scheier, H., and Siegrist, H. (1995). Cotrel-Dubousset and Harrington Instrumentation in idiopathic scoliosis: a comparison of long-term results. *Eur. Spine J.* 4, 280–283. doi: 10.1007/bf00301034

Ji, Z., Meng, G., Huang, D., Yue, X., and Wang, B. (2015). NMFBFS: a NMF-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma. *Comput. Math. Methods Med.* 2015:846942. doi: 10.1155/2015/846942

Ji, Z., Su, J., Wu, D., Peng, H., and Zhao, W. (2016). Predicting the impact of combined therapies on myeloma cell growth using a hybrid multi-scale agent-based model. *Oncotarget* 8, 7647–7665. doi: 10.18632/oncotarget.13831

Karhade, A. V., Ogink, P. T., Thio, Q., Broekman, M. L. D., Cha, T. D., Hershman, S. H., et al. (2019). Machine learning for prediction of sustained opioid prescription after anterior cervical discectomy and fusion. *Spine J.* 19, 976–983. doi: 10.1016/j.spinee.2019.01.009

Kim, H. J., Bridwell, K. H., Lenke, L. G., Park, M. S., Ahmad, A., Song, K. S., et al. (2013). Proximal junctional kyphosis results in inferior SRS pain subscores in adult deformity patients. *Spine* 38, 896–901. doi: 10.1097/BRS.0b013e3182815b42

Kim, H. J., Bridwell, K. H., Lenke, L. G., Park, M. S., Song, K. S., Piyaskulkaew, C., et al. (2014). Patients with proximal junctional kyphosis requiring revision surgery have higher postoperative lumbar lordosis and larger sagittal balance corrections. *Spine* 39, E576–E580. doi: 10.1097/brs.0000000000000246

Kim, Y. J., Bridwell, K. H., Lenke, L. G., Glattes, C. R., Rhim, S., and Cheh, G. (2008). Proximal junctional kyphosis in adult spinal deformity after segmental posterior spinal instrumentation and fusion: minimum five-year follow-up. *Spine* 33, 2179–2184. doi: 10.1097/BRS.0b013e31817c0428

Kim, Y. J., Bridwell, K. H., Lenke, L. G., Kim, J., and Cho, S. K. (2005). Proximal junctional kyphosis in adolescent idiopathic scoliosis following segmental posterior spinal instrumentation and fusion: minimum 5-year follow-up. *Spine* 30, 2045–2050. doi: 10.1097/01.brs.0000179084.45839.ad

Kim, Y. J., Lenke, L. G., Bridwell, K. H., Kim, J., Cho, S. K., Cheh, G., et al. (2007). Proximal junctional kyphosis in adolescent idiopathic scoliosis after 3 different types of posterior segmental spinal instrumentation and fusions: incidence and risk factor analysis of 410 cases. *Spine* 32, 2731–2738. doi: 10.1097/BRS.0b013e31815a7ead

Lan, L. (2017). Influencing factors of inpatient expenditure pattern for cancer in China, 2015. *Chinese J. Cancer Res.* 29, 11–17. doi: 10.21147/j.issn.1000-9604.2017.01.02

Lei, Y., Yuan, W., Wang, H., Wenhu, Y., and Bo, W. (2017). A skin segmentation algorithm based on stacked autoencoders. *IEEE Trans. Multimedia* 19, 740–749. doi: 10.1109/TMM.2016.2638204

Lei, Y., Yuan, W., Wang, H., You, W., and Bo, W. (2016). A skin segmentation algorithm based on stacked autoencoders. *IEEE Trans. Multimedia* 99, 740–749. doi: 10.1109/tmm.2016.2638204

Lonner, B. S., Ren, Y., Newton, P. O., Shah, S. A., Samdani, A. F., Shufflebarger, H. L., et al. (2017). Risk factors of proximal junctional kyphosis in adolescent idiopathic scoliosis—the pelvis and other considerations. *Spine Deformity* 5, 181–188. doi: 10.1016/j.jspd.2016.10.003

Mendoza-Lattes, S., Ries, Z., Gao, Y., and Weinstein, S. L. (2011). Proximal junctional kyphosis in adult reconstructive spine surgery results from incomplete restoration of the lumbar lordosis relative to the magnitude of the thoracic kyphosis. *Iowa Orthop. J.* 31, 199–206.

Mimura, T., Takahashi, J., Ikegami, S., Kuraishi, S., Shimizu, M., Futatsugi, T., et al. (2017). Can surgery for adolescent idiopathic scoliosis of less than 50 degrees of main thoracic curve achieve good results?. *J. Orthop. Sci.* 23, 14–19. doi: 10.1016/j.jos.2017.09.006

O'Shaughnessy, B. A., Bridwell, K. H., Lenke, L. G., Cho, W., Baldus, C., Chang, M. S., et al. (2012). Does a long-fusion "T3-sacrum" portend a worse outcome than a short-fusion "t10-sacrum" in primary surgery for adult scoliosis? *Spine* 37, 884–890. doi: 10.1097/brs.0b013e3182376414

Passias, P. G., Horn, S. R., Poorman, G. W., Daniels, A. H., Hamilton, D. K., Kim, H. J., et al. (2018). Clinical and radiographic presentation and treatment of patients with cervical deformity secondary to thoracolumbar proximal junctional kyphosis are distinct despite achieving similar outcomes: analysis of 123 prospective CD cases. *J. Clin. Neurosci.* 56, 121–126. doi: 10.1016/j.jocn.2018.06.040

Peng, H., Zhao, W., Tan, H., Ji, Z., Li, J., Li, K., et al. (2016). Prediction of treatment efficacy for prostate cancer using a mathematical model. *Sci. Rep.* 6:21599. doi: 10.1038/srep21599

Scheer, J. K., Osorio, J. A., Smith, J. S., Schwab, F., Lafage, V., Hart, R. A., et al. (2016). Development of validated computer-based preoperative predictive model for proximal junction failure (PJF) or clinically significant PJK With 86% accuracy based on 510 ASD patients with 2-year follow-up. *Spine* 41, E1328–E1335. doi: 10.1097/brs.0000000000001598

Sebaaly, A., Sylvestre, C., El Quehtani, Y., Riouallon, G., Larrieu, D., Boissiere, L., et al. (2018). Incidence and risk factors for proximal junctional kyphosis: results of a multicentric study of adult scoliosis. *Clin. Spine Surg.* 31, E178–E183. doi: 10.1097/bsd.0000000000000630

Suk, S. L., Lee, C. K., Kim, W. J., Chung, Y. J., and Park, Y. B. (1995). Segmental pedicle screw fixation in the treatment of thoracic idiopathic scoliosis. *Spine* 20, 1399–1405. doi: 10.1097/00007632-199520120-00012

Swami, A., and Jain, R. (2012). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Watanabe, K., Lenke, L. G., Bridwell, K. H., Kim, Y. J., Koester, L., and Hensley, M. (2010). Proximal junctional vertebral fracture in adults after spinal deformity surgery using pedicle screw constructs: analysis of morphological features. *Spine* 35, 138–145. doi: 10.1097/BRS.0b013e3181c8f35d

Yagi, M., Akilah, K. B., and Boachie-Adjei, O. (2011). Incidence, risk factors and classification of proximal junctional kyphosis: surgical outcomes review of adult idiopathic scoliosis. *Spine* 36, E60–E68. doi: 10.1097/BRS.0b013e3181eeaee2

Yang, S. P. (2003). Proximal kyphosis after short posterior fusion for thoracolumbar scoliosis. *Clin. Orthop. Relat. Res.* 411, 152–158. doi: 10.1097/01.blo.0000069885.72909.bb

Zhao, J., Yang, M., Yang, Y., Chen, Z., and Li, M. (2018). Proximal junctional kyphosis following correction surgery in the Lenke 5 adolescent idiopathic scoliosis patient. *J. Orthop. Sci.* 23, 744–749. doi: 10.1016/j.jos.2018.05.010

Zhao, X., Lang, R., Zhang, Z., Zhao, W., Ji, Z., Tan, H., et al. (2019). Exploring and validating the clinical risk factors for pancreatic cancer in chronic pancreatitis patients using electronic medical records datasets: three cohorts comprising 2,960 patients. *Transl. Cancer Res.* 9, 629–638. doi: 10.21037/tcr.2019.11.49

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fbioe.2020.607126

A Machine Learning Approach for Tracing Tumor Original Sites With Gene Expression Profiles

Xin Liang[1,2,3]*, Wen Zhu*[1,2,3]**, Bo Liao*[1,2,3]*, Bo Wang*[4,5]*, Jialiang Yang*[4,5]*, Xiaofei Mo*[4,5] *and Ruixi Li*[1,2,3]

[1] Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China, [2] Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, Haikou, China, [3] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [4] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [5] Geneis (Beijing) Co., Ltd., Beijing, China

Some carcinomas show that one or more metastatic sites appear with unknown origins. The identification of primary or metastatic tumor tissues is crucial for physicians to develop precise treatment plans for patients. With unknown primary origin sites, it is challenging to design specific plans for patients. Usually, those patients receive broad-spectrum chemotherapy, while still having poor prognosis though. Machine learning has been widely used and already achieved significant advantages in clinical practices. In this study, we classify and predict a large number of tumor samples with uncertain origins by applying the random forest and Naive Bayesian algorithms. We use the precision, recall, and other measurements to evaluate the performance of our approach. The results have showed that the prediction accuracy of this method was 90.4 for 7,713 samples. The accuracy was 80% for 20 metastatic tumors samples. In addition, the 10-fold cross-validation is used to evaluate the accuracy of classification, which reaches 91%.

Keywords: the ability of tissue tracing, random forest, naive Bayes, machine learning, uncertain origins

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Lei Wang,
Xiangtan University, China

****Correspondence:***
Wen Zhu
syzhuwen@163.com

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *16 September 2020*
Accepted: *26 October 2020*
Published: *24 November 2020*

Citation:
Liang X, Zhu W, Liao B, Wang B, Yang J, Mo X and Li R (2020) A Machine Learning Approach for Tracing Tumor Original Sites With Gene Expression Profiles. Front. Bioeng. Biotechnol. 8:607126. doi: 10.3389/fbioe.2020.607126

INTRODUCTION

Tumors can develop in any part of body, and some tumors even can metastasize to other parts of the body from their primary sites after developing at a certain point. In general, the occurrence of tumors at primary sites and their metastatic sites could be found deferentially, and the primary origins of metastatic cancers can be identified within a short amount of time by clinical assessments (Chen and Chen, 2001). Histological and imaging techniques are mostly employed to identify the origin of metastatic tumors. However, in some cancer patients, physicians cannot find the primary origin of tumors even after comprehensive examinations and assessment studies of patients with standard methods. These tumors are called carcinomas with unknown primary (CUP). According to statistical data, there are approximately 150,000 new cases of CUPs annually in the United States and Europe, and the numbers are still increasing though. Currently, approximately one third of cancer patients would develop metastasis after initial diagnosis and/or post-operation treatment. In many of those patients, it is relatively difficult for physicians to identify the primary origins of the metastatic cancers (Oien, 2009; Pavlidis and Pentheroudakis, 2012). To our knowledge, 2–4% of CUPs (Susman et al., 2012) account for all metastatic cancer. Even through autopsy, the primary origin of CUPs is uncertain (Myung et al., 2001; Petrushev et al., 2011). Because of limited treatment plan for CUP patients, the treatment efficacy is often unpredictable, and those patients usually have poor prognosis (Sun and Zhang, 2006; Gupta et al., 2007; Carmeliet and Jain, 2011; Petrakis et al., 2013). The immunohistochemistry assay is usually considered to be a diagnostic method for CUP patients. However, it is time-consuming and subjective. Moreover, the diagnostic accuracy is around by 30% for CUP patients, which is not reliable to design a personalized treatment plan for CUP patients. Currently, most CUP patients received radiological therapy (Stoyianni et al., 2011) or broad-spectrum chemotherapy. However, these treatments are not effective and with intolerable complications, and the prognosis is relatively poor as well. Therefore, it is urgent to develop effective clinical intervention for CUP patients (Guntinas-Lichius et al., 2006; Pavlidis and Fizazi, 2009; Hainsworth and Greco, 2014). Nowadays, identifying the primary origin of malignant tumors is critical for designing a treatment plan in clinical practices.

The targeted therapy (Tsao et al., 2005; Hudis, 2007; Miller et al., 2007; Varadhachary et al., 2008; Anderson and Weiss, 2010; Boscolo-Rizzo et al., 2015) can be used for tumors after accurately identifying the primary origin, which could greatly improve the survivals. It has been proven in the Minnie Pearl Cancer Research Network Study (Pavlidis and Pentheroudakis, 2010; Molina et al., 2012). Immunohistochemically, the marker has also been an important instrument for identifying the primary origin of cancerous tissues (Monzon et al., 2009; MacReady, 2010; Massard et al., 2011; Hashimoto et al., 2012; Oien and Dennis, 2012; Kim et al., 2013; Tang et al., 2018). Furthermore, a diagnostic method has been proposed to predict the primary origin of malignant cancers by comparing the gene expression profiles from the primary origin and the metastasis tissue (Hoadley et al., 2014). Many researchers have systematically compared the characteristics of gene expression profile across different cancers (Joyce and Pollard, 2009). Therefore, it is feasible to compare the differential gene expression to predict the primary origin of malignant cancer. There are two commercial products approved by FDA, which are Tissue of Origin (TOO) and CancerTYPE ID. Both of them are developed on the basis of differential gene expressions to predict primary origins.

TOO is a product of array-based gene expression profiles. TOO can identify 2,000 genes and 15 types of tumors, including thyroid cancer, breast cancer, non-small cell lung cancer, pancreatic cancer, gastric cancer, colorectal cancer, liver cancer, bladder cancer, kidney cancer, non-Hodgkin's lymphoma, melanoma, ovarian cancer, sarcoma, testicular germ cell tumor, and prostate cancer. The advantage of this product is that it prevents the subjective bias. It can objectively identify the primary origin of cancers no matter which is well-differentiated or not. However, TOO is time-consuming, which is not feasible for clinical practices (Brugarolas, 2007; Economopoulou et al., 2015).

CancerTYPE ID is a product that uses cancer samples based on RT-PCR data. In the study (Marquard et al., 2015), 578 labeled samples covering 39 tumor types were included in datasets for tracing origins. The results showed that there was no significant difference in the accuracy of predictions of cancer with primary or metastatic tumors. Secondly, RT-PCR was used to evaluate the 92-gene (Ma et al., 2005) expression of cancer cells from patients and then compared with labeled 50 tumors from databases to predict the primary origin of metastatic tumors and their subtypes (Pappa et al., 2006). CancerTYPE ID has been able to compare gene expression profiles from tumor samples to reference database with more than 2,000 labeled tumors, therefore identifying the most accurate results. However, CancerTYPE ID does not have the relatively good accuracy for pancreatic cancer, colorectal cancer, and gastroesophageal cancer.

Though the above two products have good performance for some types of cancers, two products are costly with up to \$3000–\$4000 (Pillai et al., 2011; Oien and Dennis, 2012; Economopoulou et al., 2015), and the accuracy is limited to other types of cancer as well. In order to facilitate the low-cost and high-efficiency product, our study aimed to use RNA-seq data, which are extracted from TCGA database, combining with random forest and naïve Bayes algorithms to develop a computational model.

RESULTS

Firstly, data were downloaded from TCGA and GEO. Secondly, after data preprocessing for raw data, genes were selected by the random forest algorithm with 10-fold cross-validation. Finally, the naive Bayes classifier was used to classify the 20 kinds of tumors, and the output of the model was shown as the evaluation index. The detailed step is shown in **Figure 1**.

Data Preparation

A total of 7,715 RNA-seq samples that covered 21 cancers and excluded metastatic cancers were extracted from TCGA. In the process of data preparation, we eliminated two samples due to the lack of clinical data. Therefore, the remaining 7,713 samples were used as either the training dataset or the validation dataset for the classification. Furthermore, the expression spectrum matrix of 7,633 samples was constructed. Each sample contained 20,501 genes. In this paper, 372 samples from metastatic cancers were selected as the test dataset, of which 352 samples belong to Skin Cutaneous Melanoma (SKCM). The ratio of SKCM was much higher than other types of metastatic cancers, and we excluded SKCM data from our selected data in order to reduce the possible effects on the results. The detailed information of selected data is shown in **Table 1**.

For the independent validation dataset, 48 samples are obtained from GEO and processed according to the description in section "Materials and Methods" and then used for the trained naive Bayesian model to make the prediction. The detailed information of selected data is shown in **Table 2**.

Gene Selection by Random Forests

Under the common condition, we use relatively low-cost panels but also include sufficient genes to determine the level of specific gene expression. However, the coverage of gene numbers would be significantly affected by the cost of panel. In order to reduce

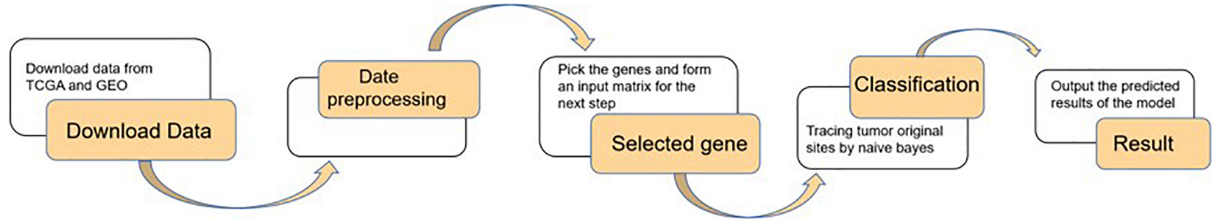


FIGURE 1 | Flow chart of the article.

the cost of panels as well as improve the accuracy of tracing ability. Random forest algorithm was employed widely in the bioinformatics researches (Lv et al., 2019, 2020; Ru et al., 2019).

TABLE 1 | Detailed information of data covering 21 cancers downloaded from TGCA.

Cancer	Total samples	Samples from women	Samples from men	Percentage (%)	Note
BLCA	301	80	221	3.9	
BRCA	1,056	1,044	11	13.7	1 person has no clinical information
CESC	258	258	0	3.3	
COAD	451	215	236	5.8	
GBM	153	53	100	2.0	
HNSC	480	128	352	6.2	
KIRC	526	184	342	6.8	
KIRP	222	63	159	2.9	
LAML	173	80	93	2.2	
LGG	439	192	247	5.7	
LIHC	294	99	195	3.8	
LUAD	486	262	224	6.3	
LUSC	428	109	319	5.5	
OV	261	261	0	3.4	
PAAD	142	64	78	1.8	
PRAD	379	0	379	4.9	
READ	153	70	82	2.0	1 person has no clinical information
SKCM	80	34	46	1.0	
STAD	415	147	268	5.4	
THCA	500	367	133	6.5	
UCEC	516	516	0	6.7	
Total	7,713	4,226	3,485	99.8	

TABLE 2 | Detailed information of data covering five cancers downloaded from GEO.

Cancer	Total samples	Percentage (%)
LIHC	9	18.75
UCEC	6	12.5
THCA	8	16.67
BLCA	11	22.92
PAAD	14	29.17
Total	48	99.98

In this study, the random forest algorithm was applied to select the features of the primary origin tumor samples, and a matrix of M*N was formed, with M representing the number of samples and N representing the numbers of genes, and all samples were labeled with the type of each cancer. The expression profile was divided into 20 types of cancer, and the combination of five genes could be used to classify this problem (Ashburner et al., 2000). The Gini average impurity method of random forest was used as the standard to evaluate the importance of genes. The importance score of genes was obtained, and the genes were sequenced according to the score. We conducted many experiments, and the precision was the highest when 2,300 genes were obtained. The experimental results are shown in **Figure 2**. Our method takes five steps and increases N up to 2,300.

Based on the above analysis, genes with high scores were selected as the features, and 2,300 genes were extracted from each sample. Because some genes were not in the GEO database, we deleted these genes and got 2,284 genes. A 7,633*2,284 matrix was constructed as the input matrix for cancer classification.

Classification Based on Naive Bayes

Since Naive Bayes is relatively consistent for classification, this study used Naive Bayes as a classifier for genomic combination. In this study, we chose 75% of the dataset for training, and the remaining 25% was chosen for validation by using our model. The algorithm used gene expression as the feature for training and predicting the labeled cancer. After the training, the model achieved the accuracy of 91% in predicting the origin of the cancer. In order to validate the accuracy of classification of model for metastatic tumors, 20 metastatic tumors with known primary origin sites were applied to the model. 7,633*2,284 was used as the input matrix for classification and applied to the naive Bayesian classification model to obtain the specific prediction results of specific cancer types with a prediction accuracy of 80% for metastatic cancer types, as shown in **Figure 3**.

In addition, ClueGO was used to identify gene ontology and enrichment analysis for selected genes. Due to the large number of 2,284 genes, we selected the top 100 genes with the highest score for analysis. The statistical significance level is set as the *p*-value of 0.001. The results of enrichment analysis are shown in **Figure 4**.

The enrichment results in **Figure 4** show that the genes are significantly enriched in cellular metabolism, especially

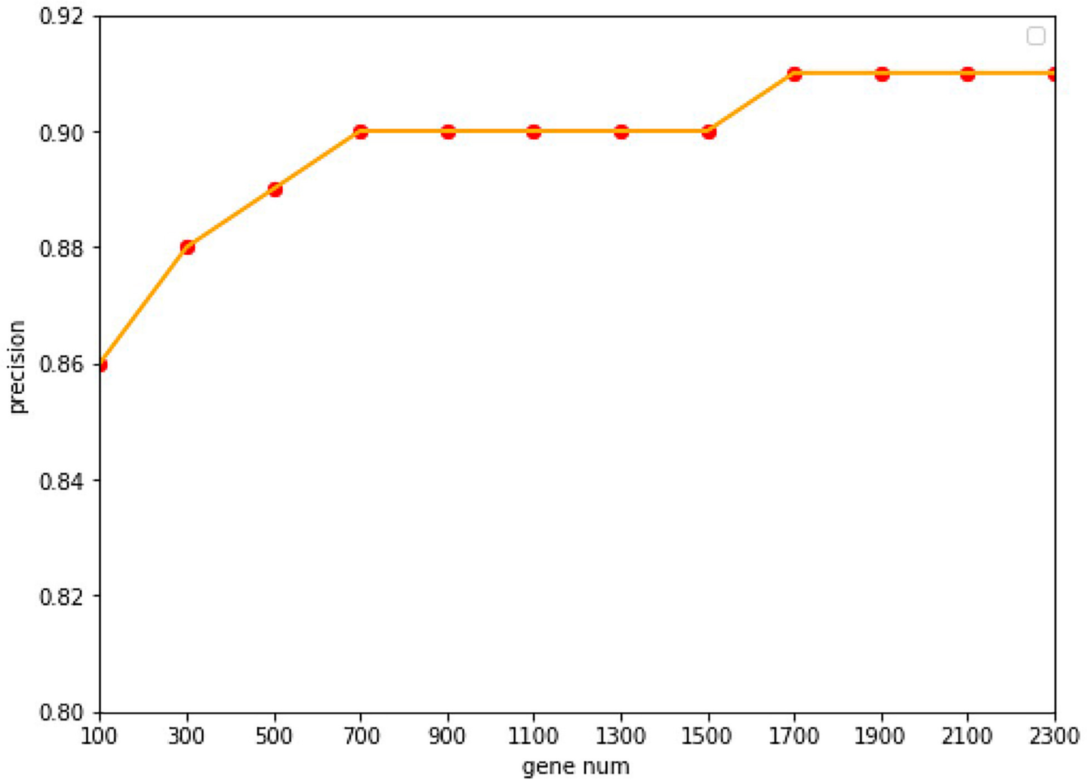
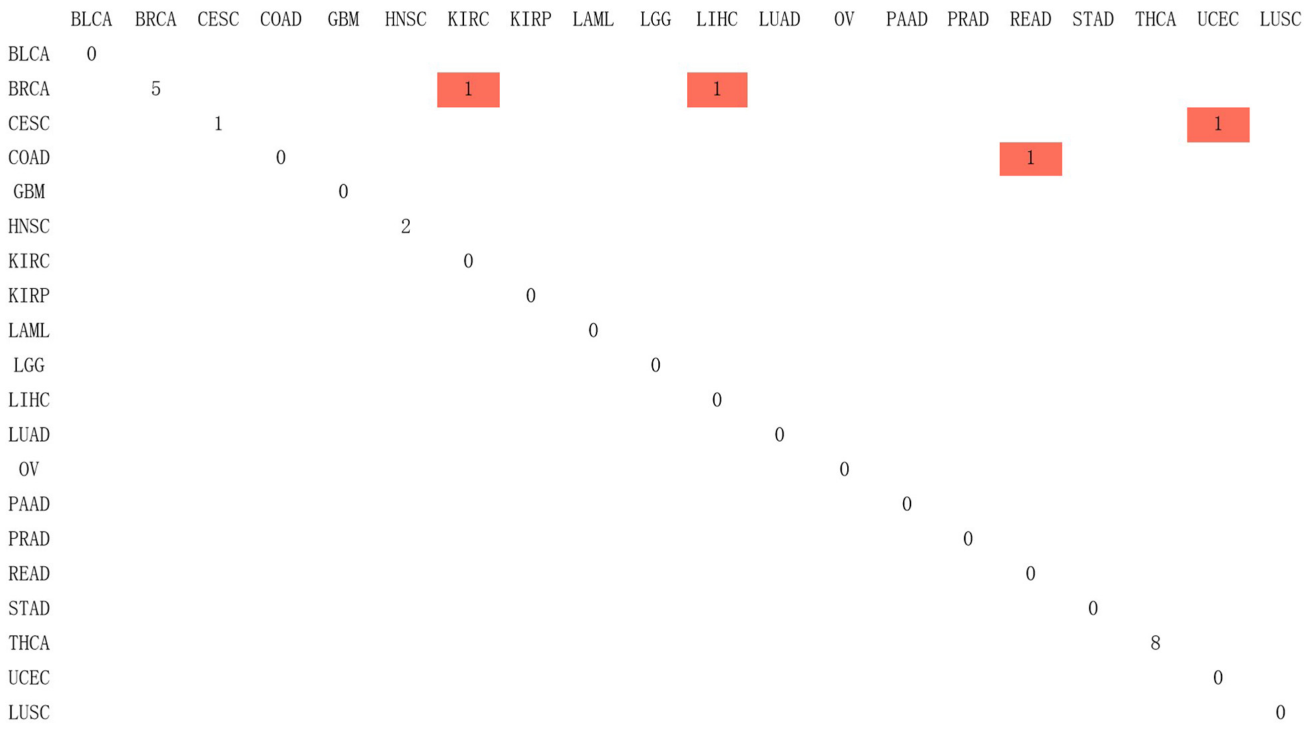


FIGURE 2 | The accuracy with the different of number genes. With a 10-fold cross-validation accuracy, the value of the accuracy is increasing up to 1,700 genes, after which it keeps stable with the value of 91.07%.



	BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LUAD	OV	PAAD	PRAD	READ	STAD	THCA	UCEC	LUSC
BLCA	0																			
BRCA		5					1				1									
CESC			1																1	
COAD				0												1				
GBM					0															
HNSC						2														
KIRC							0													
KIRP								0												
LAML									0											
LGG										0										
LIHC											0									
LUAD												0								
OV													0							
PAAD														0						
PRAD															0					
READ																0				
STAD																	0			
THCA																		8		
UCEC																			0	
LUSC																				0

FIGURE 3 | The confusion matrix of 2,284 genes in the classifier, in which red represented the result of inconsistency between primary and predicted cancer types.

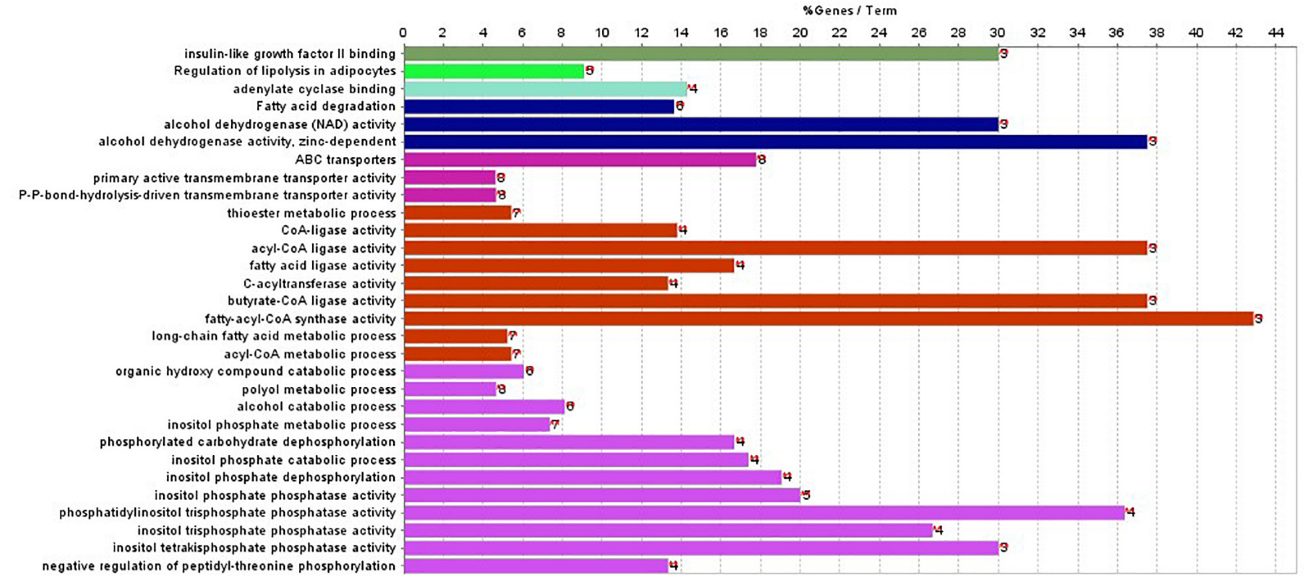


FIGURE 4 | Gene enriched in biological process, cellular component, and molecular function were drawn for first 100-gene set by ClueGO.

lipid metabolism. In addition, some genes are enriched in acetyl-CoA cycle, alcohol dehydrogenase NAD activity, etc. Almost all genes are enriched in lipid metabolism, which provides cellular energy for all cellular activity. Moreover, genes are also enriched through peroxisome proliferator-activated receptor (PPAR) signaling pathway. PPARs are nuclear hormone receptors activated by fatty acids and their derivatives and belong to ligand-activated receptors in the nuclear hormone receptor family. The PPAR signaling pathway plays a role in clearance of circulating lipid and promotes lipid oxidation and cell proliferation. The PPAR transcriptional activity can be regulated by non-gene crosstalks with phosphatases and kinases, including ERK1/2, p38-MAPK, PKC, and AMPK. The upregulated PPAR signaling pathway would lead to dysfunctional metabolic homeostasis and inflammatory response, ROS accumulation, as well as carcinogenesis across almost every tumor.

In order to further differentiate those 100 genes, the following heat map was drawn to further reveal the gene expression level in each cancer type.

The analysis shown in **Figure 5** reveals that there are expression differences of the first 100 genes in different cancers. Each small block represents a gene, and the color represents the size of gene expression. The higher level of the expression is represented with the darker color (red indicates upregulated and green indicates downregulated). The bottom horizontal line represents a different gene, while the vertical line on the right represents a different cancer.

Independent Verification

For independent tests, the model with the previous training parameters was tested on the dataset in GEO, and the probability of each sample being accurately assigned to each category was given, with an overall accuracy of 75%. The specific results are shown in **Figure 6**.

Performance Assessment

For the evaluation of classification performance, this study used the 10-fold cross-validation for the algorithm with the feature in each gene set. To be specific, the samples were randomly divided into 10 subsets; 1 of 10 subsets was selected as the test set at one time, and the other 9 was merging to 1 training set. The accuracy of cross-validation is 90%, which indicated that the algorithm achieved a good performance. The precision, recalls, and f1 scores were used to evaluate the significance of the model as well. The detailed results are shown in **Figure 7**.

The comparison among results of the *k*-nearest neighbor ($k = 5$), decision tree, and Naive Bayesian to classify 20 cancers is shown in **Figure 8**.

MATERIALS AND METHODS

Data Preparation

The TCGA RNA-seq and array data were downloaded from the ICGC Data Portal[1]. Each sample and each gene from each cancer type table were extracted to generate a matrix of M*N, where M is the number of samples, N is the number of genes, and all the samples were labeled by cancer types. All primary tumors were divided into training sets and metastatic tumors were divided into test sets.

For the independent set, 48 samples from 5 known cancer origin sites were downloaded from Gene Expression Omnibus (GEO). These tumors belong to GSE10907, GSE11222,

[1] https://dcc.icgc.org/releases/release_26/

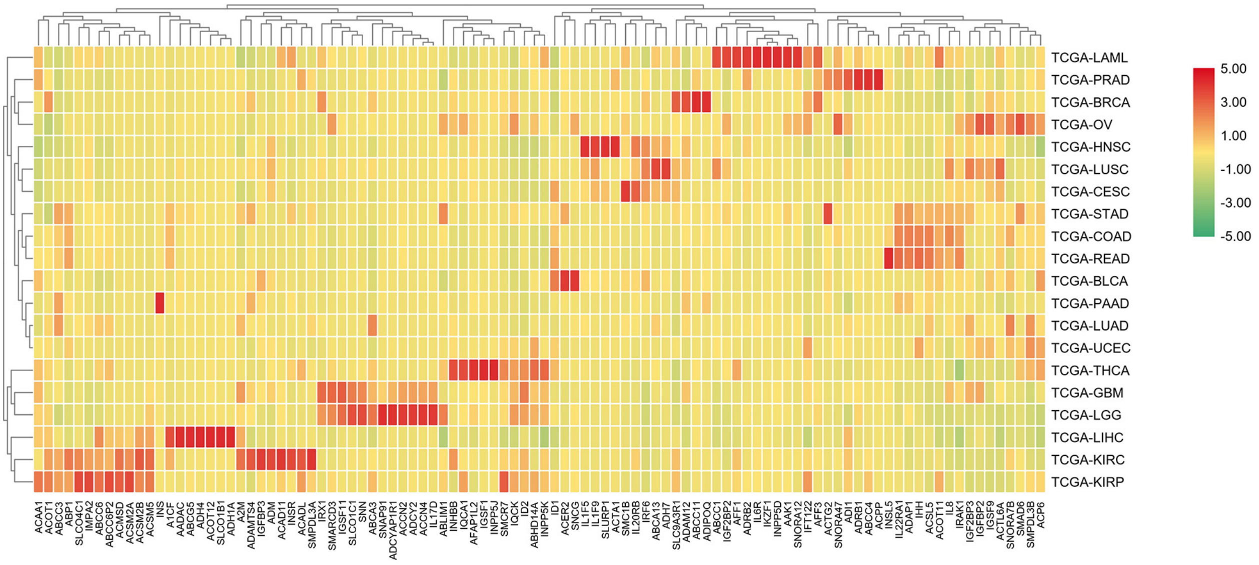


FIGURE 5 | A heat map of the first 100 genes was screened by the random forest algorithm. Where, row is cancer type, column is gene. In this part, RPKM is used to define the gene expression level, and the average value of samples in each cancer type is calculated as the gene expression difference.

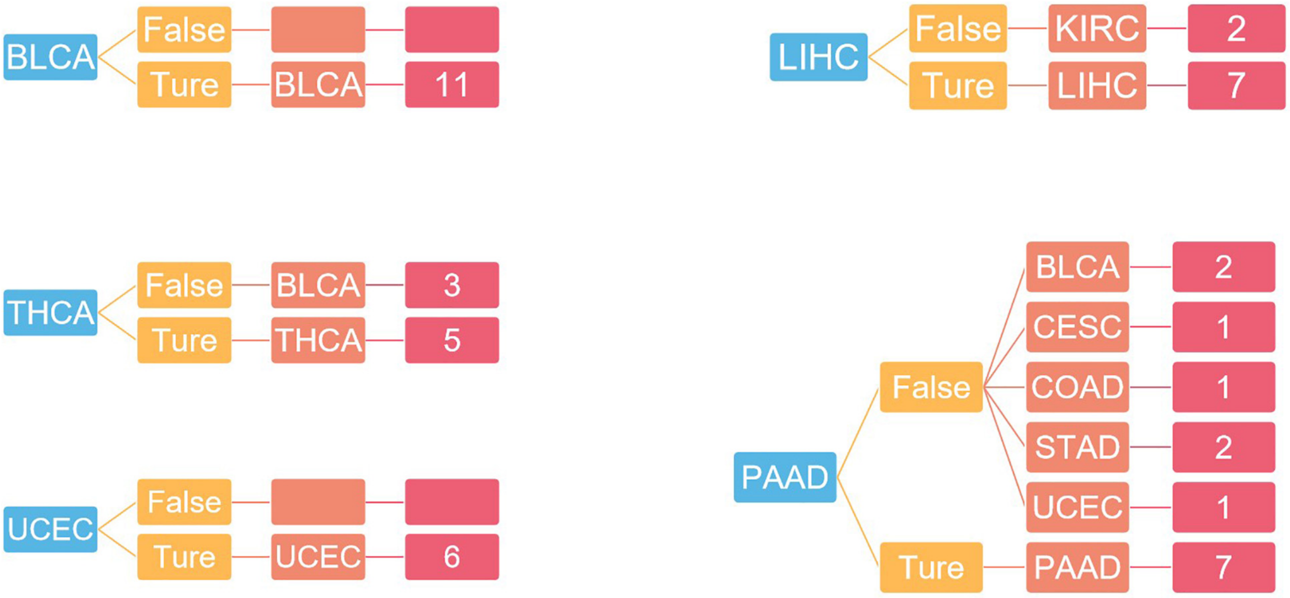


FIGURE 6 | The result of independent verification. Blue represents the primary tumor, orange represents the accuracy of the prediction, light red represents the predicted tumor type, and dark red represents the number of predicted tumor types.

GSE5608, GSE8352, GSE4895, GSE8912, GSE7966, and GSE12281. In addition, these 5 cancers belong to the 20 cancer types in this paper.

Gene Selection

In order to reduce the cost of gene number determined by gene panel, in this study, random forest algorithm was applied to select genes. The Gini average impurity in random forest was used as the criterion to estimate the importance of genes. The random forest is composed of several decision trees, which are binary decision trees. Each node in the decision tree is a condition on a single gene. As a result, we can achieve the goal by splitting the dataset into two datasets; therefore, a similar expression level can be classified in the same dataset. For random forest, the average reduction of each feature impurity can be calculated. In addition, the importance score of genes can be calculated and sorted according to the score. GI stands for Gini, S stands for importance score, $G = \{g_1, g_2, .., g_n\}$ stands for feature, and C stands

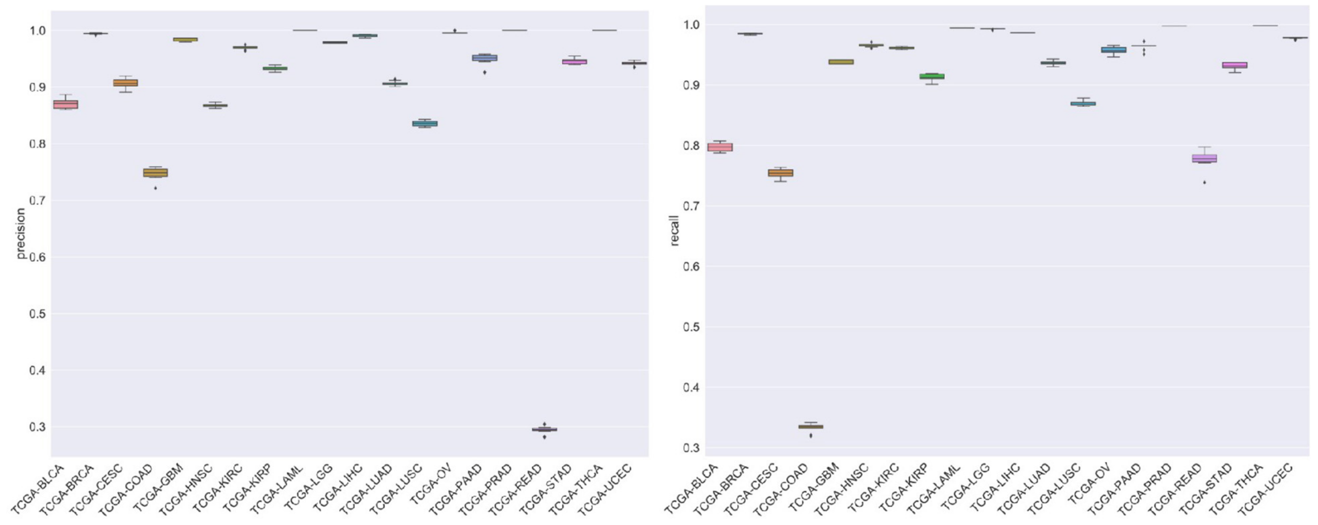
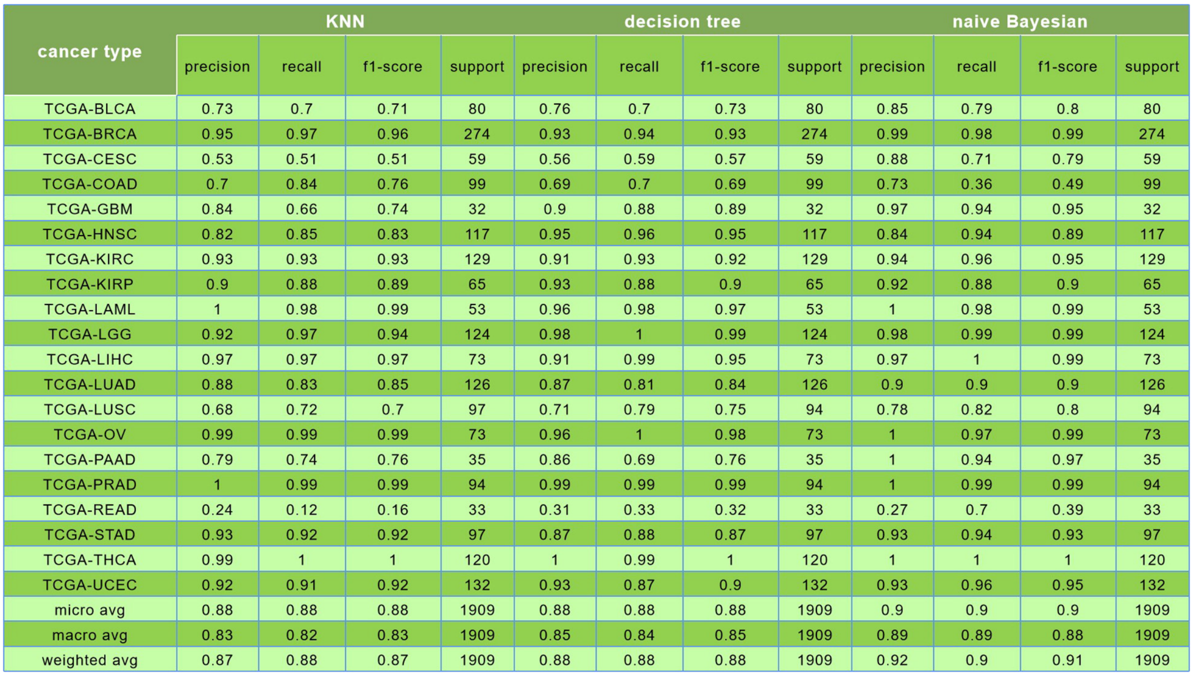


FIGURE 7 | The figure represented the recalls and precision after 10-fold cross-validation.



cancer type	KNN				decision tree				naive Bayesian			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
TCGA-BLCA	0.73	0.7	0.71	80	0.76	0.7	0.73	80	0.85	0.79	0.8	80
TCGA-BRCA	0.95	0.97	0.96	274	0.93	0.94	0.93	274	0.99	0.98	0.99	274
TCGA-CESC	0.53	0.51	0.51	59	0.56	0.59	0.57	59	0.88	0.71	0.79	59
TCGA-COAD	0.7	0.84	0.76	99	0.69	0.7	0.69	99	0.73	0.36	0.49	99
TCGA-GBM	0.84	0.66	0.74	32	0.9	0.88	0.89	32	0.97	0.94	0.95	32
TCGA-HNSC	0.82	0.85	0.83	117	0.95	0.96	0.95	117	0.84	0.94	0.89	117
TCGA-KIRC	0.93	0.93	0.93	129	0.91	0.93	0.92	129	0.94	0.96	0.95	129
TCGA-KIRP	0.9	0.88	0.89	65	0.93	0.88	0.9	65	0.92	0.88	0.9	65
TCGA-LAML	1	0.98	0.99	53	0.96	0.98	0.97	53	1	0.98	0.99	53
TCGA-LGG	0.92	0.97	0.94	124	0.98	1	0.99	124	0.98	0.99	0.99	124
TCGA-LIHC	0.97	0.97	0.97	73	0.91	0.99	0.95	73	0.97	1	0.99	73
TCGA-LUAD	0.88	0.83	0.85	126	0.87	0.81	0.84	126	0.9	0.9	0.9	126
TCGA-LUSC	0.68	0.72	0.7	97	0.71	0.79	0.75	94	0.78	0.82	0.8	94
TCGA-OV	0.99	0.99	0.99	73	0.96	1	0.98	73	1	0.97	0.99	73
TCGA-PAAD	0.79	0.74	0.76	35	0.86	0.69	0.76	35	1	0.94	0.97	35
TCGA-PRAD	1	0.99	0.99	94	0.99	0.99	0.99	94	1	0.99	0.99	94
TCGA-READ	0.24	0.12	0.16	33	0.31	0.33	0.32	33	0.27	0.7	0.39	33
TCGA-STAD	0.93	0.92	0.92	97	0.87	0.88	0.87	97	0.93	0.94	0.93	97
TCGA-THCA	0.99	1	1	120	1	0.99	1	120	1	1	1	120
TCGA-UCEC	0.92	0.91	0.92	132	0.93	0.87	0.9	132	0.93	0.96	0.95	132
micro avg	0.88	0.88	0.88	1909	0.88	0.88	0.88	1909	0.9	0.9	0.9	1909
macro avg	0.83	0.82	0.83	1909	0.85	0.84	0.85	1909	0.89	0.89	0.88	1909
weighted avg	0.87	0.88	0.87	1909	0.88	0.88	0.88	1909	0.92	0.9	0.91	1909

FIGURE 8 | In this figure, the first was the result of k-nearest neighbor (*k* = 5) algorithm, and its prediction accuracy was only 88%; the second was the result of decision tree algorithm, and the classification accuracy was only 88%; the third is the result of naive Bayesian algorithm, and the classification accuracy was reaching to 90%.

for cancer type. That is, to calculate the Gini score S_j for each feature g_j, the calculation formula of Gini index is as follows:

$$GI_m = 1 - \sum_{c=1}^{|C|} P_{mk}^2$$

where c represents C categories, and P_{mk} represents the proportion of category k in node m.

The importance of feature g_j in node m, that is, the variation of Gini impurities before and after node m branch, is calculated as follows:

$$S_{jm} = GI_m - GI_l - GI_r$$

where GI_l and GI_r, respectively, represent the Gini index of the two new nodes after branching, and S_{jm} represents the importance of feature g_j in node m.

If the node m with characteristic g_j that appears in decision tree i belongs to M, the importance of g_j in the ith tree is calculated as follows:

$$S_{ij} = \sum_{m \in M} S_{jm}$$

Assuming the random forest has t trees, the importance score formula of forest is:

$$S_j^* = \sum_{i=1}^{t} \sum_{m \in M} S_{jm}$$

The importance score is obtained by normalizing all the importance scores obtained:

$$S_j = \frac{S_j^*}{\sum_{i=1}^{n} S_i}$$

The top N genes with high scores were selected until the stopping criterion was met. Finally, the selected genes in all samples participated in the next classification.

Enrichment

Using the gene ontology (Bindea et al., 2009; Gene Ontology Consortium, 2019) as the database of enrichment analysis and annotating the function of specific gene sets to analyze their biological significance, ClueGO (Zhao et al., 2014) is used for visualization.

Classification

In this paper, naive Bayes was used as the classifier of gene combination. Naive Bayes is one of the classical machine learning algorithms. It is a classification algorithm based on Bayes theorem. Its principle is simple and easy to implement. The core idea of naive Bayesian algorithm is to assume that each feature is independent. For a given type of data to be judged, classify and predict according to the training dataset, and calculate the probability that the current type of data to be judged belongs to a certain category through Bayesian theorem. The maximum probability relationship obtained is that the algorithm judges the category of these data. Naive Bayesian algorithm can be divided into three parts:

First, determine the feature attributes; that is to say, the expression profiles of 2,284 genes corresponding to each sample were extracted. Then, it was assumed that all the features conformed to the Gauss distribution. The samples in the dataset were labeled as cancer type. G represents the characteristics and C represents the type of cancer, which can be calculated as the prior probability P(C). C_k represents the kth category, g_i represents the ith feature, and then calculate conditional probability by prior probability. The formula is as follows:

$$P(G|C_k) = P(G_1 = g_1, G_2 = g_2, \cdots, G_n = g_n|C_k)$$

The conditional probability of all the kth classes is calculated by the Bayesian formula:

$$P(C_k|G) = \frac{P(G|C_k)P(C_k)}{P(G)} = P(C_k)\prod_{i=1}^{n} P(g_i|C_k)$$

Since all the features conform to the Gaussian distribution and are independent of each other, the formula for conditional probability becomes as follows:

$$P(G|C_k) = \prod_{i=1}^{n} P(g_i|C_k) = \prod_{i=1}^{n} P(g_i|\mu_{i,C_k}, \sigma_{i,C_k})$$
$$= \frac{1}{\sigma_{i,C_k}\sqrt{2\pi}} \exp\left\{-\frac{(g_i - \mu_{i,C_k})^2}{2\sigma_{i,C_k}^2}\right\}$$

where g_i is the ith feature, and $\mu_{i,ck}$ and $\sigma_{i,ck}$ are the mean and variance of the ith feature in the K class C_k, respectively.

The conditional probability formulas for all the Kth class are calculated as follows:

$$P(C_k|G) = \frac{P(G|C_k)P(C_k)}{\sum_k P(G|C_k)P(C_k)} \propto P(G|C_k)P(C_k)$$

Finally, obtain the relationship between the maximum probability data to be classified and the category, $P(C_k| G)$, that is:

$$y = \mathrm{argmax}_{C_k} P(G|C_k)P(C_k)$$

It is meaningful to indicate that we could get the most probable type of cancer under certain gene expressions.

DISCUSSION

In **Figure 3**, 20 known primary tumors were predicted, while 4 of them were misjudged, which may be related to the naive Bayesian algorithm. Naive Bayes is one of the few algorithms based on probability theory, which is a very simple and convenient algorithm. However, the premise of this algorithm is to assume that each feature is independent of others, which is not in line with the reality. Therefore, it may produce errors in the classification results, leading to the decline of the prediction accuracy. In addition, in **Figure 3**, COAD was mislabeled as READ. It was possibly because the anatomical proximity is relatively close and may share differential gene expression. During the normal digestive process, the function of colon and rectum is not significantly different, while colon may contribute to maintaining the gut microenvironment. The epithelial cells that are usually changed in colon adenocarcinoma and rectum adenocarcinoma are not well-distinguished. It may possibly increase both the subjective and objective bias of our model. One case of CESC was misdiagnosed as UCES. Those two female malignant tumors are more commonly regulated by the female hormone, which share similar risk factors. The anatomical proximity is close as well. The above cases indicated that anatomical proximity may share oncogenic genes to drive genetic mutation, such as both cancers contain KRAS mutations (Gene Ontology Consortium, 2019), or it is difficult to differentiate epithelial or adrenal cell changes before oncogenesis. It is critical to point out that some biological factors might bring some effect for model performance. It is necessary to be considered as the model construction. In addition, there are only 20 cases of known primary tumor data used to predict the classification. The data size is relatively small, so we cannot get a certain conclusion. We need to further expand the database for classification and prediction.

In **Figure 4**, the first 100 genes with the highest score are selected by the random forest algorithm. Some genes obtained by this method may have high correlation; that is to say, these genes will provide the same information for the classifier. In addition, although we used the 10-fold cross-validation to investigate the performance of the model, in the independent validation, the accuracy of this model is only 75%. The predictive error to PAAD is large, and the independent validation dataset is small.

CONCLUSION

In this study, the random forest and naive Bayesian algorithms were employed to trace the origin of CUP sites. Through a large number of experiments, we found that 2,284 genes with the highest score achieved the best performance. Performance evaluation shows that this method can achieve good classification and prediction results. In addition, ClueGO enrichment analysis was used for the top 100 genes with the highest scores. The results showed that some genes were enriched in PPAR signaling pathway. Upregulation of PPAR signaling pathway has been proven to lead to metabolic homeostasis disorder, inflammation, ROS accumulation, and carcinogenesis. In summary, the proposed approach can reduce the cost and has high efficiency, and thus it is promising for clinical practices.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/releases/release_26/.

AUTHOR CONTRIBUTIONS

JY, BL, and WZ conceived the project. XL and BW implemented the experiments and analyzed the data. XL, XM, and RL prepared the data and performed literature search. XL and JY wrote the manuscript. All authors approved the final manuscript.

FUNDING

This work was supported by the National Nature Science Foundation of China (Grant Nos. 11926412, 11926205, 61863010, and 61873076), the Natural Science Foundation of Hainan, China (Grant No. 119MS036), and Innovative scientific research projects for graduate students in Hainan Province (Grant No. Hys2019-266).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.607126/full#supplementary-material

REFERENCES

Anderson, G. G., and Weiss, L. M. (2010). Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. *Appl. Immunohistochem. Mol. Morphol.* 18, 3–8. doi: 10.1097/pai.0b013e3181a75e6d

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Boscolo-Rizzo, P., Schroeder, L., Romeo, S., and Pawlita, M. (2015). The prevalence of human papillomavirus in squamous cell carcinoma of unknown primary site metastatic to neck lymph nodes: a systematic review. *Clin. Exp. Metast.* 32, 835–845. doi: 10.1007/s10585-015-9744-z

Brugarolas, J. (2007). Renal-cellcarcinoma: molecularpathways and therapies. *N. Engl. J. Med.* 356, 185–187.

Carmeliet, P., and Jain, R. K. (2011). Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases. *Nat. Rev. Drug Discov.* 10, 417–427. doi: 10.1038/nrd3455

Chen, L. M., and Chen, B.-S. (2001). A robust adaptive DFE receiver for DS-CDMA systems under multipath fading channels. *IEEE Trans. Signal Process.* 49, 1523–1532. doi: 10.1109/78.928705

Economopoulou, P., Mountzios, G., Pavlidis, N., and Pentheroudakis, G. (2015). Cancer of unknown primary origin in the genomic era: elucidating the dark box of cancer. *Cancer Treat. Rev.* 41, 598–604. doi: 10.1016/j.ctrv.2015.05.010

Gene Ontology Consortium (2019). The Gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Guntinas-Lichius, O., Peter Klussmann, J., Dinsh, S., Dinh, M., Schmidt, M., Semrau, R., et al. (2006). Diagnostic work-up and outcome of cervical metastases from an unknown primary. *Acta Otolaryngol.* 126, 536–544. doi: 10.1080/00016480500417304

Gupta, G. P., Perk, J., Acharyya, S., de Candia, P., Mittal, V., Todorova-Manova, K., et al. (2007). ID genes mediate tumor reinitiation during breast cancer lung metastasis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19506–19511. doi: 10.1073/pnas.0709185104

Hainsworth, J. D., and Greco, F. A. (2014). Gene expression profiling in patients with carcinoma of unknown primary site: from translational research to standard of care. *Virchows Arch.* 464, 393–402. doi: 10.1007/s00428-014-1545-2

Hashimoto, K., Sasajima, Y., Ando, M., Yonemori, K., Hirakawa, A., Furuta, K., et al. (2012). Immunohistochemical profile for unknown primary adenocarcinoma. *PLoS One* 7:e31181. doi: 10.1371/journal.pone.0031181

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classificationwithin and across tissues of origin. *Cell* 158, 929–944. doi: 10.1016/j.cell.2014.06.049

Hudis, C. A. (2007). Trastuzumab: mechanism of action and use in clinical practice. *N. Engl. J. Med.* 357, 39–51. doi: 10.1056/nejmra043186

Joyce, J. A., and Pollard, J. W. (2009). Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* 9, 239–252. doi: 10.1038/nrc2618

Kim, K. W., Krajewski, K. M., Jagannathan, J. P., Nishino, M., Shinagare, A. B., Hornick, J. L., et al. (2013). Cancer of unknown primary sites: what radiologists need to know and what oncologists want to know. *AJR Am. J. Roentgenol.* 200, 484–492. doi: 10.2214/ajr.12.9363

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Lv, Z., Zhang, J., Ding, H., and Zou, Q. (2020). RF-PseU: a random forest predictor for RNA Pseudouridine sites. *Front. Bioeng. Biotechnol.* 8:134. doi: 10.3389/fbioe.2020.00134

Ma, X. J., Patel, R., Wang, X., Salunga, R., Murage, J., Desai, R., et al. (2005). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch. Pathol. Lab. Med.* 130, 465–473.

MacReady, N. (2010). NICE issues guidance on cancer of unknown primary. *Lancet Oncol.* 11:824. doi: 10.1016/s1470-2045(10)70215-1

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2015). Tumor tracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med. Genom.* 8:58. doi: 10.1186/s12920-015-0130-0

Massard, C., Loriot, Y., and Fizazi, K. (2011). Carcinomas of an unknown primary origin–diagnosis and treatment. *Nat. Rev. Clin. Oncol.* 8, 701–710. doi: 10.1038/nrclinonc.2011.158

Miller, K., Wang, M., Gralow, J., Dickler, M., Cobleigh, M., Perez, E. A., et al. (2007). Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *N. Engl. J. Med.* 357, 2666–2676. doi: 10.1056/nejmoa072113

Molina, R., Bosch, X., Auge, J. M., Filella, X., Escudero, J. M., Molina, V., et al. (2012). Utility of serum tumor markers as an aid in the differential diagnosis of patients with clinical suspicion of cancer and in patients with cancer of unknown primary site. *Tumour Biol.* 33, 463–474. doi: 10.1007/s13277-011-0275-1

Monzon, F. A., Lyons-Weiler, M., Buturovic, L. J., Rigl, C. T., Henner, W. D., Sciulli, C., et al. (2009). Multicenter validation of a 1500-gene expression profile for identification of tumor tissue of origin. *J. Clin. Oncol.* 27, 2503–2508. doi: 10.1200/jco.2008.17.9762

Myung, J., Kim, K. B., Lindsten, K., Dantuma, N. P., and Crews, C. M. (2001). Lak of proteasome active site allostery as revealed by subunit-specific inhibitors. *Mol. Cell* 7, 411–420. doi: 10.1016/s1097-2765(01)00188-5

Oien, K. A. (2009). Pathologic evolution of unknown primary cancer. *Semin. Oncol.* 36, 8–37. doi: 10.1053/j.seminoncol.2008.10.009

Oien, K. A., and Dennis, J. L. (2012). Diagnostic work-up of carcinoma of unknown primary: from IHC to molecular profiling. *Ann. Oncol.* 23(Suppl. 10), x271–x277.

Pappa, K. I., Choleza, M., Markaki, S., Giannikaki, E., Kyroudi, A., Vlachos, G., et al. (2006). Consistent absence of BRAF mutations in cervical and endometrial cancer despite KRAS mutation status. *J. Gynecol. Oncol.* 100, 596–600. doi: 10.1016/j.ygyno.2005.09.029

Pavlidis, N., and Fizazi, K. (2009). Carcinoma of unknown primary(CUP). *Crit. Rev. Oncol. Hematol.* 69, 271–278. doi: 10.1016/j.critrevonc.2008.09.005

Pavlidis, N., and Pentheroudakis, G. (2010). Cancer of unknown primary site: 20 questions to be answered. *Ann. Oncol.* 21(Suppl. 7), vii303–vii307.

Pavlidis, N., and Pentheroudakis, G. (2012). Cancer of unknown primary site. *Lancet* 379, 1428–1435.

Petrakis, D., Pentheroudakis, G., Voulgaris, E., and Pavlidis, N. (2013). Prognostication in cancer of unknown primary (CUP): development of a prognostic algorithm in 311 cases and review of the literature. *Cancer Treat. Rev.* 39, 701–708. doi: 10.1016/j.ctrv.2013.03.001

Petrushev, B., Tomuleasa, C., Susman, S., Sorişău, O., Aldea, M., Kacsó, G., et al. (2011). The aixs of evil in the fight against cancer. *Rom. J. Intern. Med.* 49, 319–325.

Pillai, R., Deeter, R., Rigl, C. T., Nystrom, J. S., Miller, M. H., Buturovic, L., et al. (2011). Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. *J. Mol. Diagn.* 13, 48–56. doi: 10.1016/j.jmoldx.2010.11.001

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteom. Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Stoyianni, A., Pentheroudakis, G., and Pavlidis, N. (2011). Neuroendocrine carcinoma of unknown primary: a systematic review of the literature and a comparative study with other neuroendocrine tumors. *Cancer Treat. Rev.* 37, 358–365. doi: 10.1016/j.ctrv.2011.03.002

Sun, X. F., and Zhang, H. (2006). Clinicopathological significance of stromal variables: angiogenesis, lymphangiogenesis, inflammatory infiltration, MMP and PINCH in colorectal carcinomas. *Mol. Cancer* 5:43.

Susman, S., Tomuleasa, C., Soritau, O., Mihu, C., Rus-Ciuca, D., Sabourin, J. C., et al. (2012). The colorectal cancer stem-like cell hypothesis: a pathologist's point of view. *J. BUON* 17, 230–236.

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tsao, M. S., Sakurada, A., Cutz, J. C., Zhu, C. Q., Kamel-Reid, S., Squire, J., et al. (2005). Erlotinib in lung cancer: molecular and clinical predictors of outcome. *N. Engl. J. Med.* 353, 133–144.

Varadhachary, G. R., Raber, M. N., Matamoros, A., and Abbruzzese, J. L. (2008). Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions. *Lancet Oncol.* 9, 596–599. doi: 10.1016/s1470-2045(08)70151-7

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2014). Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteom.* 11, 289–299. doi: 10.2174/157016461104150121115154

Conflict of Interest: BW, JY, and XM were employed by company Geneis (Beijing) Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 23 December 2020
doi: 10.3389/fbioe.2020.605132

RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans

*Qiangguo Jin[1,2], Zhaopeng Meng[1,3], Changming Sun[2], Hui Cui[4] and Ran Su[1]**

[1] School of Computer Software, College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] CSIRO Data61, Sydney, NSW, Australia, [3] Tianjin University of Traditional Chinese Medicine, Tianjin, China, [4] Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia

Automatic extraction of liver and tumor from CT volumes is a challenging task due to their heterogeneous and diffusive shapes. Recently, 2D deep convolutional neural networks have become popular in medical image segmentation tasks because of the utilization of large labeled datasets to learn hierarchical features. However, few studies investigate 3D networks for liver tumor segmentation. In this paper, we propose a 3D hybrid residual attention-aware segmentation method, i.e., RA-UNet, to precisely extract the liver region and segment tumors from the liver. The proposed network has a basic architecture as U-Net which extracts contextual information combining low-level feature maps with high-level ones. Attention residual modules are integrated so that the attention-aware features change adaptively. This is the first work that an attention residual mechanism is used to segment tumors from 3D medical volumetric images. We evaluated our framework on the public MICCAI 2017 Liver Tumor Segmentation dataset and tested the generalization on the 3DIRCADb dataset. The experiments show that our architecture obtains competitive results.

Keywords: medical image segmentation, tumor segmentation, u-net, residual learning, attention mechanism

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Lin Gu,
National Institute of Informatics, Japan
Xiangxiang Zeng,
Hunan University, China

****Correspondence:***
Ran Su
ran.su@tju.edu.cn

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

***Received:** 11 September 2020*
***Accepted:** 01 December 2020*
***Published:** 23 December 2020*

Citation:
Jin Q, Meng Z, Sun C, Cui H and Su R (2020) RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. Front. Bioeng. Biotechnol. 8:605132. doi: 10.3389/fbioe.2020.605132

1. INTRODUCTION

Liver tumors, or hepatic tumors, are great threats to human health. The malignant tumor, also known as the liver cancer, is one of the most frequent internal malignancies worldwide (6%), and is also one of the leading death causes from cancer (9%) (WHO, 2014a,b). Even the benign (non-cancerous) tumors may grow large enough to cause health problems. Computed tomography (CT) is used to assist the diagnosis of liver tumors (Christ et al., 2017a). The extraction of liver and tumors from CT is a critical task before surgical intervention in choosing an optimal approach for treatment. Accurate segmentation of liver and tumor from medical images provides their precise locations in the human body. Then therapies evaluated by the specialists can be provided to treat individual patients (Rajagopal and Subbaiah, 2015). However, due to the heterogeneous and diffusive shapes of liver and tumor, segmenting them from CT images is challenging. Numerous efforts have been taken to tackle the segmentation task on liver/tumors. **Figure 1** shows some typical liver and tumor CT scans.

In general, liver and tumor extraction approaches can be classified into three categories: manual segmentation, semi-automated segmentation, and automated segmentation. Manual segmentation is a subjective, poorly reproducible, and time-consuming approach. It heavily depends upon human recognizable features, and requires people with high-level technical skills.

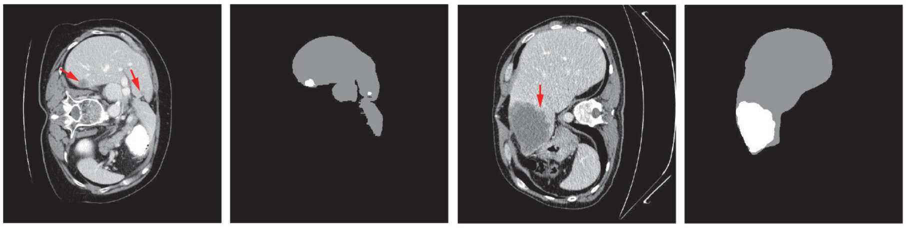


FIGURE 1 | Examples of typical 2D CT scans and the corresponding ground truth of liver/tumor extractions where red arrows indicate the tumor/lesion regions. The typical scans are from the MICCAI 2017 Liver Tumor Segmentation (LiTS) dataset.

These factors make it impractical for real applications (Li et al., 2015). Semi-automated segmentation requires initial human intervention, which may cause bias and mistakes. In order to accelerate and facilitate diagnosis, therapy planning, and monitoring, and finally help surgeons remove tumors, it is necessary to develop an automated and precise method to segment tumors from CT images. However, the large scale spatial and structural variability, low contrast between liver and tumor regions, existence of noise, partial volume effects, complexity of 3D-spatial tumor features, or even the similarity between nearby organs make the automation of segmentation quite a difficult task (Li et al., 2015). Recently, convolutional neural networks (CNN) have been applied to many volumetric image segmentations. A number of CNN models including both 2D and 3D networks have been developed. However, the 3D networks are usually not as efficient and flexible as the corresponding 2D networks. For instance, 2D and 3D fully convolutional networks (FCNs) have been proposed for semantic segmentation (Long et al., 2015). Yet due to the high computational cost and the low efficiency of 3D convolutions, the depth of the 3D FCNs is limited compared to that of 2D FCNs, which makes it impractical for 2D networks to be extended to 3D networks.

To address these issues and inspired by the residual networks (He et al., 2016) and the attention residual learning (Wang et al., 2017), we propose a hybrid residual attention-aware liver and tumor extraction neural network named RA-UNet[1], which is designed to effectively extract 3D volumetric contextual features of liver and tumor from CT images in an end-to-end manner. The proposed network integrates a residual U-Net architecture and an attention residual learning mechanism which enables the optimization and performance improvement on deep networks. The contributions of our works are listed as follows: Firstly, the attention mechanism can have the capability of focusing on specific parts of the image. Different types of attention are possible through stacking attention modules so that the attention-aware features can change adaptively. Secondly, we use the 3D U-Net as the basic architecture to capture multi-scale attention information and to integrate low-level features with high-level ones. Besides, RA-UNet, which directly segments the liver and tumor from 3D medical volumes, enlarges the U-Net family in 3D medical image analysis. What's more, our model does not depend on any pre-trained model or commonly used post processing techniques, such as 3D conditional random fields. The generalization of the proposed approach is demonstrated through testing on the 3DIRCADb dataset (Soler et al., 2010). Our architecture achieves competitive performances comparing with other state-of-the-art methods on the MICCAI 2017 Liver Tumor Segmentation (LiTS) dataset, and also shows high generalization. Our paper is organized as follows. In section 2, we briefly review the state-of-the-art automated liver tumor segmentation methods. We illustrate the methodologies in detail including the datasets, preprocessing strategy, hybrid deep learning architecture, and training procedure in section 3. In section 4, we evaluate the proposed algorithm, report the experimental results, compare with some other approaches, and extend our approach to other medical segmentation tasks. Conclusions and future works are given in section 5.

2. RELATED WORKS

In the past decades, various applications have been developed via computer-aided methods in medical/biomedical image processing, cellular biology domains (Zeng et al., 2017; Hong et al., 2020a,b; Song et al., 2020a,b, 2021). Recently, with the advance of artificial intelligence, deep learning has been used in a number of areas such as natural language processing, anti-cancer drug response prediction, and image analysis (Liu et al., 2017; Su et al., 2019; Zeng et al., 2020). Some have achieved state-of-the-art performances in medical imaging challenges (Litjens et al., 2017; Jin et al., 2019).

2.1. Deep Learning in Medical Image Analysis

Unlike the traditional methods that use hand-crafted features, deep neural networks (DNNs) are able to automatically learn discriminative features. The learned features which contain hierarchical information have the ability to represent each level of the input data. Among those methods, CNN is one of the most popular methods and has shown impressive performance for 3D medical image analysis tasks. Multi-scale patch-based and pixel-based strategies were proposed to improve the segmentation performance. For instance, Zhang et al. (2015) proposed a

[1] https://github.com/RanSuLab/RAUNet-tumor-segmentation.git

method which used a deep CNN for segmenting brain tissues using multi-modality magnetic resonance images (MRI). Li et al. (2015) presented an automatic method based on 2D CNN to segment lesions from CT slices and compared the CNN model with other traditional machine learning techniques, which included AdaBoost (Collins et al., 2002), random forests (RF) (Breiman, 2001), and support vector machine (SVM) (Furey et al., 2000). This study showed that CNN still had limitations on segmenting tumors with uneven densities and unclear borders. Pereira et al. (2016) proposed a CNN architecture with small kernels for segmenting brain tumors on MRI. This architecture reached Dice similarity coefficient metrics of 0.78, 0.65, and 0.75 for the complete, core, and enhancing regions respectively. Lee et al. (2011) presented a CNN-based architecture that could learn from provided labels to construct brain segmentation features. However, due to low memory requirements, low complexity of computation, and lots of pre-trained models, most of the latest CNN architectures including the methods reviewed above used 2D slices from 3D volumes for carrying out the segmentation task. However, the spatial structural organizations of organs are not considered, and the volumetric information is not fully utilized. Therefore, 3D automatic segmentation which makes full use of spatial information is urgently needed for surgeons.

2.2. 3D Convolutional Neural Networks

In order to sufficiently add 3D spatial structures into CNN for 3D medical image analysis, 3D CNN which considers axial direction of the 3D volumes has recently been proposed in medical imaging field. Shakeri et al. (2016) proposed a 2D CNN architecture to detect tumors from a set of brain slices. Then they additionally applied a 3D conditional random field (CRF) algorithm for post processing in order to impose volumetric homogeneity. This is one of the earliest studies that used CNN-related segmentation on volumetric images. Çiçek et al. (2016) learned from sparsely sequential volumetric images by feeding a U-Net with 2D sequential slices. 3D CNN-based segmentation methods were then employed in a large scale. Andermatt et al. (2016) used a 3D recurrent neural network (RNN) with gated recurrent units to segment gray and white matters in a brain MRI dataset. Dolz et al. (2017) investigated a 3D FCN for subcortical brain structure segmentation in MRI images. They reduced the computational and memory costs, which were quite severe issues for 3D CNN, via small kernels with a deeper network. Bui et al. (2017) proposed a deep densely convolutional network for volumetric brain segmentation. This architecture provided a dense connection between layers. They concatenated feature maps from fine and coarse blocks, which allowed to capture multi-scale contextual information. The 3D deeply supervised network (DSN), which had a much faster convergence and better discrimination capability, could be extended to other medical applications (Dou et al., 2016). Oktay et al. (2018) proposed a novel attention gate model called attention U-Net for medical imaging which could learn to concentrate on target structures of different shapes and sizes. However, due to hardware limitations, 3D convolutional medical image segmentation is still a bottleneck.

2.3. Liver Tumor Segmentation

As for liver tumor detection in 3D volumetric images, not many explorations have been made using the CNN-based methods. Lu et al. proposed a method based on 3D CNN to carry out the probabilistic segmentation task and used graph cut to refine the previous segmentation result. However, as tested only on one dataset, the generality of this architecture still needs to be validated (Lu et al., 2017). Christ et al. (2017a) proposed a cascaded FCNs (CFCNs) to segment liver and its lesions in CT and MRI images, which enabled segmentation for large scale medical trials. They trained the first FCN to segment the liver and trained the second FCN to segment its lesions based on the predicted liver region of interest (ROI). This approach reached a Dice score of 94%. Additionally, Christ et al. (2017b) also predicted hepatocellular carcinoma (HCC) malignancy using two CNN architectures. They took a CFCN as the first step to segment tumor lesions. Then they applied a 3D neural network called SurvivalNet to predict the lesions' malignancy. This method achieved an accuracy of 65% with a Dice score of 69% for lesion segmentation and an accuracy of 68% for tumor malignancy detection. Kaluva et al. (2018) proposed a fully automatic 2-stage cascaded method for liver and tumor segmentation based on the LiTS dataset, and they reached global Dice scores of 0.923 and 0.623 on liver and tumor, respectively. Bi et al. (2017) integrated 2D residual blocks into their network and gained a Dice score of 0.959. Moreover, Li et al. (2018) built a hybrid densely connected U-Net for liver and tumor segmentation, which combined both 2D and 3D features on liver and tumor. They reached Dice scores of 0.961 and 0.722 on liver and tumor segmentation, respectively. Pandey et al. (2018) reduced the complexity of a deep neural network by introducing ResNet-blocks and obtained a Dice score of 0.587 on tumor segmentation. Recently, Tang et al. (2020) proposed a two-stage framework for 2D liver and tumor segmentation. The proposed network explicitly captured complementary objects (liver and tumor) and their edge information to preserve the organ and lesion boundaries. Heker and Greenspan (2020) introduced transfer learning and joint learning to improve the network's generalization and robustness for liver lesion segmentation and classification. Seo et al. (2019) modified the U-Net with Object-Dependent high-level features for the liver tumor segmentation challenge. However, as mentioned earlier, most of them segmented the liver or lesion regions based on 2D slices from 3D volumes. The spatial information has not been taken into account to the maximum extent.

Recently, attention based image classification (Wang et al., 2017) and semantic segmentation architectures (Chen et al., 2016) have attracted a lot of attention. Some medical imaging tasks have used the attention mechanism to solve the issues in real applications. For instance, Schlemper et al. (2019) proposed an attention-gated networks for real-time automated scan plane detection in fetal ultrasound screening. The integrated self-gated soft-attention mechanisms, which can be easily incorporated into other networks, achieved good performances. Overall, it is expected that 3D deep networks combined with the attention mechanism

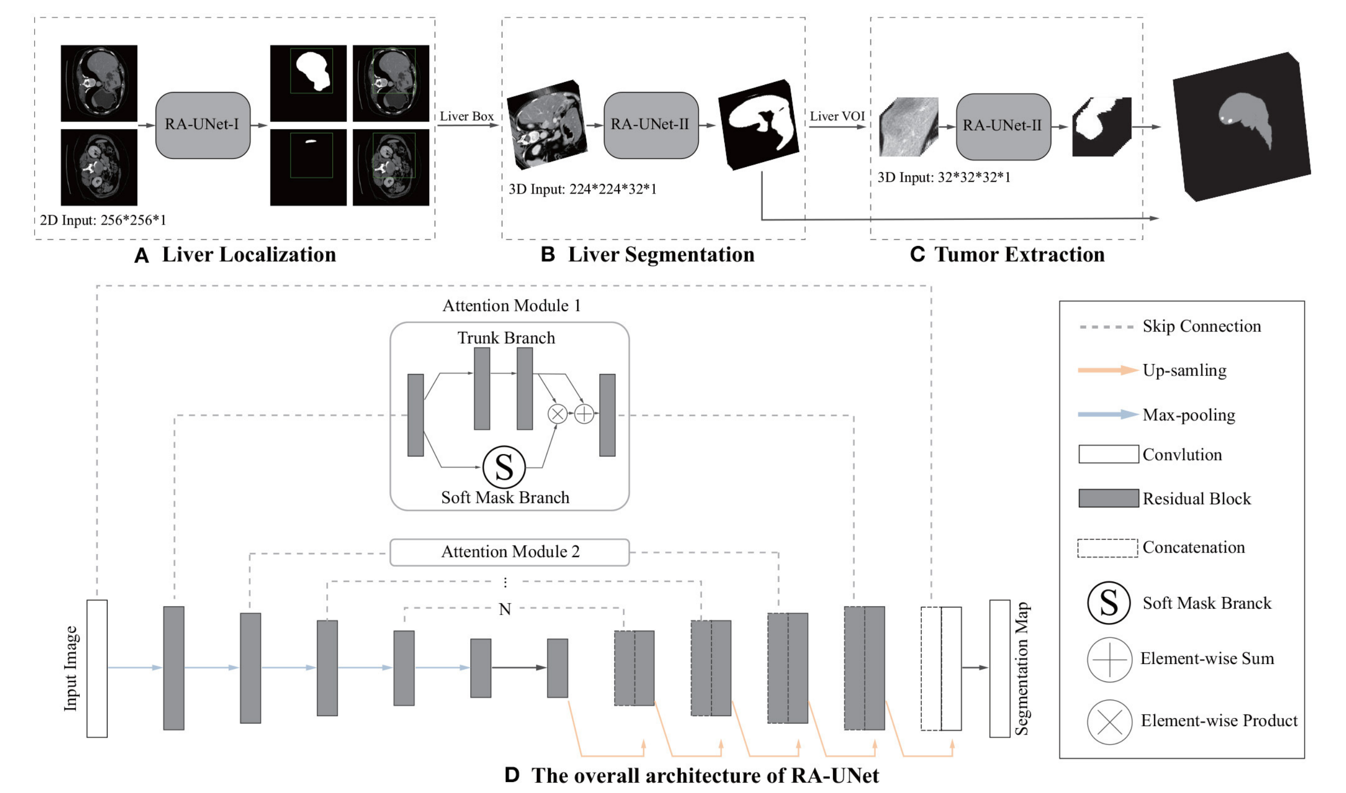


FIGURE 2 | Overview of the proposed pipeline of liver and tumor segmentation. **(A)** A simple version of 2D RA-UNet (RA-UNet-I) is employed for coarse localization of a liver region within a boundary box. **(B)** The 3D RA-UNet (RA-UNet-II) is designed for hierarchically extracting attention-aware features of liver volume of interest (VOI) inside the liver boundary box. **(C)** RA-UNet-II is responsible for an accurate tumor extraction which is inside the liver VOI. **(D)** The overall architecture of RA-UNet.

would achieve a good performance for liver/tumor extraction tasks.

3. METHODOLOGY

3.1. Overview of Our Proposed Architecture

The first time that an attention mechanism was introduced in semantic image segmentation was in Chen et al. (2016), which combined *share-net* with attention mechanisms and achieved good performances. More recently, the attention mechanism is gradually applied to medical image segmentation (Oktay et al., 2018; Schlemper et al., 2019). Inspired by residual attention learning (Wang et al., 2017) and U-Net (Ronneberger et al., 2015), we propose the RA-UNet that for the liver and tumor segmentation tasks. Our overall architecture for segmentation is depicted in **Figure 2**. The proposed architecture consists of three main stages which extract liver and tumor sequentially. Firstly, in order to reduce the overall computational time, we used a 2D residual attention-aware U-Net (RA-UNet) named RA-UNet-I to obtain a coarse liver boundary box. Next, a 3D RA-UNet, which is called RA-UNet-II, was trained to obtain a precise liver volume of interest (VOI). Finally, the obtained liver VOI was sent to a second RA-UNet-II to extract the tumor region. The designed network can handle volumes in various complicated conditions and obtain desirable results in different liver/tumor datasets.

3.2. Datasets and Materials

In our study, we used the public Liver Tumor Segmentation Challenge (LiTS) dataset to evaluate the proposed architecture. This dataset has a total of 200 CT scans containing 130 scans as training data and 70 scans as test data, both of which have the same 512 × 512 in-plane resolution but with different numbers of axial slices in each scan. These training data and their corresponding ground truth are provided by various clinical sites around the world, while the ground truth of the test data is not available.

Another dataset named 3DIRCADb is used as an external test dataset to validate the generalization and scalability of our model. It includes 20 enhanced CT scans and the corresponding manually segmented tumors from European hospitals. The number of axial slices, which have 512 × 512 in-plane resolution, differs for each scan.

3.3. Data Preprocessing

For a medical image volume, Hounsfield units (HU) is a measurement of relative densities determined by CT. Normally, the HU values range from −1,000 to 1,000. Because tumors grow on the liver tissue, the surrounding bones, air, or irrelevant tissues may disturb the segmentation result. Hence, an initial segmentation was used to filter out those noises, leaving the liver region clean which is yet to be segmented. In terms of

TABLE 1 | Typical tissues radiodensities of human body.

Tissue	HU
Air	−200+
Bone	400+
Liver	40∼50
Water	0 ± 10
Blood	3∼14

convenience and efficiency, we took a global windowing step as our data preprocessing strategy.

We list the typical radiodensities of some main tissues in **Table 1**, which shows that these tissues have a wide range of HU values. From the table, the HU value for air is typically above −200; for bone, it is the highest HU values among these tissues; for liver, it is from 40 to 50 HU; for water, it is approximately from 0 to 10 HU; and for blood, it is from 3 to 14 HU.

In this article, we set the HU window at the range from −100 to 200. With such a window, irrelevant organs and tissues were mostly removed. The first rows of **Figure 3** shows the 3D, coronal, sagittal, and axial plane views of the raw volumes of LiTS and 3DIRCADb, respectively. The second rows show the preprocessed volumes with irrelevant organ removed. It can be seen that most of the noise has been removed. The distribution of HU values before and after windowing is illustrated on the left and right of the third rows in **Figure 3** where Frequency denotes the frequency of HU values. We applied the zero-mean normalization and min-max normalization on the data after the windowing. No further image processing was performed.

3.4. RA-UNet Architecture

3.4.1. U-Net as the Basic Architecture

Our RA-UNet has an overall architecture similar to the standard U-Net, consisting of an encoder and a decoder symmetrically on the two sides of the architecture. The contextual information is propagated by the encoder within the rich skip connections which enables the extraction of hierarchical features with more complexities. The decoder receives features that have diverse complexities and reconstructs the features in a coarse-to-fine manner. An advantage is that the U-Net introduces long-range connections through the encoder part and the corresponding decoder part, so that different hierarchical features from the encoder can be merged to the decoder which makes the network much more precise and expansible.

3.4.2. Residual Learning Mechanism

The network depth is of crucial importance. However, gradient vanishing is a common problem in a very deep neural network when carrying out back propagation, which results in poor training results. In order to overcome this problem, He et al. proposed the deep residual learning framework to learn the residual of the identity map (He et al., 2016). In our study, residual blocks are stacked except the first layer and the last layer (**Figure 2D**) to unleash the capability of deep neural networks. The stacked residual blocks solve the gradient vanishing problem at the structural level of the neural network by using identity mappings as the skip connections. The residual units directly propagate features from early convolution to late convolution and consequently improve the performance of the model. The residual block is defined as:

$$\boldsymbol{OR}_{i,c}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{f}_{i,c}(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ denotes the first input of a residual block, $\boldsymbol{OR}$ denotes the output of a residual block, i ranges over all spatial positions, $c \in \{1, \ldots, C\}$ indicates the index of channels, C is the total number of channels, and $\boldsymbol{f}$ represents the residual mapping to be learned.

The residual block consists of three sets of combinations of a batch normalization (BN) layer, an activation (ReLU) layer, and a convolutional layer. A convolutional identity mapping connection is used to ensure the accuracy as the network goes "deeper" (He et al., 2016). The detailed residual unit is illustrated in **Figure 4**.

3.4.3. Attention Residual Learning Mechanism

The performance will drop if only naive stacking is used for the attention modules. This can be solved by the attention residual learning proposed by Wang et al. (2017). The attention residual mechanism divides the attention module into a trunk branch and a soft mask branch, where the trunk branch is used to process the original features and the soft mask branch is used to construct the identity mapping. The output $\boldsymbol{OA}$ of the attention module under attention residual learning can be formulated as:

$$\boldsymbol{OA}_{i,c}(\boldsymbol{x}) = (1 + \boldsymbol{S}_{i,c}(\boldsymbol{x}))\boldsymbol{F}_{i,c}(\boldsymbol{x}) \tag{2}$$

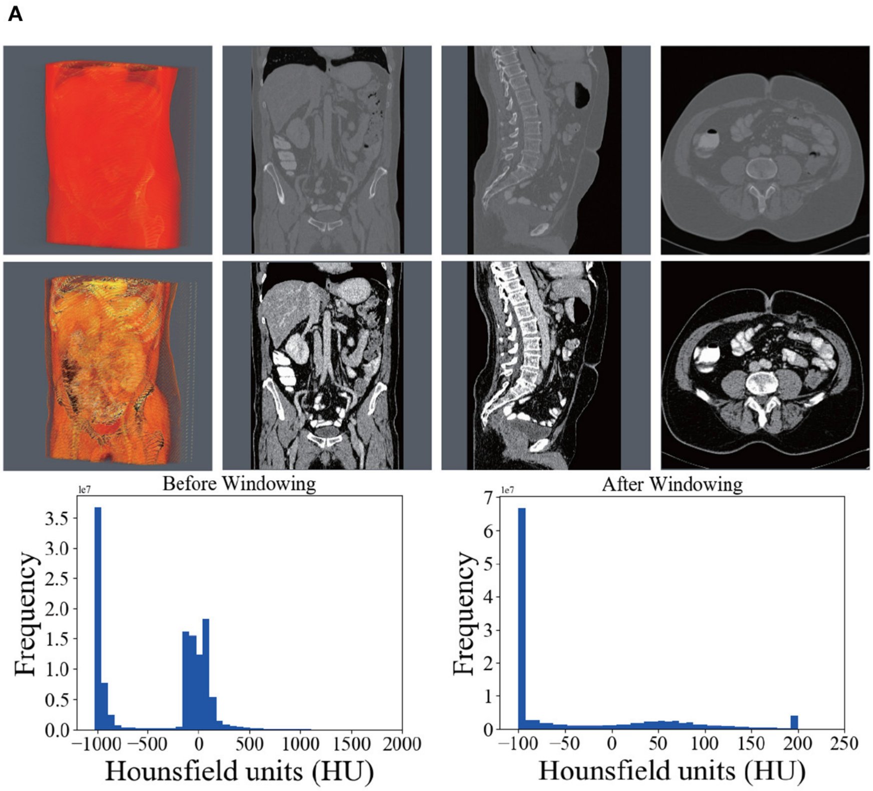
where $\boldsymbol{S}(\boldsymbol{x})$ has values in [0,1]. If $\boldsymbol{S}(\boldsymbol{x})$ is close to 0, $\boldsymbol{OA}(\boldsymbol{x})$ will approximate the original feature maps $\boldsymbol{F}(\boldsymbol{x})$. The soft mask branch $\boldsymbol{S}(\boldsymbol{x})$, which selects identical features and suppresses noised from the trunk branch, plays the most important role in the attention residual mechanism.

The soft mask branch has an encoder-decoder structure which has been widely applied to medical image segmentation (Ronneberger et al., 2015; Çiçek et al., 2016; Alom et al., 2018). In the attention residual mechanism, it is designed to enhance good features and reduce the noises from the trunk branch. The encoder in the soft mask branch contains a max-pooling operation, a residual block, and a long-range residual block connected to the corresponding decoder, where an element-wise sum is performed following a residual block and an up-sampling operation. After the encoder and decoder parts of the soft mask, two convolutional layers and one Sigmoid layer are added to normalize the output. **Figure 5** illustrates the attention residual module in detail.

In general, the attention residual mechanism can keep the original feature information through the trunk branch and pay attention to those liver tumor features by the soft mask branch.

3.4.4. Loss Function

The weights are learnt by minimizing the loss function. We employed a loss function based on the Dice coefficient proposed



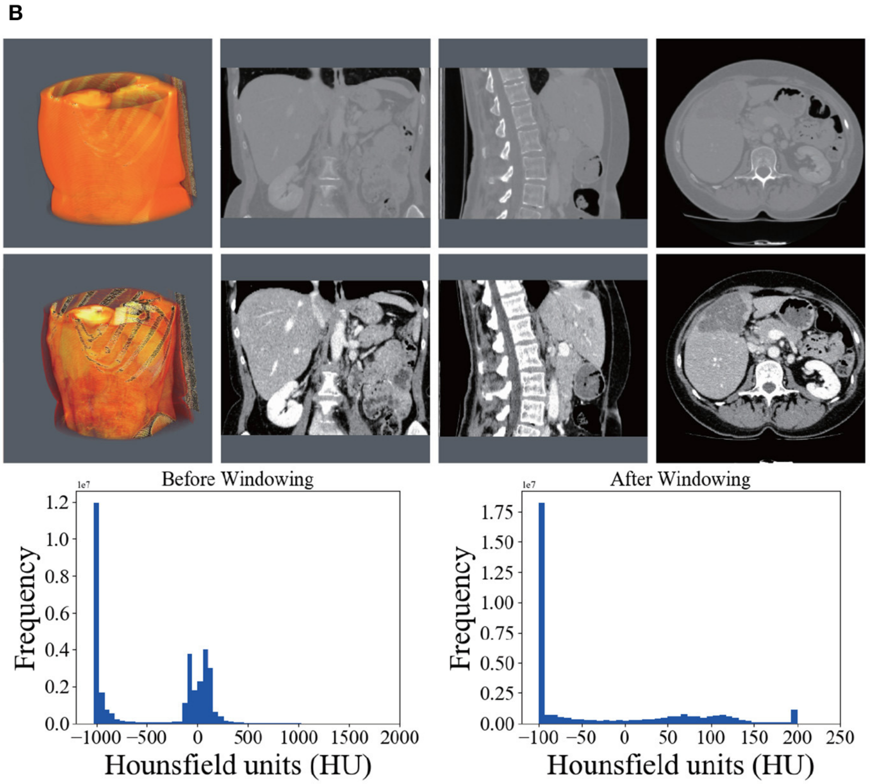


FIGURE 3 | Comparison between the raw CT scans (first row), windowed (second row) scans, and histograms of HU (third row) before and after windowing. **(A)** Shows the comparison on LiTS. **(B)** Shows the comparison on 3DIRCADb.

in Milletari et al. (2016) in this study. The loss $\boldsymbol{L}$ is defined as follows:

$$\boldsymbol{L} = 1 - \frac{2\sum_{i=1}^{N} \boldsymbol{s}_i \boldsymbol{g}_i}{\sum_{i=1}^{N} \boldsymbol{s}_i^2 + \sum_{i=1}^{N} \boldsymbol{g}_i^2} \tag{3}$$

where N is the number of voxels, $\boldsymbol{s}_i$ and $\boldsymbol{g}_i$ belong to the binary segmentation and binary ground truth voxel sets, respectively. The loss function measures the similarity of two samples directly.

3.5. Liver Localization Using RA-UNet-I

The first stage aimed to locate the 3D liver boundary box. A 2D version RA-UNet-I was introduced here to segment a coarse liver region, which can reduce the computational cost of the subsequent RA-UNet-II, remove the redundant information, and provide more effective information. It worked as a "baseline" to limit the scope of the liver.

We down sampled the slices to 256×256 and fed the preprocessed slices into the trained RA-UNet-I model. Next, we stacked all the slices in their original sequence. Afterwards, a 3D connected-component labeling (Hossam et al., 2010) was employed. The connected component labeling, which is used for determining specific regions and measure the size of regions, is a procedure for assigning a unique label to each connected component in an image. Then the largest component was chosen as the coarse liver region. Finally, we interpolated the liver region to its original volume size with a 512 × 512 in-plane resolution.

Connected component labeling is a procedure for assigning a unique label to each connected component in an image.

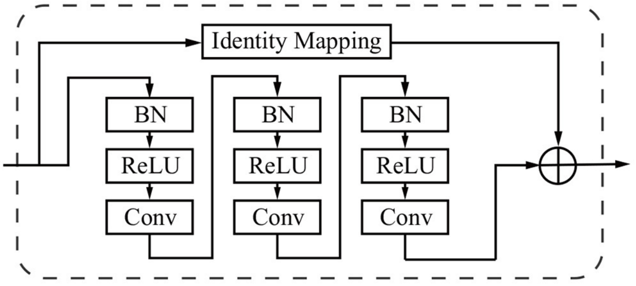


FIGURE 4 | Sample of a residual block in the dashed window. An identity mapping and convolutional blocks are added before the final feature output.

3.6. Liver Segmentation Using RA-UNet-II

The RA-UNet-II was a 3D model which fully utilized the volume information and captured the spatial information. The 3D U-Net type architecture (Çiçek et al., 2016) merges the low resolution and high resolution features to generate an accurate segmentation. Meanwhile, using large image patches (224 × 224 × 32) for training provides much richer contextual information than using small image patches, which usually leads to more global segmentation results.

As shown in **Table 2**, the network went down from the top to the bottom in the encoder, and reversed in the decoder. During the encoding phase, the RA-UNet-II received liver patches and passed them down to the bottom. During the decoding phase, lower features were passed from the bottom to the top with resolution doubled through the up-sampling operation. Note that the long-range connection between the encoder and the decoder was realized by the attention block. We then combined the features from the attention blocks with those from the corresponding up-sampling level in the decoder via concatenation. Then the concatenated features were passed on to the decoder. Finally, an activation layer (i.e., Sigmoid) was used to generate the final probability map of liver segmentation.

The RA-UNet-II has fewer parameters than the traditional U-Net (Ronneberger et al., 2015). With this architecture, the number of parameters has been largely decreased to only 4M training parameters. During the training phase, we interpolated the liver boundary box in the $x-y$ plane to a fixed size (i.e., 224×224) and randomly picked 32 slices successively in the z direction to form the training patches. The RA-UNet-II was employed on each CT patch to generate 3D liver probability patches in sequence. Then, we interpolated and stacked those probability patches to be restored to the original size of the boundary box. A voting strategy was used to generate the final liver probability of the VOI from overlapped sub-patches. A 3D connected-component labeling was used and the largest

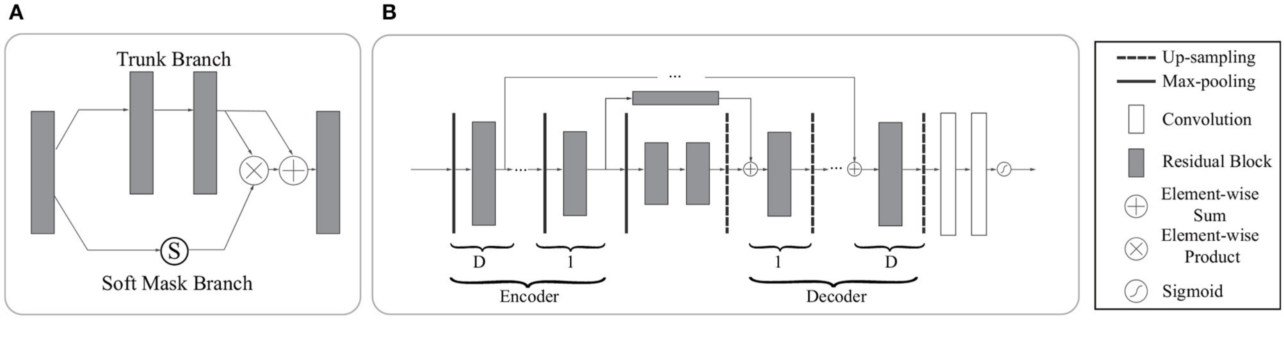


FIGURE 5 | The architecture of the attention residual module. **(A)** The attention residual module contains a trunk branch and a soft mask branch. The trunk branch learns original features while the soft mask branch focuses on reducing noises and enhancing good features. **(B)** The soft mask branch contains a stack of encoder-decoder blocks. D denotes the depth of skip connections. In our network, we set D to 0,1,2,3 according to the specific location of the attention residual block.

TABLE 2 | Architecture of the proposed RA-UNET-II in liver localization stage.

Encoder	Output size	Decoder	Pre-operation	Output size
Input	224 × 224 × 32 × 1	Att1	[Res4], depth=0	14 × 14 × 2 × 256
Conv1	224 × 224 × 32 × 32	Res7	[Up1, Att1]	14 × 14 × 2 × 256
Pooling	112 × 112 × 16 × 32	Up2		28 × 28 × 4 × 256
Res1	112 × 112 × 16 × 32	Att2	[Res3], depth=1	28 × 28 × 4 × 128
Pooling	56 × 56 × 8 × 32	Res8	[Up2, Att2]	28 × 28 × 4 × 128
Res2	56 × 56 × 8 × 64	Up3		56 × 56 × 8 × 128
Pooling	56 × 56 × 4 × 64	Att3	[Res2], depth=2	56 × 56 × 8 × 64
Res3	28 × 28 × 4 × 128	Res9	[Up3, Att3]	56 × 56 × 8 × 64
Pooling	14 × 14 × 2 × 128	Up4		112 × 112 × 16 × 64
Res4	14 × 14 × 2 × 256	Att4	[Res1], depth=3	112 × 112 × 16 × 32
Pooling	7 × 7 × 1 × 256	Res10	[Up4, Att4]	112 × 112 × 16 × 32
Res5	7 × 7 × 1 × 512	Up5		224 × 224 × 32 × 32
Res6	7 × 7 × 1 × 512	Conv2	[Up5, Conv1]	224 × 224 × 32 × 32
Up1	14 × 14 × 2 × 512	Conv3		224 × 224 × 32 × 1

Here [ ], long range connection; [,], concatenate operation; Conv, convolution; Up, up-sampling; Res, residual block; Att, attention block.

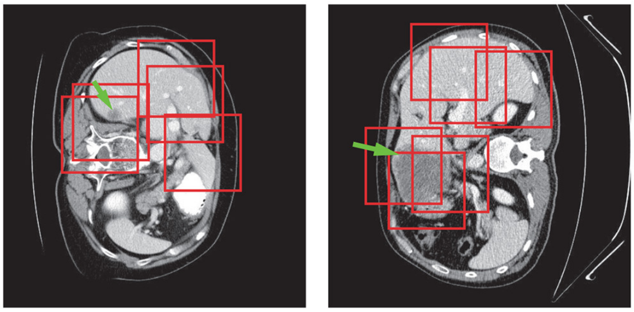


FIGURE 6 | Tumor patch extraction results. The green arrows point to the tumor regions and the red boxes show the patches used for training.

component was chosen on the merged VOI to yield the final liver region.

3.7. Extraction of Tumors Based on RA-UNet-II

Tumor region extraction was similar to liver segmentation but no interpolation and resizing were performed. Because the size of the tumor is much smaller than that of the liver, the original tumor resolution was used to avoid losing small lesions. Furthermore, in order to solve the data imbalance issue and learn more effective tumor features, we picked patches on both tumor and its surroundings non-tumor regions for training as shown in **Figure 6**. Note that only those in the liver VOIs would be the candidate patches for training. We extracted the tumors following a similar routine as for the liver segmentation step except the use of interpolation. Subsequently, a voting strategy is used again on the merged VOI to yield the final tumor segmentation. At last, we filtered out those voxels which were not in the liver region.

3.8. Evaluation Metrics

We evaluated the performance of the proposed approach using the metrics introduced in Heimann et al. (2009). The evaluation metrics include the Dice score (DS) (Wu et al., 2016) consist of Dice global (Dice score computed on all combined volumes denoted with DG) and Dice per case (mean Dice score per volume denoted with DC), Jaccard similarity coefficient (Jaccard), volumetric overlap error (VOE), relative volume difference (RVD), average symmetric surface distance (ASSD), and maximum surface distance (MSD).

3.9. Implementation Details

The RA-UNet architecture was constructed using the Keras (Chollet, 2015) and the TensorFlow (Abadi et al., 2015) libraries. All the models were trained from scratch. The parameters of the network were initialized with random values and then they were trained with back-propagation based on Adam (Kingma and Ba, 2014) with an initial learning rate (LR) of 0.001, β_1=0.9, and β_2=0.999. The learning rate would be reduced to LR×0.1 if the network went to plateau after 20 epoches. We used 5-fold cross-training on the LiTS training dataset, and evaluated the performance on the LiTS test dataset. To demonstrate the generalization of our RA-UNet, we also evaluated the performance on the 3DIRCADb dataset using the well-trained weights from the LiTS training dataset. For the liver and tumor training, the total numbers of epoches were set at 50 and 50 for each fold, respectively. An integration operation by a voting strategy is implemented to ensemble all the prediction results of 5 models. The training of all the models was performed with an NVIDIA 1080Ti GPU. In our experiments, it took about 100/40 min to train an epoch of our 3D RAUNet for liver/tumor segmentation, respectively.

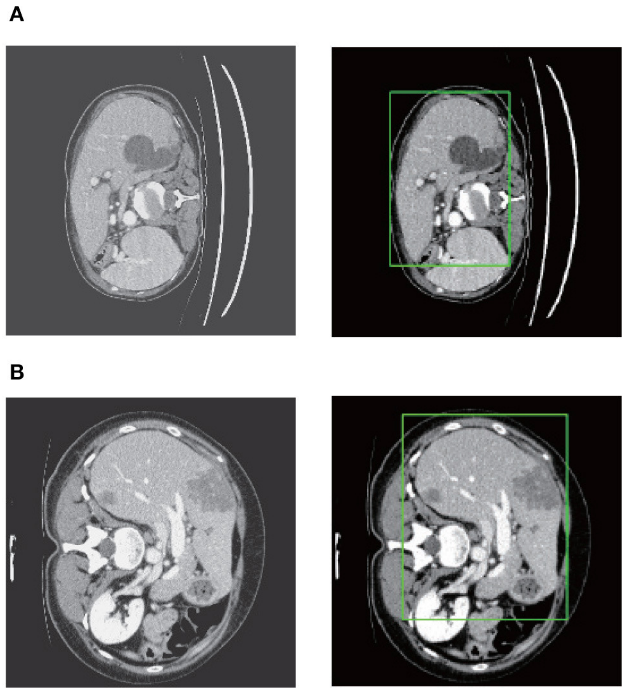


FIGURE 7 | Liver localization using RA-UNet-I. From left to right the figure shows the preprocessed slice, and the final boundary box which restricts the liver region. **(A)** A typical slice from the LiTS validation dataset. **(B)** A typical slice from the 3DIRCADb dataset. The RA-UNet-I enables the coarse localization of liver regions.

4. EXPERIMENTS AND RESULTS

4.1. Liver Volume of Interest Localization

In order to reduce the computational cost, we first down-sampled the input slices to a 256 × 256 pixel in-plane resolution. Secondly, we used all the slices which have liver in the images together with 1/3 of those randomly picked slices without liver as the training data. There are a total of 32,746 slices with liver which were used, including 23,283 slices for training and 9,463 slices for validation. Note that 5-fold training was not employed at this stage, because our goal at this stage was to obtain a coarse liver boundary box and reduce the computational time.

After stacking all the slices and employing the 3D connected-component labeling, we calculated the 3D boundary box of the slices with liver, and extended 10 pixels in coronal, sagittal, and axial directions to ensure that the entire liver region was included. **Figure 7** shows the liver localization results from RA-UNet-I. It demonstrates that the attention mechanism has successfully constrained the liver region. Note that this stage aims to reduce the computational cost for precisely segmenting liver and tumor by RA-UNet-II.

4.2. Liver Segmentation Using RA-UNet-II

RA-UNet-II allowed the network to go "deeper." However, the implementation of a 3D convolution is still limited by the hardware and memory requirements (Prasoon et al., 2013). In order to balance the computational cost and efficiency, we first carried out interpolation in the region inside the liver boundary box to the size of 224×224×M, where M was the axial length of the liver boundary box. Then we cropped the volumetric patches (224×224×32) randomly from each boundary box, which was constrained by the liver boundary box. Totally, 4,077/1,019 patches were selected for training/validation.

Figure 8 shows the liver segmentation based on RA-UNet-II, which indicates that our proposed network has the ability to learn 3D contextual information and could successfully extract the liver from adjacent slices in an image volume. After the 3D connected-component labeling was carried out, the liver region was precisely extracted by selecting the largest region.

As shown in **Table 3**, our method reached up to 0.961 and 0.977 Dice scores on the LiTS test dataset and the 3DIRCADb dataset, respectively. It reveals that RA-UNet yields remarkable liver segmentation results. Then we can extract tumors from the segmented liver regions.

4.3. Extraction of Tumors Based on RA-UNet-II

Tumors were tiny structures compared to livers. Therefore, no interpolation or resizing was applied to tumor patch sampling to avoid information loss from image scaling. It was difficult to decide what size of patch for training could reach a desirable performance. In order to determine the patch size, we set the patch size of 32×32×32, 64×64×32, and 128×128×32, respectively to test the performance of tumor segmentation. Results showed that 128×128×32 patch-sized data achieved the best tumor segmentation performance. The larger the patch size was, the richer context in formation the patches could provide. Due to the limitation of computational resources, 128×128×32 was chosen empirically for tumor patches. We randomly picked 150 patches from each liver volume in the boundary box. Totally, 14,160/3,540 patches were chosen from LiTS as training/validation datasets. As shown in **Table 4**, our method reached 0.595 and 0.830 Dice scores on the LiTS test dataset and the 3DIRCADb dataset, respectively. **Figure 9** shows the tumor segmentation results in detail.

Figure 10 shows the liver/tumor segmentation results. It shows that liver regions which are large in size are successfully segmented and tumors that are tiny and hard to detect can be identified by the proposed method as well. Due to the low contrast with the surrounding livers and the extremely small size of some tumors, the proposed method still has some false positives and false negatives for tumor extraction.

4.4. Comparison With Other Methods

There were several submissions about liver and tumor segmentations to the 2017 ISBI and MICCAI LiTS challenges. We reached a Dice per case of 0.961, Dice global of 0.963, Jaccard of 0.926, VOE of 0.074, RVD of 0.002, ASSD of 1.214, and MSD of 26.948, which is a desirable performance on the LiTS challenge for liver segmentation. For tumor segmentation evaluation, our method reached a Dice per case of 0.595, Dice global of 0.795, Jaccard of 0.611, VOE of 0.389, RVD of −0.152, ASSD of 1.289, and MSD of 6.775. Compared with other methods, Bellver

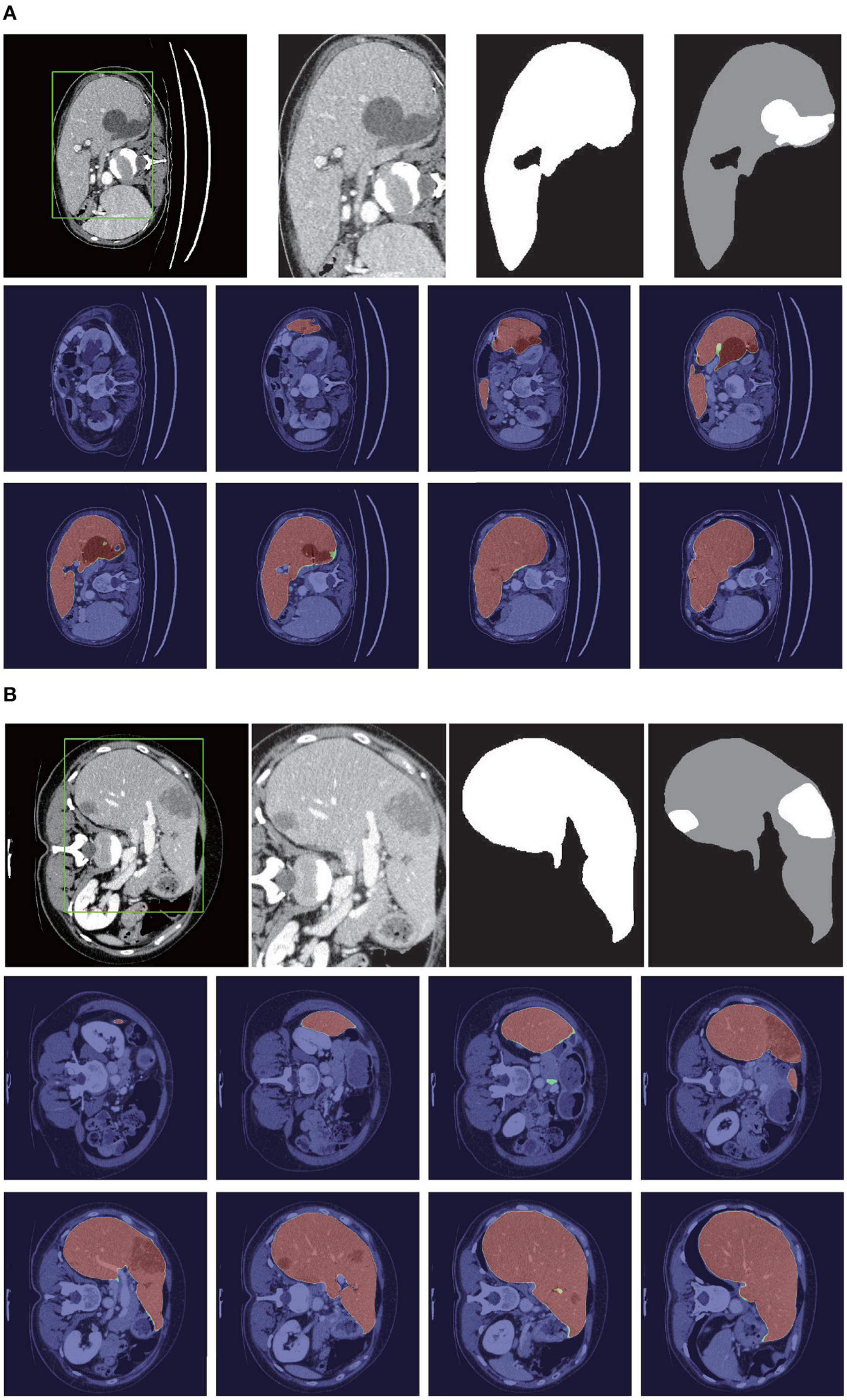


FIGURE 8 | Liver segmentation results based on RA-UNet-II. **(A)** From the LiTS validation dataset and **(B)** is from the 3DIRCADb dataset. From left to right, the first row of each subplot shows the liver in the green boundary box, magnified liver region, the liver segmentation results, and the corresponding ground truth. The second and the third rows show the probability heat map of liver segmentation results. The darker the color, the higher the probability of the liver region. Note that the ground truth contains liver in gray and tumor in white.

TABLE 3 | Evaluation results of the liver segmentation on the LiTS test dataset and the 3DIRCADb dataset.

	LiTS	3DIRCADb
DC	0.961	0.977
Jaccard	0.926	0.977
VOE	0.074	0.045
RVD	0.002	−0.001
ASSD	1.214	0.587
MSD	26.948	18.617

TABLE 4 | Scores of the tumor segmentation on the LiTS test dataset and the 3DIRCADb dataset.

	LiTS	3DIRCADb
DC	0.595	0.830
Jaccard	0.611	0.744
VOE	0.389	0.255
RVD	−0.152	0.740
ASSD	1.289	2.230
MSD	6.775	53.324

et al. (2017) and Pandey et al. (2018) methods reached tumor Dice per case at 0.587 and 0.59, respectively, which were 2D segmentation methods. Our approach outperformed these two methods. The detailed results and all the performances are listed in **Table 5**. It is worth mentioning that our method for precise segmentation of liver and tumor was a full 3D technique with a much deeper network.

4.5. Generalization of the Proposed RA-UNet

To show the generalization of the proposed method, we used the weights well-trained on LiTS and tested on the 3DIRCADb dataset. Some works concentrated on liver segmentation, and there were a few about tumor segmentation. Hence, we listed the results of some approaches in **Table 6**. Our methods reached a Dice per case of 0.977, Jaccard of 0.977, VOE of 0.045, RVD of −0.001, ASSD of 0.587, and MSD of 18.617, which quantitatively show that our method performed significantly better than all the other methods on liver segmentation. Since most of the works aimed at liver segmentation, few of them displayed tumor segmentation results, we only compared with Christ et al. (2017a) on the 3DIRCADb dataset. It was worth mentioning that our method reached a mean Dice score of 0.830 on livers with tumors compared to a mean Dice score of 0.56 for the method by Christ et al. (2017a). The visualization of typical performance was illustrated in **Figures 8B**, **9B**, **10B**, which qualitatively indicated that our method produced precise segmentation performance.

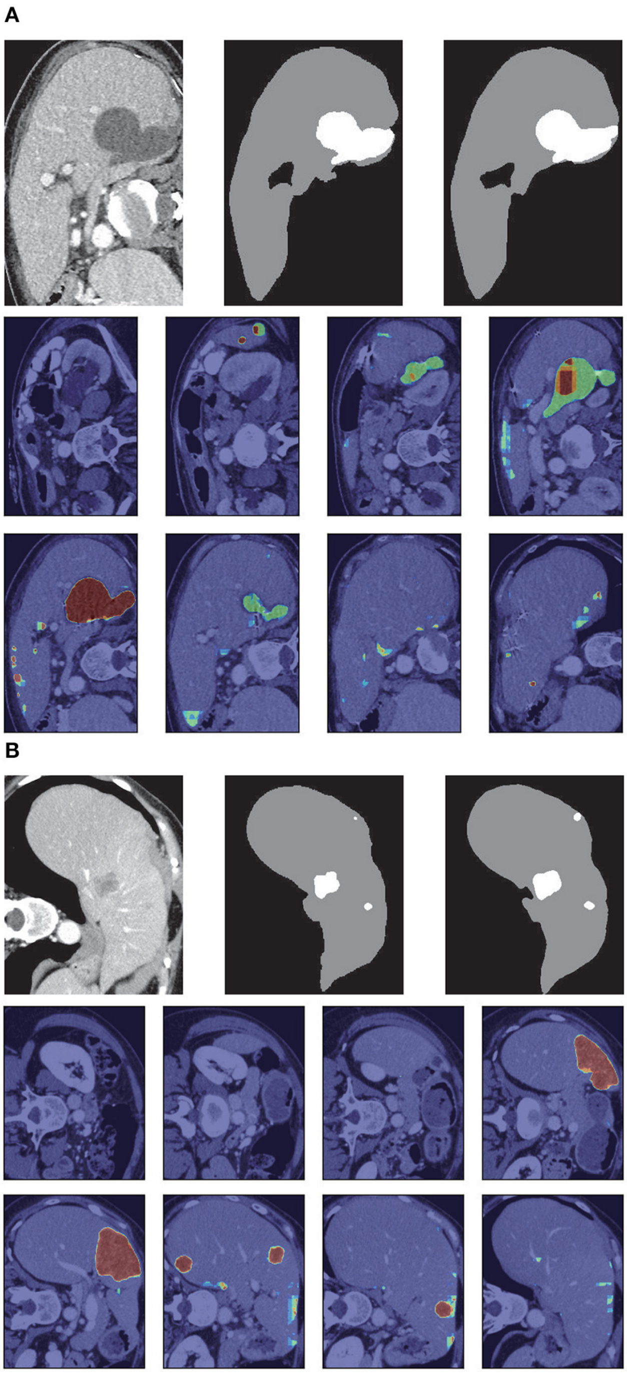


FIGURE 9 | Tumor segmentation results based on RA-UNet-II. **(A)** From the LiTS validation dataset, and **(B)** is from the 3DIRCADb dataset. From left to right, the first row of each subplots indicates the raw images, segmentation results of liver tumor, and the corresponding ground truth. The second and the third rows show the probability heat map of tumor segmentation results.

5. CONCLUSION

To summarize our work, we have proposed an effective and efficient hybrid architecture for automatic extraction of liver and tumor from CT volumes. We introduce a new 3D residual attention-aware liver and tumor segmentation neural network named RA-UNet, which allows the extraction

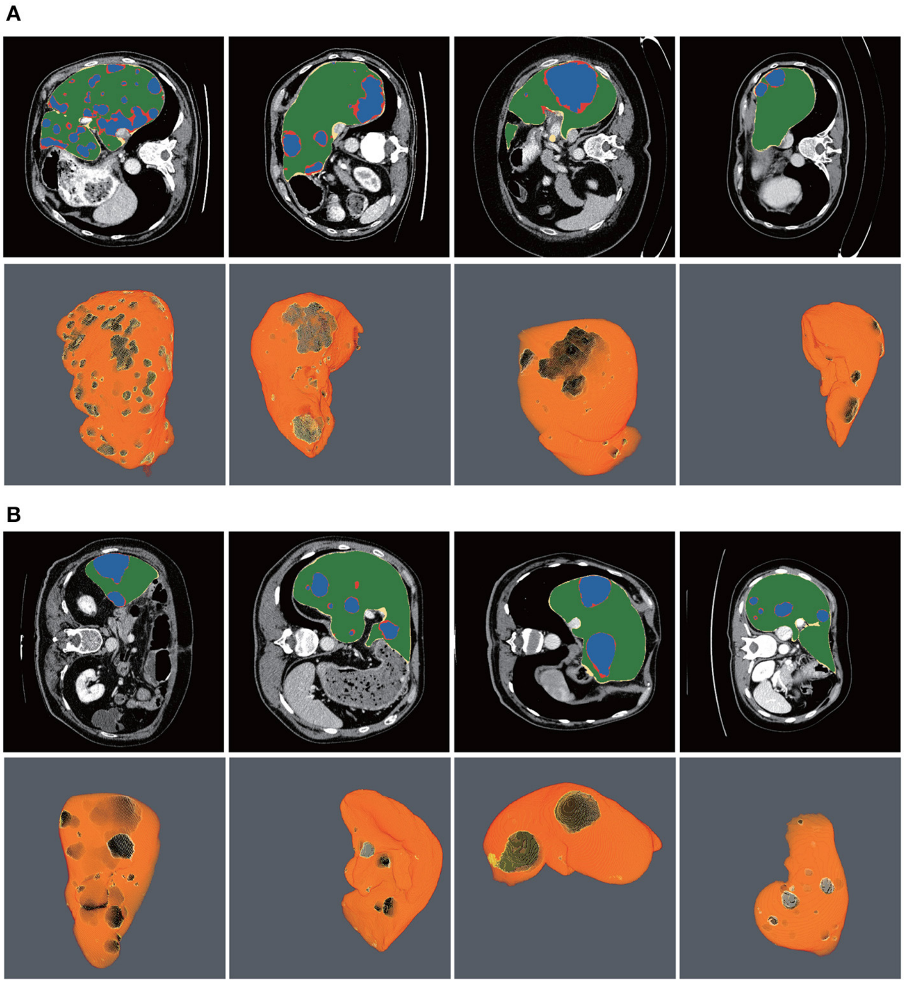


FIGURE 10 | Automatic liver and tumor segmentation with RA-UNet. The green regions indicate the correctly extracted liver, the yellow regions are the wrongly extracted liver, the blue color depicts the correctly extracted tumor regions, and the red color means wrongly extracted tumor. The first row of each subplot shows four slices from different volumes in the axial view and the second row of each subplot shows the corresponding 3D view of the entire liver/tumor segmentation results. **(A)** From the LiTS dataset. **(B)** From the 3DIRCADb dataset.

TABLE 5 | Segmentation results compared with other methods on the LiTS test dataset.

		LiTS liver							**LiTS tumor**						
	Dimension	**DC**	**DG**	**Jaccard**	**VOE**	**RVD**	**ASSD**	**MSD**	**DC**	**DG**	**Jaccard**	**VOE**	**RVD**	**ASSD**	**MSD**
Kaluva et al. (2018)	2D	0.912	0.923	0.850	0.150	−0.008	6.465	45.928	0.492	0.625	0.589	0.411	19.705	1.441	7.515
Bi et al. (2017)	2D	0.959	–	0.922	–	–	–	–	0.500	–	0.388	–	–	–	–
Li et al. (2018)	2.5D	0.961	0.965	–	0.074	−0.018	1.450	27.118	0.722	0.824	–	0.366	4.272	1.102	6.228
MEDDIIR	Unknown	0.950	0.955	–	0.094	0.047	1.597	28.911	0.658	0.819	–	0.380	−0.129	1.113	6.323
Yuan (2017)	2D	0.963	0.967	–	0.071	−0.010	1.104	23.847	0.657	0.820	–	0.378	0.288	1.151	6.269
Summer	Unknown	0.941	0.945	–	0.108	−0.066	6.552	152.350	0.631	0.786	–	0.400	−0.181	1.184	6.367
Proposed method	3D	0.961	0.963	0.926	0.074	0.002	1.214	26.948	0.595	0.795	0.611	0.389	−0.152	1.289	6.775

of 3D structures in a pixel-to-pixel fashion. The proposed network takes advantage of the strengths from the U-Net, the residual learning, and the attention residual mechanism. Firstly, attention-aware features change adaptively with the use of attention modules. Secondly, the residual blocks are stacked into our architecture which allows the architecture to

TABLE 6 | Segmentation results compared with other methods on the 3DIRCADb dataset.

		3DIRCADb liver						3DIRCADb tumor
	Dimension	**DC**	**Jaccard**	**VOE**	**RVD**	**ASSD**	**MSD**	**DC**
Christ et al. (2017a)	2D	0.943	–	0.107	−0.014	1.6	24	0.56
Ronneberger et al. (2015)	2D	0.729	–	0.39	0.87	19.4	119	–
Li et al. (2013)	2D	0.945	–	0.068	−0.112	1.6	28.2	–
Eapen et al. (2015)	3D	–	–	0.0554	0.0093	0.78	15.6	–
Lu et al. (2017)	3D	–	–	0.0936	0.0097	1.89	33.14	–
Proposed method	3D	0.977	0.977	0.045	−0.001	0.587	18.617	0.83

go deeply and solve the gradient vanishing problem. Finally, the U-Net is used to capture multi-scale attention information and integrate low-level features with high-level features. To the best of our knowledge, this is the full 3D model and the first time that the attention residual mechanism is implemented in the medical imaging tasks. Fewer parameters are trained by the attention residual mechanism. The proposed method enlarges the U-Net family for 3D liver and tumor segmentation tasks, which is crucial for real-world applications. The effective system includes three stages: liver localization by the RA-UNet-I, precise segmentation of liver, and tumor lesion by the RA-UNet-II. More importantly, the trained network is a general segmentation model working on both the LiTS and the 3DIRCADb datasets.

Overall, our method achieved competitive performances in liver tumor challenge, and exhibits high extension and generalization ability in another tumor segmentation dataset. The proposed model has great potential to be applied to other modalities of medical images. It may also assist surgeons to find treatment for novel tumors. The limitation of the proposed method is the training time because the 3D convolutions require larger parameters than the 2D convolutions. In future work, we aim to further improve the architecture, making the architecture much more general to other tumor segmentation datasets and more flexible to common medical imaging tasks. What's more, reducing computational cost and developing a lightweight architecture for speeding training time are also under consideration.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://competitions.codalab.org/competitions/17094.

AUTHOR CONTRIBUTIONS

QJ conducted the experiments. ZM, CS, and HC participated in manuscript writing. RS designed the experiments and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 61702361), the Science and Technology Program of Tianjin, China (Grant No. 16ZXHLGX00170), the National Key Technology R&D Program of China (Grant No. 2015BAH52F00), the National Key Technology R&D Program of China (Grant No. 2018YFB1701700), and the program of China Scholarships Council.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at arXiv (Jin et al., 2018).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]. arXiv:1603.04467*.

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv [Preprint]. arXiv:1802.06955*. doi: 10.1109/NAECON.2018.8556686

Andermatt, S., Pezold, S., and Cattin, P. (2016). "Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data," in *Deep Learning and Data Labeling for Medical Applications* (Athens: Springer), 142–151. doi: 10.1007/978-3-319-46976-8_15

Bellver, M., Maninis, K., Ponttuset, J., Nieto, X. G. I., Torres, J., and Van Gool, L. (2017). Detection-aided liver lesion segmentation using deep learning. *arXiv [Preprint]. arXiv:1711.11069*.

Bi, L., Kim, J., Kumar, A., and Feng, D. (2017). Automatic liver lesion detection using cascaded deep residual networks. *arXiv [Preprint]. arXiv:1704.02703*.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bui, T. D., Shin, J., and Moon, T. (2017). 3D densely convolution networks for volumetric segmentation. *arXiv [Preprint]. arXiv:1709.03199*.

Chen, L. C., Yang, Y., Wang, J., Xu, W., and Yuille, A. L. (2016). "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 3640–3649. doi: 10.1109/CVPR.2016.396

Chollet, F. (2015). *Keras*. GitHub.

Christ, P. F., Ettlinger, F., Grun, F., Elshaer, M. E. A., Lipkova, J., Schlecht, S., et al. (2017a). Automatic liver and tumor segmentation of CT and MRI Volumes using cascaded fully convolutional neural networks. *arXiv: Computer Vision and Pattern Recognition*.

Christ, P. F., Ettlinger, F., Kaissis, G., Schlecht, S., Ahmaddy, F., Grün, F., et al. (2017b). "SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D convolutional neural networks," in *IEEE International Symposium on Biomedical Imaging (ISBI 2017)*, 839–843. doi: 10.1109/ISBI.2017.7950648

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Mach. Learn.* 48, 253–285. doi: 10.1023/A:1013912006537

Dolz, J., Desrosiers, C., and Ayed, I. B. (2017). 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. *NeuroImage* 170, 456–470. doi: 10.1016/j.neuroimage.2017.04.039

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., and Heng, P.-A. (2016). "3D deeply supervised network for automatic liver segmentation from CT volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 149–157. doi: 10.1007/978-3-319-46723-8_18

Eapen, M., Korah, R., and Geetha, G. (2015). Swarm intelligence integrated graph-cut for liver segmentation from 3D-CT Volumes. *Sci. World J.* 2015:823541. doi: 10.1155/2015/823541

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914. doi: 10.1093/bioinformatics/16.10.906

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Heimann, T., Van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., et al. (2009). Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans. Med. Imag.* 28, 1251–1265. doi: 10.1109/TMI.2009.2013851

Heker, M., and Greenspan, H. (2020). Joint liver lesion segmentation and classification via transfer learning. *arXiv [Preprint]. arXiv:2004.12352*.

Hong, Q., Shi, Z., Sun, J., and Du, S. (2020a). Memristive self-learning logic circuit with application to encoder and decoder. *Neural Comput. Appl.* 1–13. doi: 10.1007/s00521-020-05281-z

Hong, Q., Yan, R., Wang, C., and Sun, J. (2020b). Memristive circuit implementation of biological nonassociative learning mechanism and its applications. *IEEE Trans. Biomed. Circuits Syst.* 14, 1036–1050. doi: 10.1109/TBCAS.2020.3018777

Hossam, M. M., Hassanien, A. E., and Shoman, M. (2010). "3D brain tumor segmentation scheme using K-mean clustering and connected component labeling algorithms," in *International Conference on Intelligent Systems Design and Applications (ISDA)* (Cairo: IEEE), 320–324. doi: 10.1109/ISDA.2010.5687244

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., and Su, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

Jin, Q., Meng, Z., Sun, C., Wei, L., and Su, R. (2018). RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *arXiv [Preprint]. arXiv:1811.01328*.

Kaluva, K. C., Khened, M., Kori, A., and Krishnamurthi, G. (2018). 2D-densely connected convolution neural networks for automatic liver and tumor segmentation. *arXiv [Preprint]. arXiv:1802.02182*.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]. arXiv:1412.6980*.

Lee, N., Laine, A. F., and Klein, A. (2011). "Towards a deep learning approach to brain parcellation," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (Chicago, IL), 321–324. doi: 10.1109/ISBI.2011.5872414

Li, C., Wang, X., Eberl, S., Fulham, M., Yin, Y., Chen, J., et al. (2013). A likelihood and local constraint level set model for liver tumor segmentation from CT volumes. *IEEE Trans. Biomed. Eng.* 60, 2967–2977. doi: 10.1109/TBME.2013.2267212

Li, W., Jia, F., and Hu, Q. (2015). Automatic segmentation of liver tumor in CT images with deep convolutional neural networks. *J. Comput. Commun.* 3, 146–151. doi: 10.4236/jcc.2015.311023

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C. -W., and Heng, P. -A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans Med. Imag.* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing* 234, 11–26. doi: 10.1016/j.neucom.2016.12.038

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Lu, F., Wu, F., Hu, P., Peng, Z., and Kong, D. (2017). Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *Int. J. Comput. Assist. Radiol. Surg.* 12, 171–182. doi: 10.1007/s11548-016-1467-3

Milletari, F., Navab, N., and Ahmadi, S. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision* (Palo Alto, CA), 565–571. doi: 10.1109/3DV.2016.79

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-net: learning where to look for the pancreas. *arXiv [Preprint]. arXiv:1804.03999*.

Pandey, R. K., Vasan, A., and Ramakrishnan, A. (2018). Segmentation of liver lesions with reduced complexity deep models. *arXiv [Preprint]. arXiv:1805.09233*.

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imag.* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., and Nielsen, M. (2013). "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Nagoya-shi: Springer), 246–253. doi: 10.1007/978-3-642-40763-5_31

Rajagopal, R., and Subbaiah, P. (2015). A survey on liver tumor detection and segmentation methods. *ARPN J. Eng. Appl. Sci.* 10, 2681–2685.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: learning to leverage salient regions in medical images. *Med. Image Anal.* 53, 197–207. doi: 10.1016/j.media.2019.01.012

Seo, H., Huang, C., Bassenne, M., Xiao, R., and Xing, L. (2019). Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. *IEEE Trans. Med. Imag.* 39, 1316–1325. doi: 10.1109/TMI.2019.2948320

Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., et al. (2016). "Sub-cortical brain structure segmentation using F-CNN's," in *International Symposium on Biomedical Imaging*, 269–272. doi: 10.1109/ISBI.2016.7493261

Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., et al. (2010). "3D Image reconstruction for comparison of algorithm database: a patient specific anatomical and medical image database," in *IRCAD* (Strasbourg: Tech. Rep).

Song, B., Li, K., Orellana-Martín, D., Valencia-Cabrera, L., and Pérez-Jiménez, M. J. (2020a). Cell-like P systems with evolutional symport/antiport rules and membrane creation. *Inf. Comput.* 275:104542. doi: 10.1016/j.ic.2020.104542

Song, B., Zeng, X., Jiang, M., and Pérez-Jiménez, M. J. (2020b). Monodirectional tissue P systems with promoters. *IEEE Trans. Cybern.* 1–13. doi: 10.1109/TCYB.2020.3003060

Song, B., Zeng, X., and Rodríguez-Patón, A. (2021). Monodirectional tissue P systems with channel states. *Inform. Sci.* 546, 206–219. doi: 10.1016/j.ins.2020.08.030

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009

Tang, Y., Tang, Y., Zhu, Y., Xiao, J., and Summers, R. M. (2020). "E2Net: an edge enhanced network for accurate liver and tumor segmentation on CT scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 512–522. doi: 10.1007/978-3-030-597 19-1_50

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). "Residual attention network for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6450–6458. doi: 10.1109/CVPR.2017.683

WHO (2014a). *World Cancer Report 2014. Chapter 1.1.*

WHO (2014b). *World Cancer Report 2014. Chapter 5.6.*

Wu, W., Zhou, Z., Wu, S., and Zhang, Y. (2016). Automatic liver segmentation on volumetric CT images using supervoxel-based graph cuts. *Comput. Math. Methods Med.* 2016:9093721. doi: 10.1155/2016/9093721

Yuan, Y. (2017). Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation. *arXiv [Preprint]. arXiv:1710.04540.*

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi: 10.1039/C9SC04336E

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., et al. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage* 108, 214–224. doi: 10.1016/j.neuroimage.2014.12.061

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 February 2021
doi: 10.3389/fgene.2021.608313

Exploring the Differential Expression and Prognostic Significance of the COL11A1 Gene in Human Colorectal Carcinoma: An Integrated Bioinformatics Approach

*Ritwik Patra, Nabarun Chandra Das and Suprabhat Mukherjee**

Integrative Biochemistry & Immunology Laboratory, Department of Animal Science, Kazi Nazrul University, Asansol, India

Colorectal cancer is one of the most common cancers of humans and the second highest in cancer-related death. Genes used as prognostic biomarkers play an imperative role in cancer detection and may direct the development of appropriate therapeutic strategies. Collagen type XI alpha 1 (COL11A1) is a minor fibrillary collagen that has an essential role in the regulation of cell division, differentiation, proliferation, migration, growth, and apoptosis of intestinal and colon cells. The present study seeks to evaluate the significance of the COL11A1 gene in the progression of colorectal cancer in humans across the various parameters using advanced bioinformatics approaches. The application of various databases and servers like ONCOMINE, UALCAN, and GEPIA were accessed for analyzing the differential expression of the COLL11A1 gene and its relative influence over the survival of the transformed subjects. In addition, oncogenomics of COL11A1 gene, mutations associated with this gene and interacting partners of the gene in the context of oncogenesis were studied using COSMIC, cBioPortal, GeneMANIA, and NetworkAnalyst. Our experimental data indicate that the COL11A1 gene is overexpressed in the transformed tissues across the various clinicopathological parameters reduces the probability of survival in both overall and disease-specific survival cases. Mutational studies imply that it can induce perturbations in various signaling pathways viz. RTK-RAS-PI3K, Wnt, TGF-β, and TP53 pathways influencing cancer development. Also, a positive association and correlation amongst the THBS2, COL10A1, COL5A2, and COL1A2 genes were observed, which most likely to contribute to the upregulation of carcinogenesis. Conclusively, this comprehensive study indicates the COL11A1 gene to be a significant contributor in the etiology of colorectal cancer, henceforth this gene can be considered as a prognostic biomarker for the conception of diagnostic and therapeutic strategies against colorectal cancer in the near future.

Keywords: bioinformatics, COL11A1 gene, colorectal cancer, mutation, prognosis, survival assay

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Biju Issac,
Leidos Biomedical Research, Inc.,
United States
Mamoon Rashid,
King Abdullah International Medical
Research Center (KAIMRC),
Saudi Arabia

***Correspondence:**
Suprabhat Mukherjee
suprabhat.mukherjee@knu.ac.in

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *20 September 2020*
Accepted: *14 January 2021*
Published: *01 February 2021*

Citation:
Patra R, Das NC and Mukherjee S
(2021) Exploring the Differential
Expression and Prognostic
Significance of the COL11A1 Gene
in Human Colorectal Carcinoma: An
Integrated Bioinformatics Approach.
Front. Genet. 12:608313.
doi: 10.3389/fgene.2021.608313

INTRODUCTION

Colorectal cancer is considered as the third most common cancer in the world and is in the second position for cancer-related death of humans worldwide (Siegel et al., 2017). It is a multi-stage process that gradually develops with the initiation of transformation in normal colon tissue to an adenomatous intermediate by the consequences of mutation, epigenetic changes, DNA damage, uncontrolled growth with gene and chromosomal instability as well as defects leading to invasive adenocarcinoma (Zhang et al., 2011). It is imperative to understand the appropriate mechanism of prognosis, pathogenesis, and genomic alterations associated with colorectal cancer for the development of appropriate therapeutic strategies.

The intestinal extracellular matrix (ECM) is majorly constituted of collagen and is vital for the regulation of cell division, differentiation, proliferation, migration, growth, and apoptosis which signify its cruciality across the development and progression of cancer (Fischer et al., 2001). Collagen type XI alpha 1 (COL11A1) is a minor fibrillary collagen protein, that represents one of the two alpha chains of type XI collagen. Mutations in the COL11A1 gene and/or translational overexpression of COL11A1 protein due to the signaling defects are considered as the essential contributors of carcinogenesis in human colorectal cancer (Raglow and Thomas, 2015). In this context, higher expression of COL11A1 protein has been reported in the cancerous tissue and has been found to be linked with poor progression-free and overall survival across the various types of cancers (Raglow and Thomas, 2015). A microarray-based study reveals that the COL11A1 gene is associated with the disease progression and poor survival in ovarian cancer and regulates cell invasiveness required for tumor formation (Wu et al., 2014). Further studies have also established that COL11A1 gene attributes as a prognostic biomarker for human carcinoma-associated stromal cells and also stimulates cancer progression in lungs, breast, gastrointestinal tract, and pancreas (García-Pravia et al., 2013; Vázquez-Villa et al., 2015; Shen et al., 2016; Li A. et al., 2017; Toss et al., 2019). All these reports collectively suggest that overexpression of COL11A1 in different cancerous tissues results in metastasis and recurrence of several human cancers (García-Pravia et al., 2013; Vázquez-Villa et al., 2015; Shen et al., 2016; Li A. et al., 2017; Toss et al., 2019). COL11A1 is a highly specific biomarker of activated cancer-associated fibroblasts (CAFs) which remains conserved for epithelial cancer irrespective of the site and transformation within the cell undergoing neoplastic transformation, indicating that targeting fibroblast activation could be an effective therapeutic strategy for various cancer (Jia et al., 2016). In an another study, the COL11A1 along with the other two genes viz. THBS2 and INHBA have been found to be overexpressed in colon tissue indicating invasion-facilitated alteration in proteolysis of the extracellular matrix and used for developing high specificity biomarkers sensing cancer invasion and determining response against potential multi-cancer metastasis and therapeutic target (Kim et al., 2010). Particularly for colorectal cancer, previous researchers revealed that the expression of the COL11A1 gene is upregulated up to several folds in the stromal cells of affected colonic mucosa in comparison to the normal tissue (Fischer et al., 2001). Studies on left-sided and right-sided colon cancer, it has been found that COL11A1, TWIST1, insulin-like 5, and chromogranin A were upregulated across the right-sided colon cancer more significantly than that of the left-sided cancer, with a sharp downregulation in 3β-hydroxysteroid dehydrogenase protein (Su et al., 2019). Several experiments on the transformed cells also display significant alteration in a number of cellular signaling pathways, including Wnt, TGF-β, RTK-RAS-PI3K, and TP53 signaling pathways which might be the crucial contributors of the neoplastic transformation (Li et al., 2015; Koveitypour et al., 2019). Although all these various studies imply that the COL11A1 gene is crucial in the progression of various cancer, however, the actual significance across the various clinicopathological factors including cancer-stage, nodal metastasis status, age group, etc., have not been documented comprehensively till date.

The mutations in the COL11A1 gene and resultant impact on the oncogenomic and metabolic pathways are indeed very much essential in understanding the etiology of human colorectal cancer and are yet unclear, thus it provides an area for new research in understanding the actual significance of the COL11A1 gene in the progression of colorectal carcinoma. Regarding this, the application of various bioinformatics tools using the huge dataset of well-established cancer data from different demographic and clinicopathologic patients provides a comprehensive area for further research and development of therapeutic strategies. Considering the background, the objective of the present study is to collectively examine the differential expression, survival, co-expression, correlation, mutations, and protein-protein interaction network that result in the alteration of various pathways related to the COL11A1 gene playing a key role in the transformation of human colon tissue to colorectal cancer using an integrated bioinformatics approach. In addition, our study also aggregates all the available discrete data to identify the significance of the COL11A1 gene as a prognosis biomarker for colorectal cancer which may be useful in designing future research for the conception of appropriate therapeutic strategies.

MATERIALS AND METHODS

Analysis of the Differential Expression of COL11A1 Gene Across Healthy and Transformed Colon Tissues

Differential expression of COL11A1 gene was studied to identify the expression pattern of the COL11A1 gene between tumor and normal tissues across all TCGA (The Cancer Genome Atlas) datasets was performed using TIMER 2.0[1]. It is a comprehensive online resource for systematic analysis of immune infiltrates and gene expression across diverse cancer types (Li T. et al., 2017; Li et al., 2020).

[1] http://timer.cistrome.org/

Next, the Oncomine server[2] was searched for human colorectal cancer and the differential gene analysis section (Cancer vs. Normal Analysis) was selected to retrieve the results. It is a publicly accessible cancer microarray database and web-based data mining platform, containing 715 datasets and 86,733 samples (Rhodes et al., 2004, 2007). The dataset selected for differential expression of mRNA include TCGA colorectal cancer and Kaiser Colon cancer, and recorded within a threshold value of *P*-value- 1E-4, fold change- 2, Gene rank- Top 10 and are shown in **Supplementary Table 1**.

Expression Profile and Correlation Analysis

The functional expression of COL11A1 gene in colon carcinoma is analyzed using UALCAN[3], a public server to analyze the cancer OMICS data (TCGA and MET500), built upon PERL-CGI with high-quality graphics through javascript and CSS to provide graphs and plots depicting gene expression, survival information, epigenetic regulation, and also correlation among gene (Chandrashekar et al., 2017). It is used here to analyze the expression and promoter methylations of the COL11A1 gene in colon adenocarcinoma based on clinicopathological features including sample type, individual cancer stage, patients' sex and age, histological subtype, nodal metastasis status, and TP53 mutation status and are listed in **Supplementary Tables 2,3**. The correlation of expression between the COL11A1 with THBS2, COL10A1, COL5A2, and COL1A2 genes for colon adenocarcinoma is performed using the GEPIA[4] and UCSC Xena[5] servers (Tang et al., 2017; Goldman et al., 2020).

Survival Assay of COL11A1 and Its Correlated Genes

The survival analysis for overall survival and disease-free survival is determined by generating Kaplan-Meier (KM) plot using the GEPIA server. It is a web server for analyzing the RNA sequencing expression data of 9,736 tumors and 8,587 normal samples from the TCGA and the GTEx projects (Tang et al., 2017). On the other hand, the KM-plot for disease-specific and overall survival of these genes in the TCGA COAD dataset is performed using the UCSC Xena server.

Oncogenomics and Mutational Study

cBioPortal[6] is an online server for exploration, visualization, and analysis of multidimensional cancer genomics data (Cerami et al., 2012). We use it to analyze the impact of the COL11A1 gene in the Colorectal Adenocarcinoma TCGA PanCancer dataset containing 594 samples. It provides a wide range of analysis tab within its server. The oncoprint demonstrates the overview of the COL11A1 gene across the dataset and also generate the heatmap of the correlated gene. Further using the mRNA expression data of the top 25 positively correlated genes, a clustered heatmap is generated using the delimited data on the Clustviz server[7]. The cancer type summary tab provides a detailed overview of the COL11A1 gene across the different subtypes of colorectal cancer i.e., mucinous adenocarcinoma of colon and rectum, colon adenocarcinoma, and rectal adenocarcinoma. It also shows the mutation of the COL11A1 gene for colorectal cancer and the mutational correlation within the associated gene set. The different types of mutations associated with the COL11A1 gene for colorectal cancer were analyzed using COSMIC-"Catalogue of Somatic Mutations in Cancer"[8] which is the world's largest source of expert for manually curated somatic mutation information related to human cancers (Tate et al., 2019).

Analysis for Pathways Associated With the COL11A1 Gene in Colorectal Carcinoma

We have explored PathwayMapper in the cBioPortal server shows the alteration frequencies of selected genes (COL11A1, THBS2, COL10A1, COL5A2, and COL1A1) along with the various pathways overlaid on a TCGA pathway using a white to a red color scale. Furthermore, the top 25 correlated genes belonging to the COL11A1 gene cluster were used to reveal the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways in Colorectal cancer using DAVID (Database for Annotation, Visualization and Integrated Discovery) available at https://david.ncifcrf.gov/.

Network and Enrichment Analysis

GeneMANIA[9] is a web-based platform to determine the association between the gene of interest with other genes using an extensive of functional association data. Herein, this platform was used to analyze the association of the COL11A1 gene with others genes, based on the protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity.

After screening, the top 25 significantly correlated gene along with the COL11A1 were used in NetworkAnalyst[10] for the enrichment analysis including Gene Ontology (GO) enrichment analysis, KEGG and Reactome pathways analysis, and to construct the protein-protein interaction at a generic level using International Molecular Exchange Consortium (IMEx) protein interactions database.

RESULTS

Expression of COL11A1 Gene Is Upregulated in Colorectal Cancer

The role of the COL11A1 gene in colorectal cancer is significantly upregulated in colorectal cancer (**Figure 1A**). The TIMER analysis reveals that the comparison of the COL11A1 gene across various cancer types including colon cancer and displays that it is significantly upregulated for colon adenocarcinoma

[2] https://www.oncomineorg/

[3] http://ualcan.path.uab.edu/

[4] http://gepia.cancer-pku.cn/

[5] https://xena.ucsc.edu/

[6] https://www.cbioportal.org

[7] https://biit.cs.ut.ee/clustvis/

[8] https://cancer.sanger.ac.uk/cosmic

[9] https://genemania.org

[10] https://www.networkanalyst.ca/

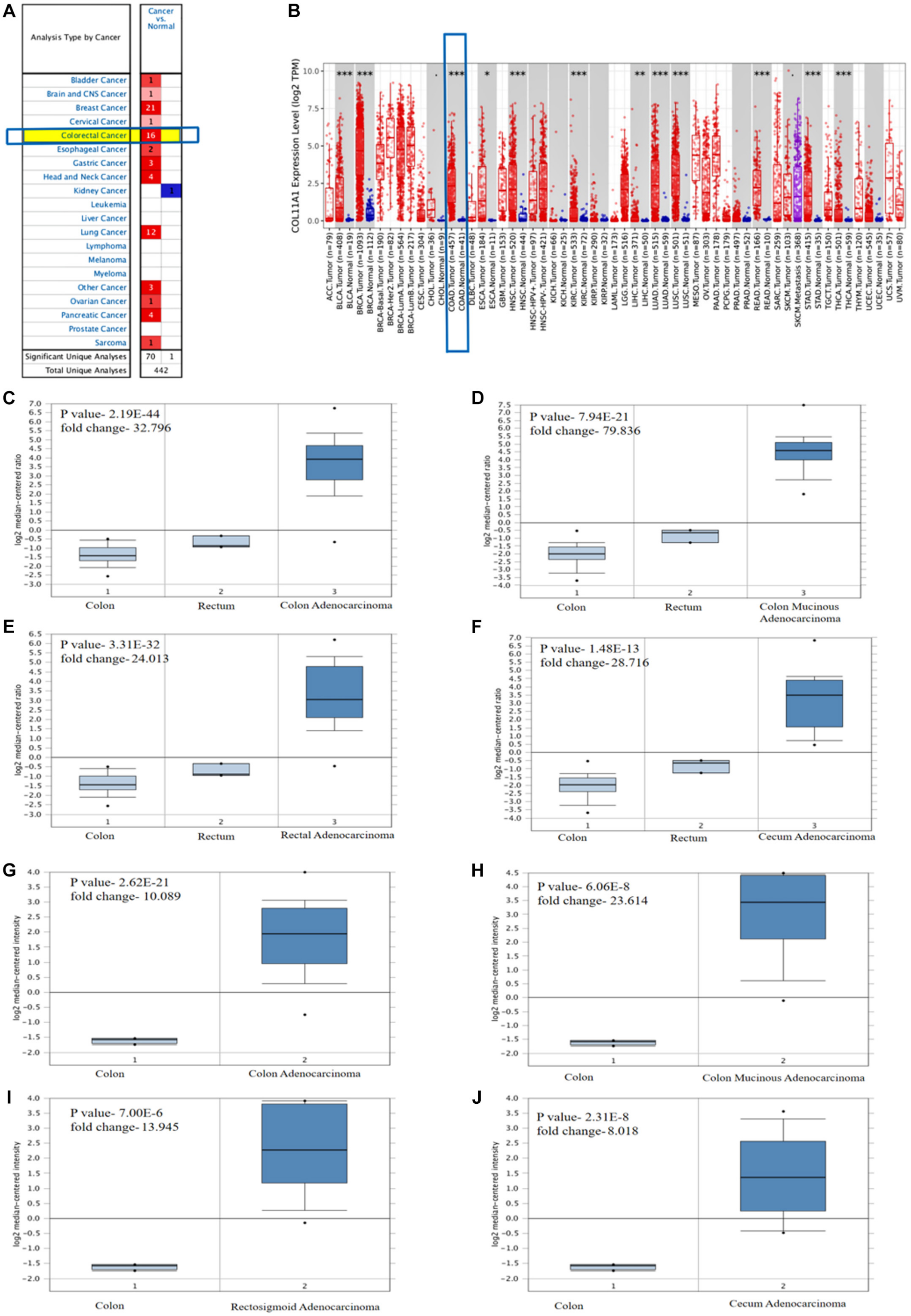


FIGURE 1 | Differential expression of COL11A1 gene **(A)** Expression of COL11A1 mRNA across different cancers where red and blue represent the upregulation and downregulation, respectively. **(B)** Comparative expression of COL11A1 mRNA between colon adenocarcinoma tumor tissue and normal tissue (statistical significance computed by differential analysis, **P* < 0.05; ***P* < 0.01; ****P* < 0.001). **(C–F)** Box plot comparison of COL11A1 expression for TCGA colorectal cancer dataset in panel **(C)**. Colon adenocarcinoma, **(D)** Colon Mucinous Adenocarcinoma, **(E)** Rectal Adenocarcinoma, **(F)** Cecum Adenocarcinoma. **(G–J)** Box plot comparison of COL11A1 expression for Kaiser colon cancer dataset in panel **(G)**. Colon adenocarcinoma, **(H)** Colon Mucinous Adenocarcinoma, **(I)** Rectosigmoid Adenocarcinoma, **(J)** Cecum Adenocarcinoma.

(**Figure 1B**). Further analyses of the mRNA expression profiles of the COL11A1 gene in normal and transformed tissue in ONCOMINE server reveal significant upregulation of COL11A1 mRNA in both the subtypes of cancer datasets i.e., TCGA colorectal cancer and Kaiser colon cancer (**Figures 1C–J** and **Supplementary Table 1**). It includes colon adenocarcinoma (*p*-value- 2.19E-44, fold change- 32.796), colon mucinous adenocarcinoma (*p*-value-7.94E-21, fold change- 79.836), rectal adenocarcinoma (*p*-value-3.31E-32, fold change- 24.013), and cecum adenocarcinoma (*p*-value-1.48E-13, fold change- 28.716) for TCGA colorectal cancer (**Figures 1C–F**), and is somehow greater than that of the Kaiser Colon cancer dataset (**Figures 1G–J**). All these data collectively indicate that human colorectal cancer samples display significantly higher expression of COL11A1 mRNA in comparison to normal colon and rectum tissues, indicating COL11A1 could have a crucial role in the neoplastic transformation of colorectal cancer.

Transcriptional Expression and Epigenetic Regulation of COL11A1 Across Various Clinicopathological Parameters

The expression of COL11A1 in colon adenocarcinoma was analyzed based on the different clinicopathological parameters like sample type, individual cancer stage, patient's sex and age, histological subtype, nodal metastasis status, and TP53 mutation status using the UALCAN server (**Figure 2** and **Supplementary Table 2**). The results support the inference depicted in the earlier section by demonstrating that COL11A1 expression is higher in the colorectal cancer tissue at different clinical stages than in normal tissue (**Figure 2A**). It tends to increase the expression of COL11A1 at advanced stages of cancer (Stage 3 > Stage 2 > Stage 1) (**Figure 2B**) and decrease along with the increase in the age group of patients (**Figure 2C**). It was also found that the expression of the COL11A1 gene increases along with the nodal metastasis status (N2 > N1 > N0) (**Figure 2E**).

DNA methylation is relatively associated with the development of cancer within the human body (Greenberg and Bourc'his, 2019). From our data, it was evident that the promoter methylation of the COL11A1 gene is overexpressed in the colon cancer tissue than that of the normal tissue, and is negatively regulated for all other clinicopathological parameters (**Figures 2G–L** and **Supplementary Table 3**). It is reflected that along with the development of cancer stages and nodal metastasis status, the expression of promoter methylation decreases in the tissues (Stage 1 > Stage 2 > Stage 3; N0 > N1 > N2) (**Figures 2H,K**). These results indicate that the promoter methylation is negatively associated with the expression of COL11A1 mRNA, and the hypermethylation of the promoter of COL11A1 may inhibit COL11A1 in upgrading cancer development.

Survival Assay of the COL11A1 Gene in Colorectal Cancer

Survival analysis is one of the key components in analyzing the influence of any cancer-associated gene (Clark et al., 2003). In this study, the survival assay of the COL11A1 gene is explained by the KM-plots which show a reciprocal correlation between the expression of COL11A1 and overall survival (log-rank p-0.055) or disease-free survival (log-rank p-0.053), which signifies the COL11A1 gene as a poor prognostic indicator for colorectal cancer (**Figures 3A,B**). Also, the disease-specific survival plot of COL11A1, obtained from the UCSC XENA server indicates that higher expression leads to lower survival probability (*p*-value- 0.1059) (**Figure 3C**). Therefore, low COL11A1 expression in colorectal cancer patients is correlated with prolonged survival, but high COL11A1 expression in colorectal cancer is associated with poor survival.

The survival assay of the correlated genes shows similar significance to that of the COL11A1 gene in the colon adenocarcinoma dataset. The KM-plot obtained for overall survival at higher expression of THBS2 (*p*-value- 0.021), COL10A1 (*p*-value- 0.129), COL5A2 (*p*-value- 0.714), and COL1A2 (*p*-value- 0.221) is related with lower survival probability (**Figures 3D–G**). Similarly, the disease-specific survival is also decrease with the increase in expression of THBS2 (*p*-value- 0.015), COL10A1 (*p*-value- 0.126), COL5A2 (*p*-value- 0.925), and COL1A2 (*p*-value- 0.602) (**Figures 3H–K**).

Co-expression and Correlation Amongst the Other Genes Associated With COL11A1 in Colorectal Cancer

The top 25 positively co-expressed genes were analyzed via cBioPortal, containing the Spearman's correlation coefficient, *p*-value from two-sided *t*-test, and also *q*-value derived from the Benjamini-Hochberg FDR correction procedure (**Supplementary Table 4**). Further mRNA expression data was used for generating a clustered heatmap showing expression between +3/−3 with mean-centered to 0 (**Figure 4A**). From these above two analyses, it is found that the co-expression of THBS2, COL10A1, COL5A2, and COL1A2 is most likely to be positively correlated with the COL11A1 gene in colorectal cancer (**Table 1**). To further validate the co-expression, another heatmap was generated using UCSC XENA server to correlate the gene expression of the associated genes with respect to the COL11A1 gene, represented as a histogram with the z score transformation (**Supplementary Figure 1**). Moreover, correlation graph was obtained using the Pearson's correlation coefficient amongst COL11A1 gene with THBS2 (*R*-value- 0.90), COL10A1 (*R*-value- 0.89), COL5A2 (*R*-value- 0.69) and COL1A2 (*R*-value- 0.65) (**Figures 4B–E**). Collectively all these results reveal that the COL11A1 gene has a positive association and correlation with THBS2, COL10A1, COL5A2, and COL1A2 to upregulate the gene expression to induce the development of colorectal cancer.

Genomic Alteration and Mutation Associated With COL11A1 Gene in Colorectal Cancer

The COL11A1 gene mutation was analyzed on COSMIC database comprising more than 2406 samples of colorectal cancer out of which 249 were recorded for mutations, among them the missense substitution is highest with 51.81% followed

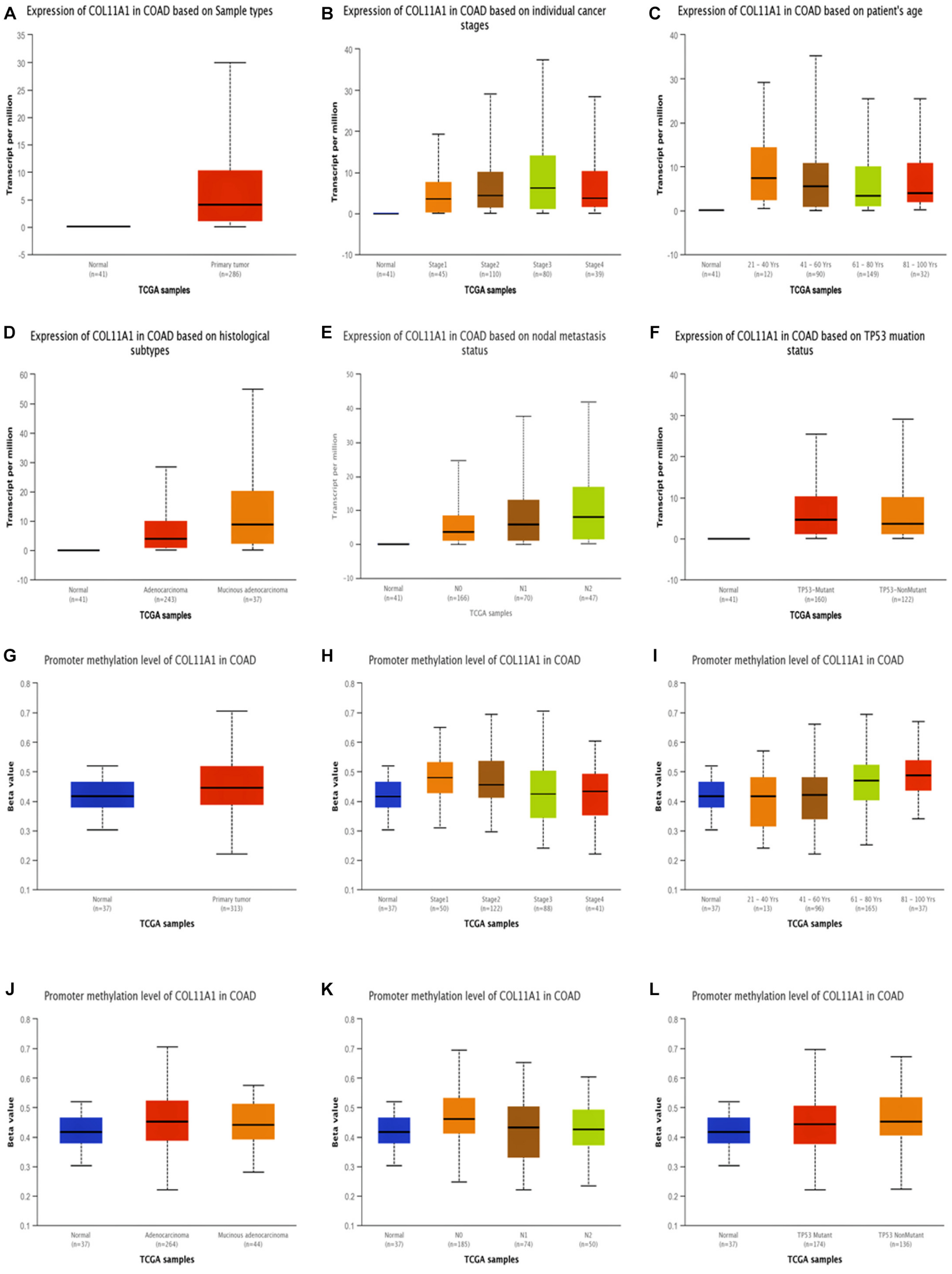


FIGURE 2 | Expression and promoter methylation of the COL11A1 gene in colon adenocarcinoma for different clinicopathological parameters. **(A–F)** Box-plot showing relative expression of COL11A1 mRNA in panel **(A)**. cancer tissues and normal tissues, **(B)** individual cancer stage, **(C)** patient's age, **(D)** histological subtypes, **(E)** nodal metastasis status, **(F)** TP53 mutation status. **(G–L)** Box-plot showing promoter methylation of COL11A1 mRNA in, **(G)** cancer tissues and normal tissues, **(H)** individual cancer stage, **(I)** patient's age, **(J)** histological subtypes, **(K)** nodal metastasis status, **(L)** TP53 mutation status.

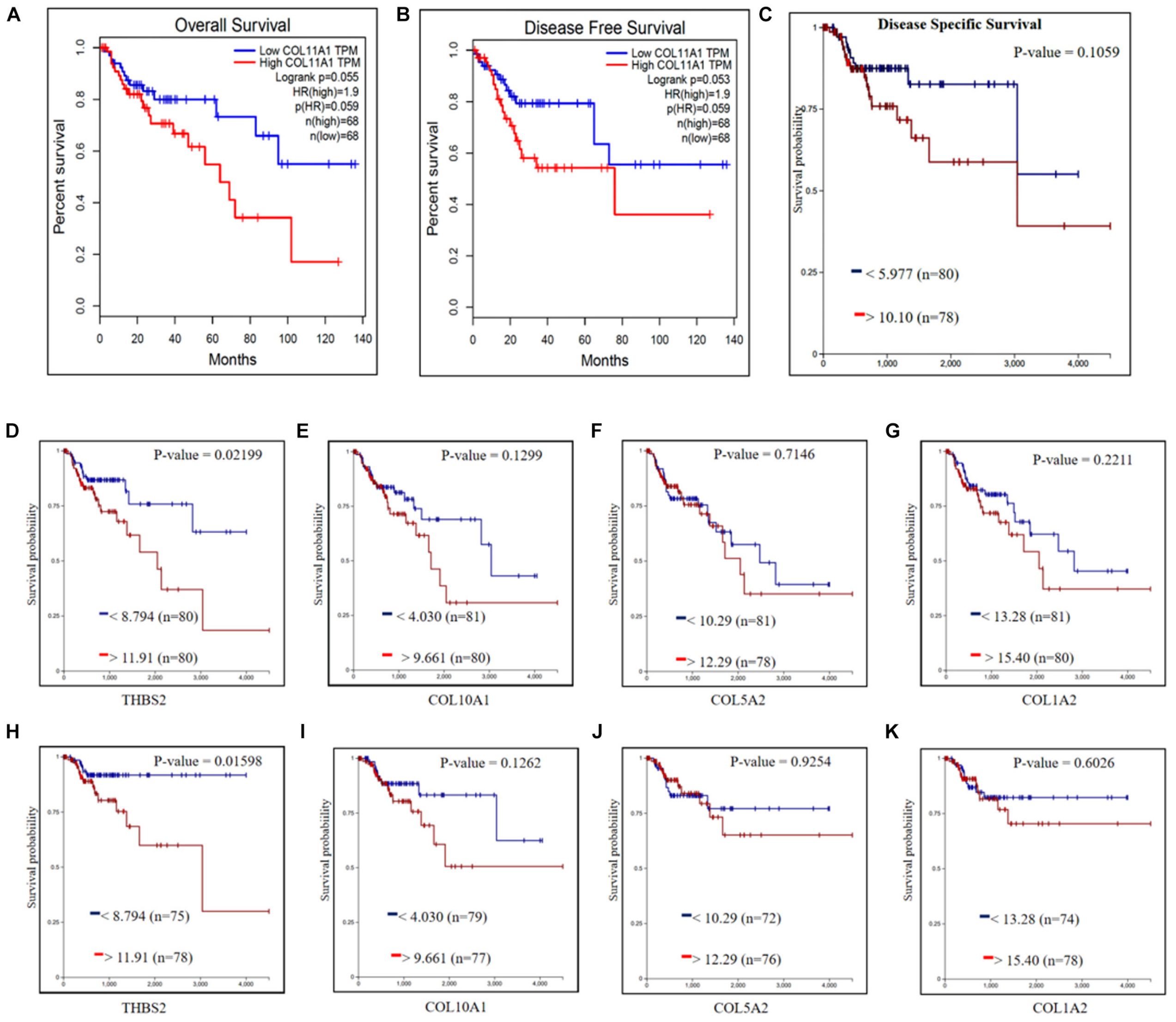


FIGURE 3 | KM-plot for survival assay of COL11A1 and other associated genes. **(A–C)** Effect of COL11A1 expression on **(A)** overall survival, **(B)** Disease-free, **(C)** Disease-specific survival. **(D–G)** Overall survival concerning gene expression of **(D)** THBS2, **(E)** COL10A1, **(F)** COL5A2, **(G)** COL1A2. **(H–K)** Disease-specific survival for the gene expression of **(H)** THBS2, **(I)** COL10A1, **(J)** COL5A2, **(K)** COL1A2. (Red and blue indicate the higher and lower expression of the gene, respectively).

by synonymous substitution (15.66%), frameshift mutation (15.66%), nonsense substitution (4.42%) and other types (4.02%) (**Figure 5A**). The breakdown of various substitution mutation is shown in **Figure 5B**, representing the highest type of G > A (25.73%) and lowest showing T > A (0.58%). To determine and analyze the frequency and type of mutation, cBioPortal server was used where the cancer type summary indicates the mutation along with the various subtypes of colorectal cancer showing mucinous adenocarcinoma of colon and rectum (>12%), colon adenocarcinoma (<12%), and rectal adenocarcinoma (∼6%) (**Figure 5C**). The Oncoprint and Mutation tab shows that the COL11A1 gene is altered in 10% of the total 526 patients in TCGA colorectal cancer dataset along with the heatmap for the associated genes (**Figure 5D**). Additionally, a mutational study for the correlation among the COL11A1 gene with THBS2, COL10A1, COL5A2, and COL1A2 (**Figures 5F–I**) showing a significant coefficient value for both Spearman and Pearson Correlation test and the regression line. It is observed that the mutation of COL11A1 is much more expressive for COL1A2 > COL5A2 > THBS2 > COL10A1.

Gene Network and Pathways Alteration

GeneMANIA server provides a complete network of COL11A1 gene with its neighboring gene of interaction in colorectal cancer displaying the physical interactions (67.64%), coexpression (13.50%), predicted (6.35%), co-localization (6.17%), pathways (4.35%), genetic interaction (1.40%), and shared protein domains (0.59%) (**Figure 6A**). The Gene Ontology (GO) enrichment

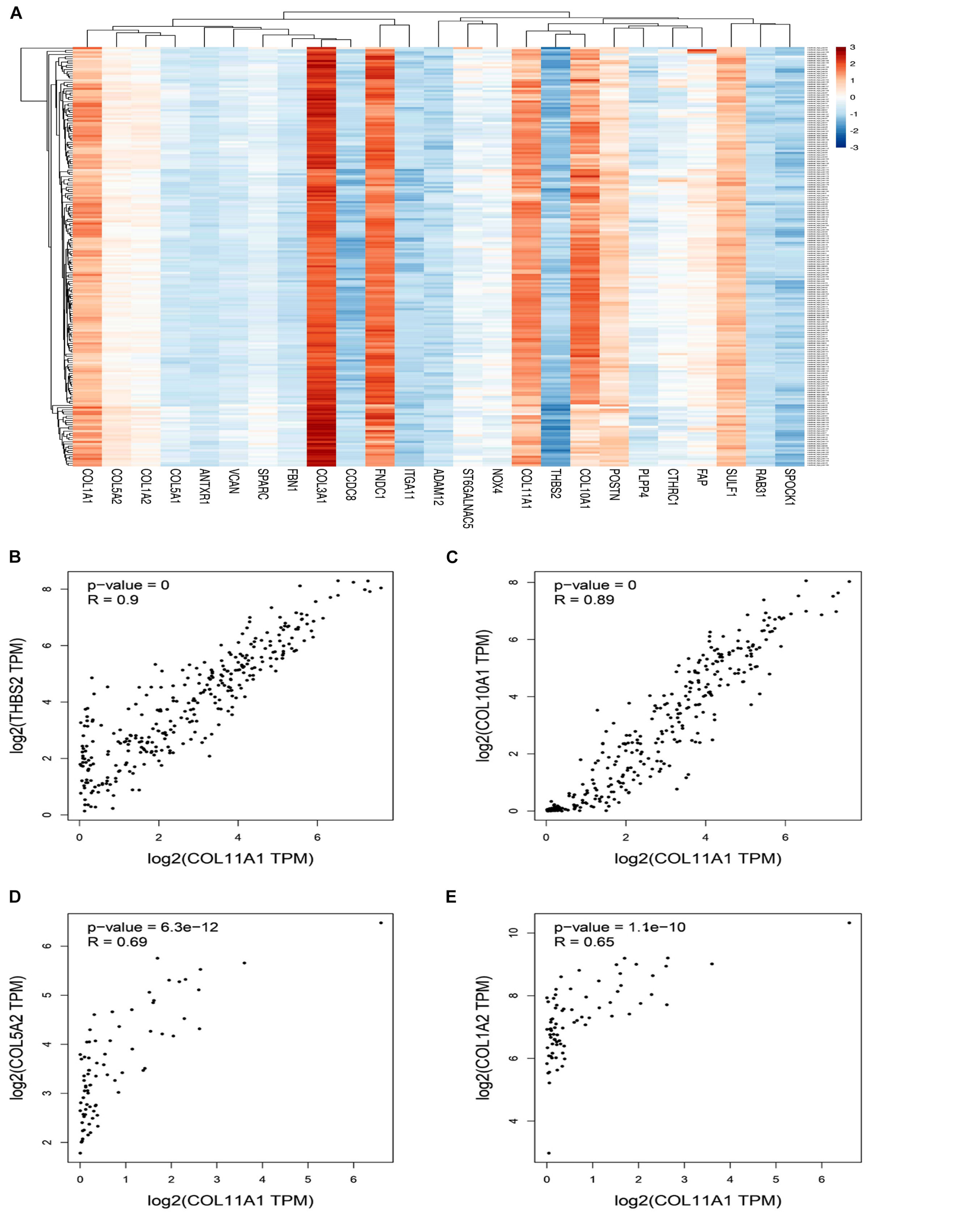


FIGURE 4 | Coexpression and correlation of genes functionally associated with COL11A1. **(A)** Clustered heatmap of the top 25 correlated genes (Scaling in –3/3 with mean-centered to 0). **(B–E)** Graphical representation of Pearson's correlation test of COL11A1 gene with, **(B)** THBS2, **(C)** COL10A1, **(D)** COL5A2, **(E)** COL1A2.

analysis was performed on NetworkAnalyst to obtain the network of GO: biological pathway (**Figure 6C**), and molecular function (**Figure 6D**) showing the significance of the genes in extracellular structure organization, collagen fibril organization, protein complex subunit organization, collagen metabolic process, cell migration, etc., and are listed in **Supplementary Table 5**. It was

TABLE 1 | The correlation among the genes associated with COL11A1 gene in colorectal cancer.

Correlated Gene	Spearman's correlation	Pearson's correlation
THBS2	0.922	0.90
COL10A1	0.913	0.89
COL5A2	0.909	0.69
COL1A2	0.903	0.65

further used to generate the network for Reactome (**Figure 6E**) and KEGG pathway (**Figure 6F**) analysis. Moreover, the protein-protein interaction (PPI) network was constructed based on the International Molecular Exchange Consortium (IMEx) protein interactions database using NetworkAnalyst represented the crucial protein and helps to further establish the genes promoting in colorectal cancer prognosis and development. As shown in the PPI network (**Figure 6B**), the degree of a node is the number of connections among the node, and betweenness is the smallest path amongst nodes showing RAB31 (Degree:19, Betweeness:2401.3), COL1A1 (Degree:36, Betweeness:5357.51), COL1A2 (Degree:26, Betweeness:2955.04), COL3A1 (Degree:9, Betweeness:657.29), COL11A1 (Degree:7, Betweeness:537.23), and VCAN (Degree:18, Betweeness:2701.11) as the important proteins of the network.

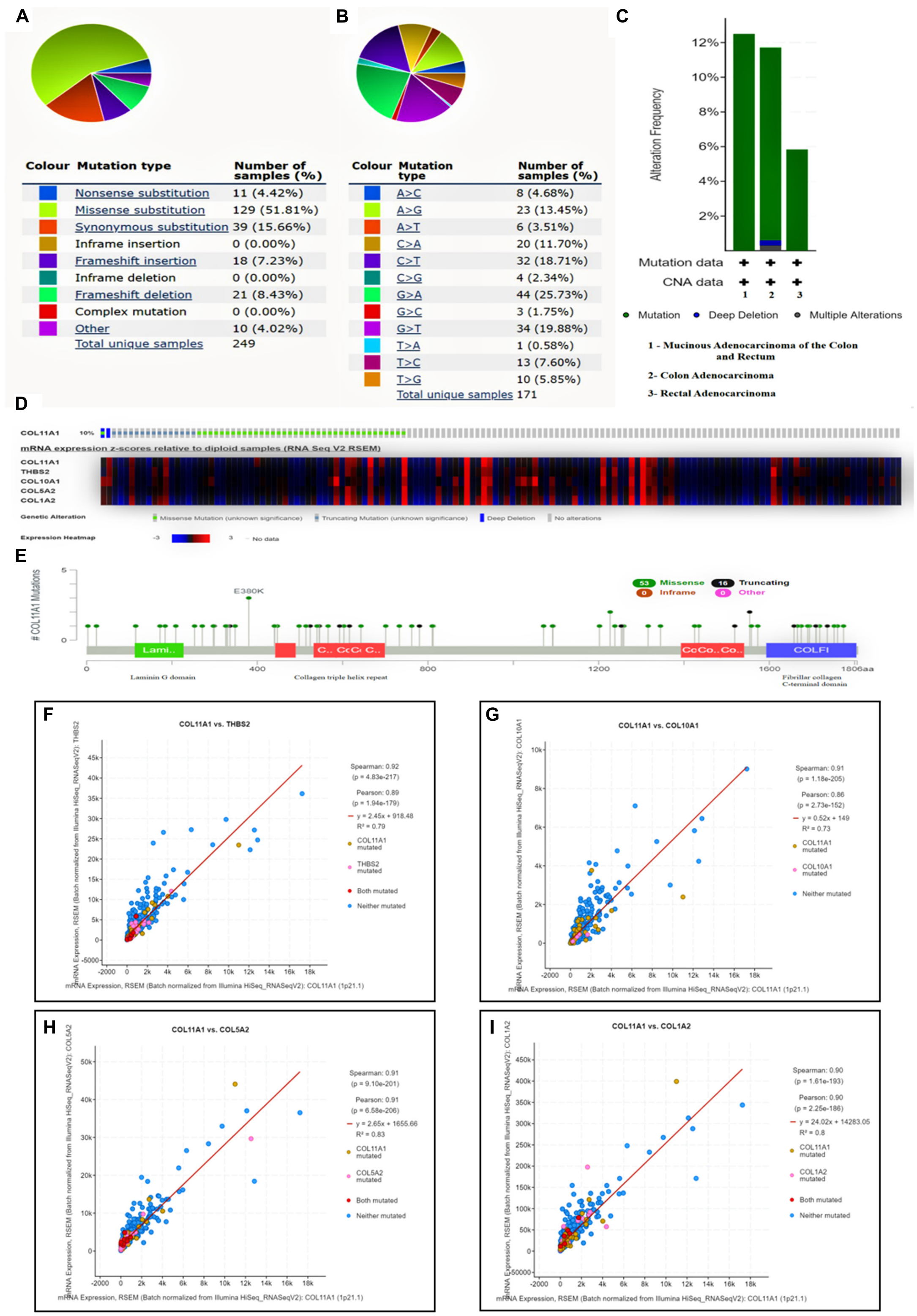
The KEGG pathways established from the DAVID analysis indicate the intervention of COL11A1 and associated genes in the ECM-receptor interaction (**Supplementary Figure 2A**), Protein digestion and absorption (**Supplementary Figure 2B**), Focal-adhesion (**Supplementary Figure 2C**), and PI3K-Akt signaling pathway (**Supplementary Figure 2D**), and are listed in **Table 2**. The PathwayMapper tab in cBioPortal servers shows the alteration frequency of COL11A1, THBS2, COL10A1, COL5A2, and COL1A2 over the various pathways on the colorectal cancer dataset using a white to red color scale where the more frequently altering gene shows greater intensity of the red color (**Figures 7A–D**). COL11A1 associated alteration mainly induces changes of PTEN (8.1%), PIK3CA (24.8%), KRAS (37.4%), and BRAF (10.8%) for regulation of RTK-RAS-PI3K signaling pathway (**Figure 7A**); APC (66.7%) in regulation of Wnt signaling pathway (**Figure 7B**); SMAD4 (15.5%) for TGF-β signaling pathway (**Figure 7C**); and ATM (12.5%) and TP53 (53.0%) in alteration of TP53 pathway (**Figure 7D**) to proliferate the cancer development.

DISCUSSION

In this modern era, the change in lifestyle, food habits, consumption of carcinogens, and several altered environmental factors are collectively considered as the major concerns of colorectal cancer and related deaths. The functional association amongst the various genetic and epigenetic processes are known to play a remarkable role in the initiation and progression of colorectal cancer (Pancione et al., 2012). In particular, overexpression and differentiation of ECM molecules, including collagen in the intestine, are considered as the key determinants of the proliferation and development of colorectal cancer (Fischer et al., 2001). The COL11A1 gene is a minor fibrillary collagen and plays an essential role in the fibrillogenesis and skeletal morphogenesis by controlling the lateral growth, and interfibrillar spacing of collagen II fibrils (Brown et al., 2011). Hitherto, studies available in the literatures and databases provide discrete evidences on the regulation of COL11A1 gene expression in the onset of various types of carcinomas (Vázquez-Villa et al., 2015; Toss et al., 2019). Our present study is a maiden attempt to provide a comprehensive knowledge of the various clinical relevance of the COL11A1 gene in the expression profile, methylation, survivability, and mutation in association with the colorectal cancer.

The mRNA expression profile of COL11A1 gene obtained from TCGA dataset of colorectal cancer from the various servers like ONCOMINE, UALCAN, and GEPIA collectively discloses significant upregulations at transcriptional level in cancer tissue than the normal colon tissue across various cancer subtypes including colon adenocarcinoma, colon mucinous adenocarcinoma, rectal adenocarcinoma, and cecum adenocarcinoma (**Figures 1C–J**); and even in the various clinicopathological parameters including patients' age, cancer stage, nodal metastasis status, and TP53 mutation (**Figures 2A–F**). Epigenetic changes in the gene are known to be the leading causes of neoplastic transformation, and regarding this, our result on the promoter methylation of the COL11A1 gene across various parameters indicates negative relation with the expression profile in a way suggesting the hypermethylation of the COL11A1 gene may regulate the of development cancer (**Figures 2G–L**). The KM-plots obtained for the overall survival (**Figure 3A**) and disease-free survival (**Figure 3B**) show poor prognosis of colorectal cancer i.e., the higher expression of the COL11A1 gene signifies poor survivability. The coexpression and correlation of the top 25 positively correlated genes with the COL11A1 gene are depicted on the heatmap (**Figure 4A**). Herein, we have found that THBS2, COL10A1, COL5A2, and COL1A2 are the most significant gene having the highest positive correlation (**Supplementary Figure 1** and **Figures 4B–E**). Further, upon the survival assays of THBS2, COL10A1, COL5A2, and COL1A2 genes it has been found a similar pattern of lower survival probability on overexpression (**Figures 3D–K**). Collectively all these experimental data clearly reveal that the COL11A1 gene along with its associated THBS2, COL10A1, COL5A2, and COL1A2 might serve as a prognostic biomarker for colorectal cancer.

The genomic alteration and mutation are the major inducers for the initiation and development of several cancers (Loeb et al., 2008). In our study, it has been observed that up to 12% mutation that relates to the COL11A1 gene contributes to the development of colorectal cancer with the highest alteration in mucinous adenocarcinoma of colon and rectum (**Figure 5C**). Further analysis from the COSMIC server illustrates that substitution mutation is the most prevalent mutation that constitutes the highest frequency of G > A types of changes (**Figures 5A,B**). In addition, the prevalence of THBS2, COL10A1, COL5A2, and COL1A2 enhances the frequency of alteration and depicting a positive correlation with COL11A1 mutation (**Figures 5F–I**). A functional network of the interaction among the



Colour	Mutation type	Number of samples (%)
🟦	Nonsense substitution	11 (4.42%)
🟩	Missense substitution	129 (51.81%)
🟧	Synonymous substitution	39 (15.66%)
🟨	Inframe insertion	0 (0.00%)
🟪	Frameshift insertion	18 (7.23%)
🟦	Inframe deletion	0 (0.00%)
🟩	Frameshift deletion	21 (8.43%)
🟥	Complex mutation	0 (0.00%)
🟪	Other	10 (4.02%)
	Total unique samples	249

Colour	Mutation type	Number of samples (%)
🟦	A>C	8 (4.68%)
🟩	A>G	23 (13.45%)
🟧	A>T	6 (3.51%)
🟨	C>A	20 (11.70%)
🟪	C>T	32 (18.71%)
🟦	C>G	4 (2.34%)
🟩	G>A	44 (25.73%)
🟥	G>C	3 (1.75%)
🟪	G>T	34 (19.88%)
🟦	T>A	1 (0.58%)
🟪	T>C	13 (7.60%)
🟧	T>G	10 (5.85%)
	Total unique samples	171

A
B
C
Alteration Frequency
12%
10%
8%
6%
4%
2%
Mutation data + + +
CNA data + + +
1 2 3
● Mutation ● Deep Deletion ● Multiple Alterations
1 - Mucinous Adenocarcinoma of the Colon and Rectum
2- Colon Adenocarcinoma
3- Rectal Adenocarcinoma
D
COL11A1 10%
mRNA expression z-scores relative to diploid samples (RNA Seq V2 RSEM)
COL11A1
THBS2
COL10A1
COL5A2
COL1A2
Genetic Alteration
Missense Mutation (unknown significance)
Truncating Mutation (unknown significance)
Deep Deletion
No alterations
Expression Heatmap
-3
3
No data
E
# COL11A1 Mutations
5
0
E380K
53 Missense
16 Truncating
0 Inframe
0 Other
Lami..
C.. CcCc C..
CcCo..Co..
COLFI
0
400
800
1200
1600
1806aa
Laminin G domain
Collagen triple helix repeat
Fibrillar collagen C-terminal domain
F
COL11A1 vs. THBS2
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): THBS2
45k
40k
35k
30k
25k
20k
15k
10k
5k
0
-5000
-2000 0 2k 4k 6k 8k 10k 12k 14k 16k 18k
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL11A1 (1p21.1)
Spearman: 0.92
(p = 4.83e-217)
Pearson: 0.89
(p = 1.94e-179)
y = 2.45x + 918.48
R² = 0.79
COL11A1 mutated
THBS2 mutated
Both mutated
Neither mutated
G
COL11A1 vs. COL10A1
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL10A1
10k
8k
6k
4k
2k
0
-2000 0 2k 4k 6k 8k 10k 12k 14k 16k 18k
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL11A1 (1p21.1)
Spearman: 0.91
(p = 1.18e-205)
Pearson: 0.86
(p = 2.73e-152)
y = 0.52x + 149
R² = 0.73
COL11A1 mutated
COL10A1 mutated
Neither mutated
H
COL11A1 vs. COL5A2
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL5A2
50k
40k
30k
20k
10k
0
-2000 0 2k 4k 6k 8k 10k 12k 14k 16k 18k
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL11A1 (1p21.1)
Spearman: 0.91
(p = 9.10e-201)
Pearson: 0.91
(p = 6.58e-206)
y = 2.65x + 1655.66
R² = 0.83
COL11A1 mutated
COL5A2 mutated
Both mutated
Neither mutated
I
COL11A1 vs. COL1A2
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL1A2
450k
400k
350k
300k
250k
200k
150k
100k
50k
0
-50000
-2000 0 2k 4k 6k 8k 10k 12k 14k 16k 18k
mRNA Expression, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2): COL11A1 (1p21.1)
Spearman: 0.90
(p = 1.61e-193)
Pearson: 0.90
(p = 2.25e-186)
y = 24.02x + 14283.05
R² = 0.8
COL11A1 mutated
COL1A2 mutated
Both mutated
Neither mutated

FIGURE 5 | Mutational analysis of COL11A1 gene. **(A)** Summary of various types of mutations associated with COL11A1 gene. **(B)** Bar-graph depicting various types of substitutional mutation occurring within the gene. **(C)** Mutation along the subtype of cancer including mucinous adenocarcinoma of colon and rectum (>12%), colon adenocarcinoma (<12%), and rectal adenocarcinoma (~6%). **(D)** Oncoprint showing mutational rate of COL11A1 gene and the heatmap for mRNA expression of associated genes. **(E)** Genomic information of COL11A1 mutation. **(F–I)** Graphical representation of correlation between COL11A1 gene showing mutation, Pearson correlation coefficient, Spearman correlation coefficient and regression line with, **(F)** THBS2, **(G)** COL11A1, **(H)** COL5A2, **(I)** COL1A2.

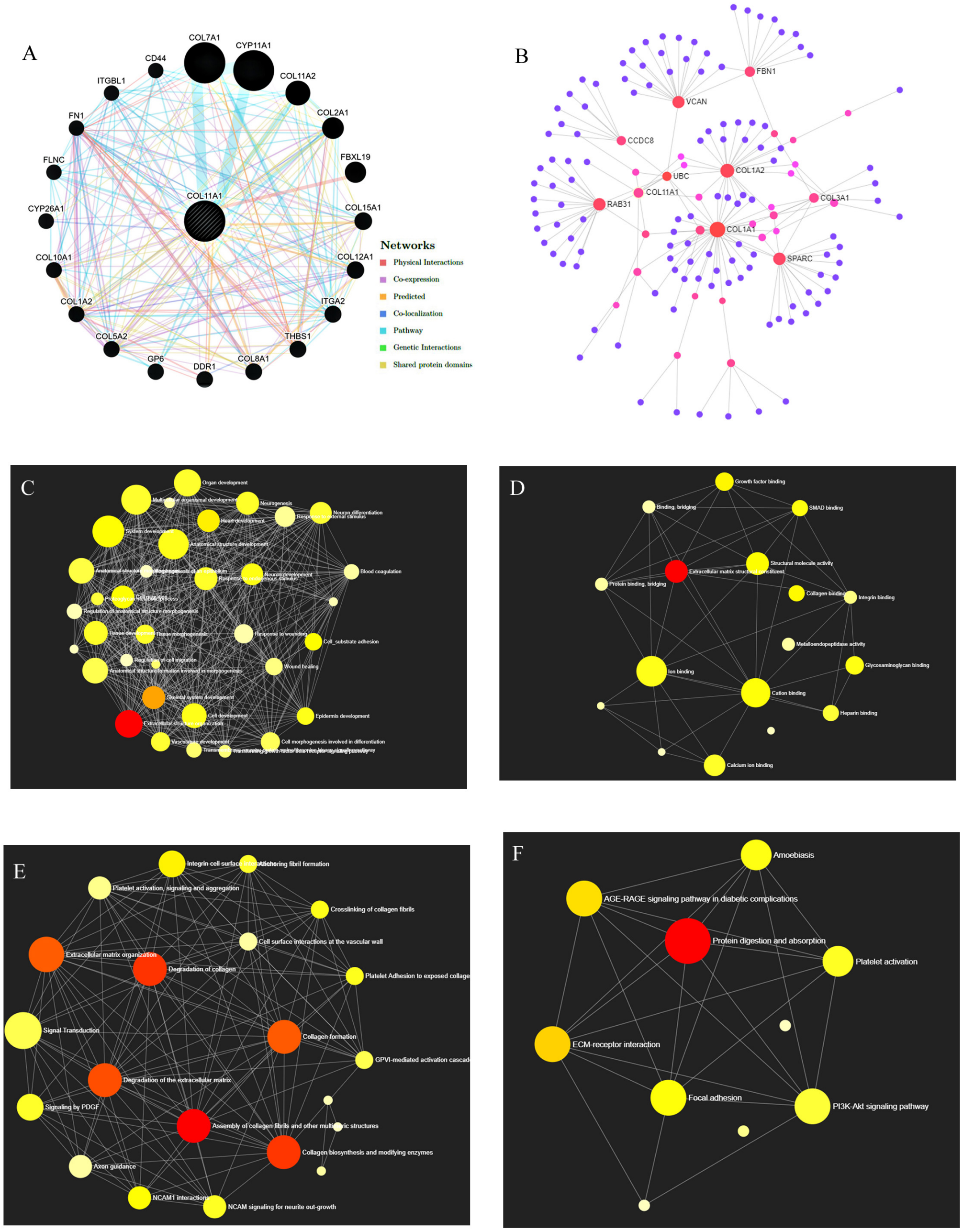

A
COL7A1
CYP11A1
CD44
COL11A2
ITGBL1
COL2A1
FN1
FBXL19
FLNC
COL11A1
CYP26A1
COL15A1
COL10A1
COL12A1
COL1A2
ITGA2
COL5A2
THBS1
GP6
COL8A1
DDR1
Networks
Physical Interactions
Co-expression
Predicted
Co-localization
Pathway
Genetic Interactions
Shared protein domains
B
FBN1
VCAN
CCDC8
COL1A2
UBC
COL11A1
RAB31
COL3A1
COL1A1
SPARC
C
Organ development
Multicellular organismal development
Neurogenesis
Neuron differentiation
Response to external stimulus
Heart development
System development
Anatomical structure development
Anatomical structure morphogenesis
Morphogenesis of an epithelium
Response to endogenous stimulus
Neuron development
Blood coagulation
Cell migration
Proteoglycan metabolic process
Regulation of anatomical structure morphogenesis
Tissue development
Tissue morphogenesis
Response to wounding
Cell_substrate adhesion
Regulation of cell migration
Anatomical structure formation involved in morphogenesis
Wound healing
Skeletal system development
Cell development
Extracellular structure organization
Epidermis development
Vasculature development
Cell morphogenesis involved in differentiation
Transmembrane receptor protein serine/threonine kinase signaling pathway
Transforming growth factor beta receptor signaling pathway
D
Growth factor binding
Binding, bridging
SMAD binding
Structural molecule activity
Extracellular matrix structural constituent
Protein binding, bridging
Collagen binding
Integrin binding
Metalloendopeptidase activity
Glycosaminoglycan binding
Ion binding
Cation binding
Heparin binding
Calcium ion binding
E
Integrin cell surface interactions
Anchoring fibril formation
Platelet activation, signaling and aggregation
Crosslinking of collagen fibrils
Cell surface interactions at the vascular wall
Extracellular matrix organization
Degradation of collagen
Platelet Adhesion to exposed collagen
Signal Transduction
Collagen formation
GPVI-mediated activation cascade
Degradation of the extracellular matrix
Signaling by PDGF
Assembly of collagen fibrils and other multimeric structures
Collagen biosynthesis and modifying enzymes
Axon guidance
NCAM1 interactions
NCAM signaling for neurite out-growth
F
Amoebiasis
AGE-RAGE signaling pathway in diabetic complications
Protein digestion and absorption
Platelet activation
ECM-receptor interaction
Focal adhesion
PI3K-Akt signaling pathway

FIGURE 6 | Gene Network Analysis. **(A)** COL11A1 gene with its neighboring genes showing physical interactions (67.64%), coexpression (13.50%), predicted (6.35%), co-localization (6.17%), pathways (4.35%), genetic interaction (1.40%), and shared protein domains (0.59%) **(B)** Protein-protein interaction network based on IMEx protein interactions database. **(C–F)** Network enrichment analysis for **(C)** GO: Biological process, **(D)** GO: Molecular function, **(E)** Reactome pathways, and **(F)** KEGG pathways.

TABLE 2 | KEGG pathways analysis using the DAVID server for top 25 correlated genes of COL11A1 in Colorectal cancer.

Pathways	Gene count	Percentage	Fold enrichment	*p*-value	*q*-value
ECM-receptor interaction	8	30.8	52.7	1.3E-11	2.1E-10
Protein digestion and absorption	7	26.9	45.6	1.6E-9	1.3E-8
Focal-adhesion	8	30.8	22.3	5.8E-9	3.1E-8
PI3K-Akt signaling pathway	8	30.8	13.3	2.1E-7	8.3E-7

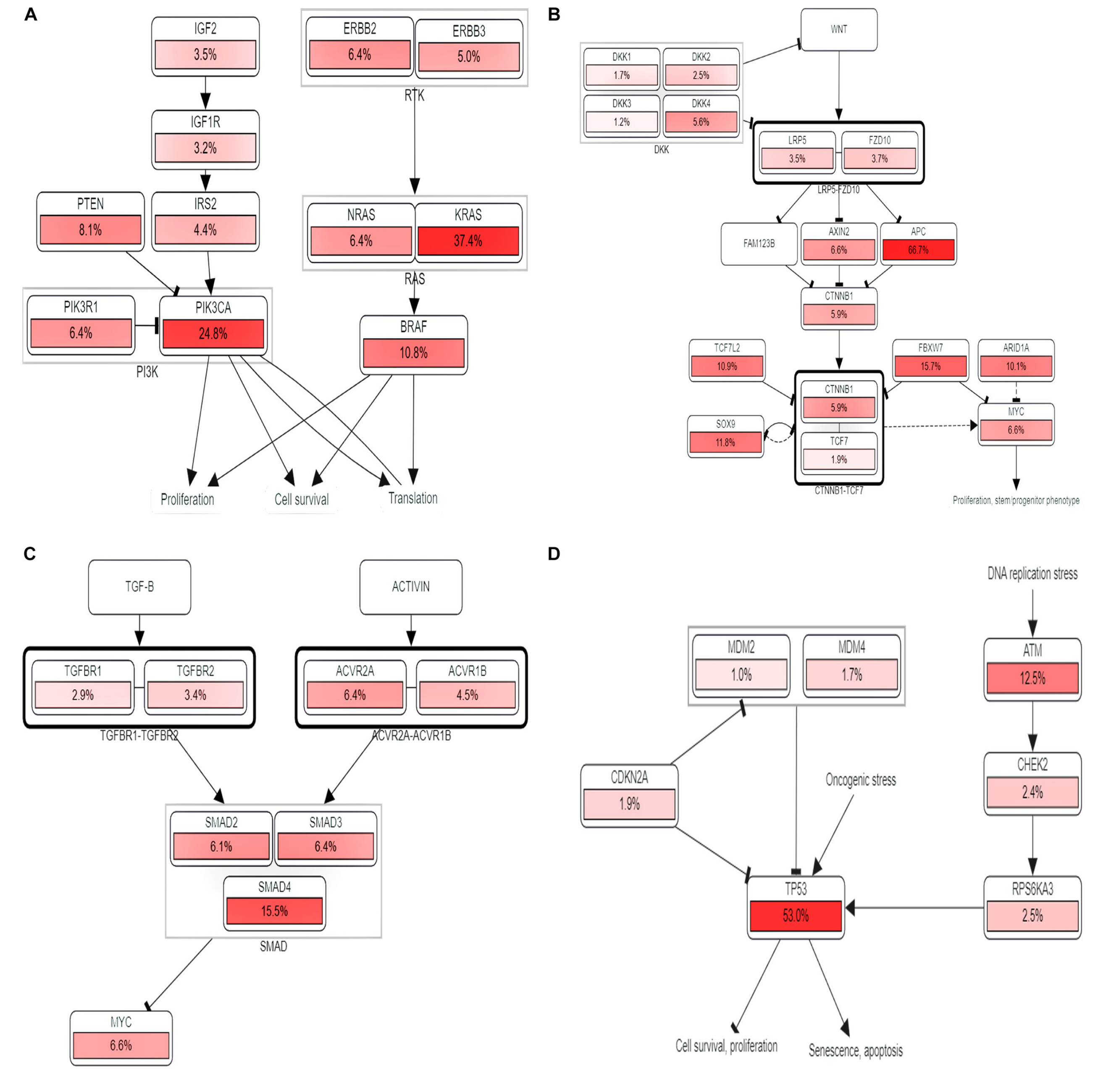


FIGURE 7 | Pathway analysis. **(A–D)** Impact of COL11A1 and associated gene in regulating alteration frequency of **(A)** RTK-RAS-PI3K signaling pathway, **(B)** Wnt signaling pathway, **(C)** TGF-β signaling pathway, **(D)** TP53 pathway.

neighboring genes of the COL11A1 in colorectal cancer displays physical interactions, co-expression, predicted co-localization, pathways, genetic interaction, and shared protein domains (**Figure 6A**). Collectively, we can postulate that the COL11A1 gene interacts with the neighboring mediators to induce the downregulation of various biological signaling pathways. Enrichment network created through NetworkAnalyst shows GO enrichment of various biological and molecular pathways where the genes significantly associated with extracellular structure organization, collagen fibril organization, protein complex subunit organization, collagen metabolic process, and cell migration (**Figures 6C,D** and **Supplementary Table 5**). The PPI networks indicate the RAB31, COL1A1, COL1A2, COL3A1, COL11A1, and VCAN as the most important protein network that likely to be connected and show betweenness among themselves to significantly promote the prognosis of colorectal cancer (**Figure 6B**).

The KEGG pathways established from DAVID analysis reveals that it shows the highest intimacy with the ECM-receptor interaction and PI3K-Akt signaling pathway

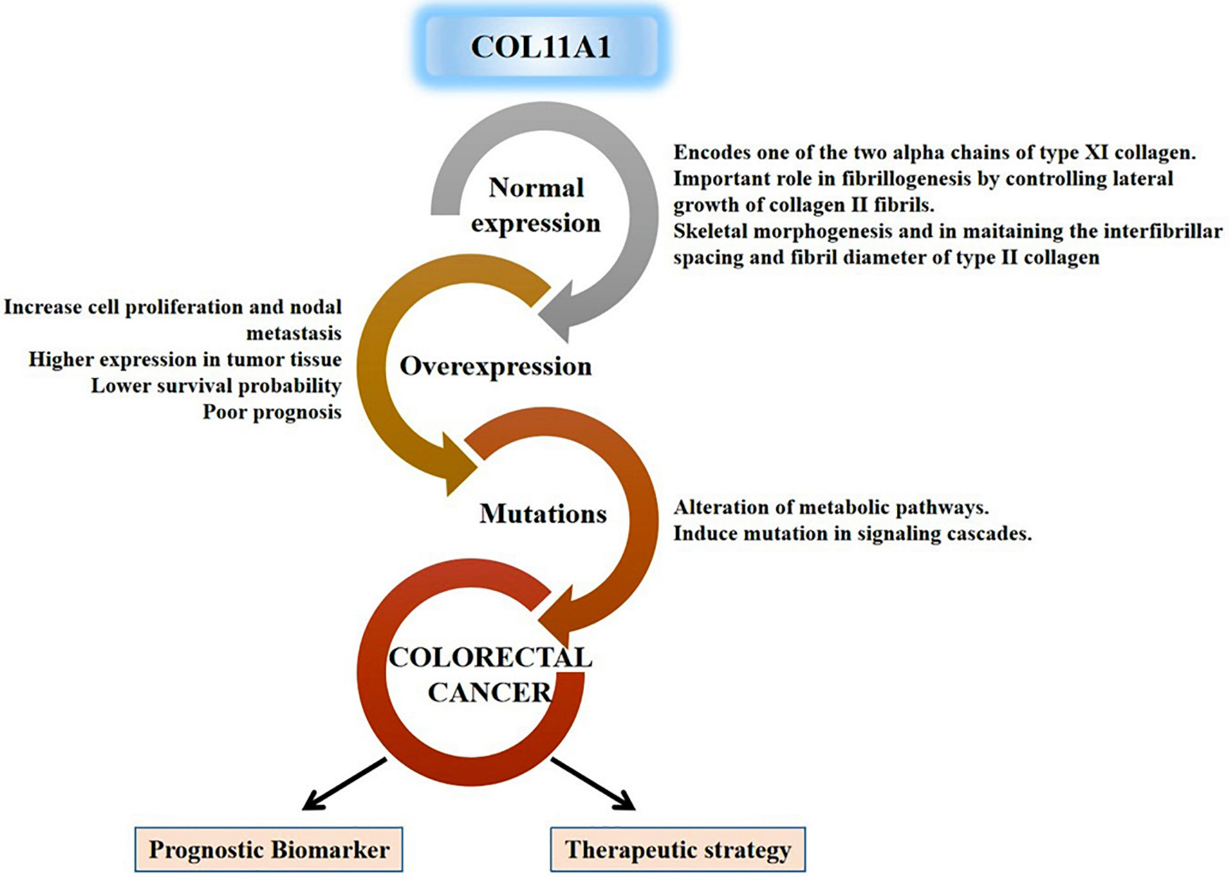

COL11A1
Normal expression
Encodes one of the two alpha chains of type XI collagen.
Important role in fibrillogenesis by controlling lateral growth of collagen II fibrils.
Skeletal morphogenesis and in maitaining the interfibrillar spacing and fibril diameter of type II collagen
Overexpression
Increase cell proliferation and nodal metastasis
Higher expression in tumor tissue
Lower survival probability
Poor prognosis
Mutations
Alteration of metabolic pathways.
Induce mutation in signaling cascades.
COLORECTAL CANCER
Prognostic Biomarker
Therapeutic strategy

FIGURE 8 | Schematic representation for functional relevance of COL11A1 gene in the oncogenesis of colorectal cancer and its candidature as a prognostic biomarker and therapeutic target.

(**Supplementary Figure 2**). Moreover, our study through the PathwayMapper tab of the cBioPortal website indicates the frequency of alteration of the various signaling cascades of RTK-RAS-PI3K, Wnt, TGF-β, and TP53 pathways that consequently leads to colorectal cancer. The RTK-RAS-PI3K signaling axis is important in regulating the cell growth and survival (Xu et al., 2020). Perturbation in these signaling cascades is known to contribute in the induction as well as in the development of cancer. The mutation of KRAS is found to be higher in colorectal cancer and thought to enhance the malignancy character of the transformed cells (Zenonos and Kyprianou, 2013). The alteration of the PI3K pathway mainly including the RTK upstream regulator of PI3K, catalytic subunit PIK3CA, PTEN negative regulator, and the downstream regulator of PI3K lead to the surge of cancer development (Yuan and Cantley, 2008). Herein, our study reveals the impact of COL11A1 gene product in the alteration of PTEN, PIK3CA, KRAS, and BRAF which might downregulate the RTK-RAS-PI3K signaling pathways to induce cancer development (**Figure 7A**). On the other hand, the Wnt signaling pathway is associated with the regulation of various developmental and physiological processes including cell division, specification, proliferation, and even maintenance of tissues and abnormal signaling leading to colorectal cancer (Clevers, 2006). The mutation of APC leads to overactivation of Wnt signaling pathways resulting in 80% of colorectal cancer prognosis (Koveitypour et al., 2019). The influence of COL11A1 and its associated gene triggers alteration of APC for around 66.7% that resulted in the overactivation of Wnt signaling pathways leading to cancer development (**Figure 7B**). TGF-β signaling pathway plays a vital role in tissue maintenance and is associated with inflammation and carcinogenesis by restraining the cell growth, differentiation, and apoptosis (Koveitypour et al., 2019). Mutation of TGF-β receptor type 2 (TGFBR2) leads to the microsatellite instability causing colorectal cancer, and also the loss of function of SMAD4 in the TGF-β signaling pathway promotes the tumor progression and poor survival in colorectal cancer (Itatani et al., 2019). The alteration of the TGF-β signaling pathway by the COL11A1 gene indicates that the SMAD4 alteration frequency of 15.5% might drive the formation of cancer (**Figure 7C**). TP53 pathway is the regulator of the cell cycle, DNA replication, apoptosis, and response to a wide range of stresses and safeguards maintenance of genomic integrity and acts as a tumor suppressor gene (Aubrey et al., 2016). The mutation of TP53 leads to colorectal cancer elevating the invasiveness, metastasis, and poor survival (Li et al., 2015). The association of the COL11A1 gene with its correlated gene from our study influences the alteration of TP53 by 53.0% disrupting the pathway results in uncontrolled cell proliferation and metastasis (**Figure 7D**).

All these discrete pieces of evidences from our experimental results designate the significance of the COL11A1 gene along with its highly correlated genes (THBS2, COL10A1, COL5A2, and COL1A2) in the progression of colorectal cancer across the various parameters using a wide range of data available in the cancer databases globally. Taken together, this study comprehensively enlightens the validation of the COL11A1 gene in the initiation, progression, and development of colorectal cancer using the bioinformatic approach, and the overall mechanism is schematized in **Figure 8**.

CONCLUSION

Our study provides several important pieces of evidences on the significance of the COL11A1 gene in the prognosis of human colorectal cancer. The overexpression of COL11A1 is positively upregulated in the cancer tissue across the various clinicopathological conditions, while negatively regulated in the case of promoter methylation indicating that the hypermethylation can induce the inhibition of cancer development. The survival assay signifies poor prognosis in both overall and disease-free survival. Our *in silico* study reveals that an abundance of COL11A1 mRNA could induce the transcriptional upregulation of THBS2, COL10A1, COL5A2, and COL1A2 genes cooperatively, to promote the neoplasia. The dysregulation in the expression of COL11A1 and mutations alters various critical regulatory pathways to influence the oncogenesis of colorectal cancer in humans. Therefore, our experimental data firmly claims the candidature of the COL11A1 gene as a potential biomarker for the prognosis of colorectal cancer and opens new areas of research for the diagnosis and development of appropriate therapeutic strategies. However, further *in vitro* and *in vivo* experimental validations are required to determine the efficacy of the COL11A1 gene in the prognosis of colorectal cancer and the development of a therapeutic strategy. Regarding this, we are in order to work on the cancer cell-lines and the murine model of colorectal cancer for validating the present study and developing efficacious therapeutic strategy by targeting COL11A1 gene.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RP performed all the experiments, analyzed the data, and wrote the manuscript. NCD reviewed the data and manuscript. SM analyzed the data, edited the manuscript, and supervised the study. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

RP thanks the Department of Higher Education, Government of West Bengal for awarding Swami Vivekananda Merit Cum Means Fellowship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.608313/full#supplementary-material

REFERENCES

Aubrey, B. J., Strasser, A., and Kelly, G. L. (2016). Tumor-suppressor functions of the TP53 pathway. *Cold Spring Harb. Perspect. Med.* 6:a026062. doi: 10.1101/cshperspect.a026062

Brown, R. J., Mallory, C., McDougal, O. M., and Oxford, J. T. (2011). Proteomic analysis of Col11a1-associated protein complexes. *Proteomics* 11, 4660–4676. doi: 10.1002/pmic.201100058

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401L–404. doi: 10.1158/2159-8290.CD-12-0095

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *Br. J. Cancer* 89, 232–238. doi: 10.1038/sj.bjc.6601118

Clevers, H. (2006). Wnt/β-catenin signaling in development and disease. *Cell* 127, 469–480. doi: 10.1016/j.cell.2006.10.018

Fischer, H., Stenling, R., Rubio, C., and Lindblom, A. (2001). Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis* 22, 875–878. doi: 10.1093/carcin/22.6.875

García-Pravia, C., Galván, J. A., Gutiérrez-Corral, N., Solar-García, L., García-Pérez, E., García-Ocaña, M., et al. (2013). Overexpression of COL11A1 by cancer-associated fibroblasts: clinical relevance of a stromal marker in pancreatic cancer. *PLoS One* 8:e78327. doi: 10.1371/journal.pone.0078327

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38, 675–678. doi: 10.1038/s41587-020-0546-548

Greenberg, M. V. C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* 20, 590–607. doi: 10.1038/s41580-019-0159-6

Itatani, Y., Kawada, K., and Sakai, Y. (2019). Transforming Growth Factor-β signaling pathway in colorectal cancer and its tumor microenvironment. *Int. J. Mol. Sci.* 20:5822. doi: 10.3390/ijms20235822

Jia, D., Liu, Z., Deng, N., Tan, T. Z., Huang, R. Y.-J., Taylor-Harding, B., et al. (2016). A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. *Cancer Lett.* 382, 203–214. doi: 10.1016/j.canlet.2016.09.001

Kim, H., Watkinson, J., Varadan, V., and Anastassiou, D. (2010). Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics* 3:51. doi: 10.1186/1755-8794-3-51

Koveitypour, Z., Panahi, F., Vakilian, M., Peymani, M., Forootan, F. S., Esfahani, M. H. N., et al. (2019). Signaling pathways involved in colorectal cancer progression. *Cell Biosci.* 9:97.

Li, A., Li, J., Lin, J., Zhuo, W., and Si, J. (2017). COL11A1 is overexpressed in gastric cancer tissues and regulates proliferation, migration and invasion of HGC-27 gastric cancer cells in vitro. *Oncol. Rep.* 37, 333–340. doi: 10.3892/or.2016.5276

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77, e108–e110.

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2. 0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514.

Li, X.-L., Zhou, J., Chen, Z.-R., and Chng, W.-J. (2015). P53 mutations in colorectal cancer-molecular pathogenesis and pharmacological reactivation. *World J. Gastroenterol. WJG* 21:84. doi: 10.3748/wjg.v21.i1.84

Loeb, L. A., Bielas, J. H., and Beckman, R. A. (2008). Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res.* 68, 3551–3557. doi: 10.1158/0008-5472.can-07-5835

Pancione, M., Remo, A., and Colantuoni, V. (2012). Genetic and epigenetic events generate multiple pathways in colorectal cancer progression. *Patholog. Res. Int.* 2012:509348. doi: 10.1155/2012/509348

Raglow, Z., and Thomas, S. M. (2015). Tumor matrix protein collagen XIα1 in cancer. *Cancer Lett.* 357, 448–453. doi: 10.1016/j.canlet.2014.12.011

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9:166. doi: 10.1593/neo.07112

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6:1. doi: 10.1016/s1476-5586(04)80047-2

Shen, L., Yang, M., Lin, Q., Zhang, Z., Zhu, B., and Miao, C. (2016). COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. *Oncol. Rep.* 36, 877–885. doi: 10.3892/or.2016.4869

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., et al. (2017). Colorectal cancer statistics, 2017. *CA. Cancer J. Clin.* 67, 177–193. doi: 10.3322/caac.21395

Su, C., Zhao, J., Hong, X., Yang, S., Jiang, Y., and Hou, J. (2019). Microarray-based analysis of COL11A1 and TWIST1 as important differentially-expressed pathogenic genes between left and right-sided colon cancer. *Mol. Med. Rep.* 20, 4202–4214. doi: 10.3892/mmr.2019.10667

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Toss, M. S., Miligy, I. M., Gorringe, K. L., Aleskandarany, M. A., Alkawaz, A., Mittal, K., et al. (2019). Collagen (XI) alpha-1 chain is an independent prognostic factor in breast ductal carcinoma in situ. *Mod. Pathol. Off. J. U.S. Can. Acad. Pathol. Inc.* 32, 1460–1472. doi: 10.1038/s41379-019-0286-289

Vázquez-Villa, F., García-Ocaña, M., Galván, J. A., García-Martínez, J., García-Pravia, C., Menéndez-Rodríguez, P., et al. (2015). COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumor Biol.* 36, 2213–2222. doi: 10.1007/s13277-015-3295-3294

Wu, Y.-H., Chang, T.-H., Huang, Y.-F., Huang, H.-D., and Chou, C.-Y. (2014). COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene* 33, 3432–3440. doi: 10.1038/onc.2013.307

Xu, F., Na, L., Li, Y., and Chen, L. (2020). Roles of the PI3K/AKT/mTOR signalling pathways in neurodegenerative diseases and tumours. *Cell Biosci.* 10:54. doi: 10.1186/s13578-020-00416-410

Yuan, T. L., and Cantley, L. C. (2008). PI3K pathway alterations in cancer: variations on a theme. *Oncogene* 27, 5497–5510. doi: 10.1038/onc.2008.245

Zenonos, K., and Kyprianou, K. (2013). RAS signaling pathways, mutations and their role in colorectal cancer. *World J. Gastrointest. Oncol.* 5, 97–101. doi: 10.4251/wjgo.v5.i5.97

Zhang, J., Roberts, T. M., and Shivdasani, R. A. (2011). Targeting PI3K signaling as a therapeutic approach for colorectal cancer. *Gastroenterology* 141, 50–61. doi: 10.1053/j.gastro.2011.05.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fgene.2021.606816

Identifying Hypoxia Characteristics to Stratify Prognosis and Assess the Tumor Immune Microenvironment in Renal Cell Carcinoma

Zhenan Zhang [1,2†], Qinhan Li [1,2†], Feng Wang [3], Binglei Ma [1,2], Yisen Meng [1,2] and Qian Zhang [1,2]*

[1] Department of Urology, Peking University First Hospital, Beijing, China, [2] National Research Center for Genitourinary Oncology, Institute of Urology, Peking University, Beijing, China, [3] Department of Urology, People's Hospital of Tibet Autonomous Region, Lhasa, China

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Gurmeet Kaur,
National Center for Biotechnology Information (NLM), United States
Jun Zhang,
Anhui University, China

Correspondence:
Yisen Meng
mysbmu@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 September 2020*
Accepted: *05 May 2021*
Published: *14 June 2021*

Citation:
Zhang Z, Li Q, Wang F, Ma B, Meng Y and Zhang Q (2021) Identifying Hypoxia Characteristics to Stratify Prognosis and Assess the Tumor Immune Microenvironment in Renal Cell Carcinoma.
Front. Genet. 12:606816.
doi: 10.3389/fgene.2021.606816

Background: Renal cell carcinoma (RCC) is a common malignant tumor worldwide, and immune checkpoint inhibitors are a new therapeutic option for metastatic RCC. Infiltrating immune cells in the tumor microenvironment (TME) play a critical part in RCC biology, which is important for tumor therapy and prediction. Hypoxia is a common condition that occurs in the TME and may lead to RCC immunosuppression and immune escape. This study was conducted to analyze the extent of the hypoxia immune microenvironment in the TME of RCC and develop a hypoxia-related risk model for predicting the prognosis of patients with RCC.

Methods: The gene expression profiles of 526 patients with RCC were downloaded from The Cancer Genome Atlas database. Combined with the hallmark-hypoxia gene dataset downloaded from Gene Set Enrichment Analysis, prognosis-related hypoxia genes were selected by survival analysis. A protein–protein interaction network and functional enrichment analysis were performed. A hypoxia-related risk model predicting the prognosis of patients with RCC was established using the least absolute shrinkage and selection operator. Data of 91 cases downloaded from the International Cancer Genome Consortium (ICGC) database were used for validation. CIBERSORT was applied to analyze the fractions of 22 immune cell types in the TME of RCC between low- and high-risk groups. The expression profiles of immunomodulators and immunosuppressive cytokines were also analyzed.

Results: Ninety-three genes were significantly associated with poor overall survival of patients with RCC and were mainly involved in 10 pathways. Using the established hypoxia-related risk model, the receiver operating characteristic curves showed an accuracy of 76.1% (95% CI: 0.719–0.804), and Cox proportional hazards regression analysis revealed that the model was an independent predictor of the prognosis of patients with RCC [hazard ratio (HR) = 2.884; 95% CI: 2.090–3.979] ($p < 0.001$). Using the ICGC database, we verified that the low-risk score group had a better overall survival outcome than the high-risk group. Additionally, dividing the hypoxia risk score into high-risk and low-risk groups could predict the immune microenvironment of RCC.

Conclusions: We demonstrated that a hypoxia-related risk model can be used to predict the outcomes of patients with RCC and reflect the immune microenvironment of RCC, which may help improve the overall clinical response to immune checkpoint inhibitors.

Keywords: renal cell carcinoma, immune response, tumor microenvironment, hypoxia, risk model

INTRODUCTION

Kidney cancer is a common malignant tumor worldwide, with an estimated 403,000 new cases and 175,000 deaths in 2018 (Bray et al., 2018). Renal cell carcinoma (RCC) is the most common form of kidney cancer, and ∼70% of these cases show clear-cell tumors in histological analysis (Lipworth et al., 2016). Surgical resection, including radical nephrectomy and nephron-sparing surgery, remains the most effective therapy for clinically localized RCC. Once metastasis of RCC occurs, clinical treatment is challenging and patients show a 5-year survival rate of approximately 12% (Siegel et al., 2017). Cytokines [interferon (IFN)-α, interleukin (IL)-2], targeted therapy [tyrosine kinase inhibitors, anti-vascular endothelial growth factor (VEGF) antibodies, agents targeting the mammalian target of rapamycin (mTOR)], and immune checkpoint inhibitors are used as therapies for metastatic RCC. However, it is important to elevate the overall clinical response rate of cancer immunotherapy and identify biomarkers for response prediction.

Multiple factors contribute to cancer initiation and progression. The tumor microenvironment (TME) is an important regulator of tumor progression and metastasis (McAllister and Weinberg, 2014). Infiltrating immune cells are among the major normal cells in tumor tissues and play a crucial role in tumor biology, tumor prognosis, drug resistance, and immunotherapeutic efficacy (Straussman et al., 2012; van Dijk et al., 2019; Guo et al., 2020). A better understanding of the TME, particularly infiltrating immune cells, is important for improving tumor therapy and tumor prediction.

Hypoxia is a common condition found in the TME, playing a vital role in tumor genetic instability and prognosis (LaGory and Giaccia, 2016). The hypoxia-inducible transcription factor (HIF) signaling pathway can be activated by tumor-induced hypoxia (Fallah and Rini, 2019). In clear-cell RCC (ccRCC), HIF is particularly important, with HIF-1α and HIF-2α exerting opposing effects on tumor development (Schödel et al., 2016). Small-molecule inhibitors of HIF-2 may serve as another therapeutic option for ccRCC in the future (Martínez-Sáez et al., 2017). Hypoxia can lead to tumor immunosuppression and immune escape. It has been reported that hypoxia promotes suppressive immune cells and immunosuppressive cytokines in the TME (Terry et al., 2017). Therefore, hypoxia-related genes may be useful for predicting immunotherapy outcomes.

This study was conducted to analyze the gene expression profiles of RCC downloaded from The Cancer Genome Atlas (TCGA) database and hypoxia-related genes (hallmark-hypoxia genes) downloaded from Gene Set Enrichment Analysis (GSEA). We selected prognosis-related hypoxia genes to develop a hypoxia-related risk model for predicting the prognosis and immune microenvironment landscape of patients with RCC in high/low hypoxia risk score groups. The workflow of the study design is shown in **Figure 1**.

MATERIALS AND METHODS

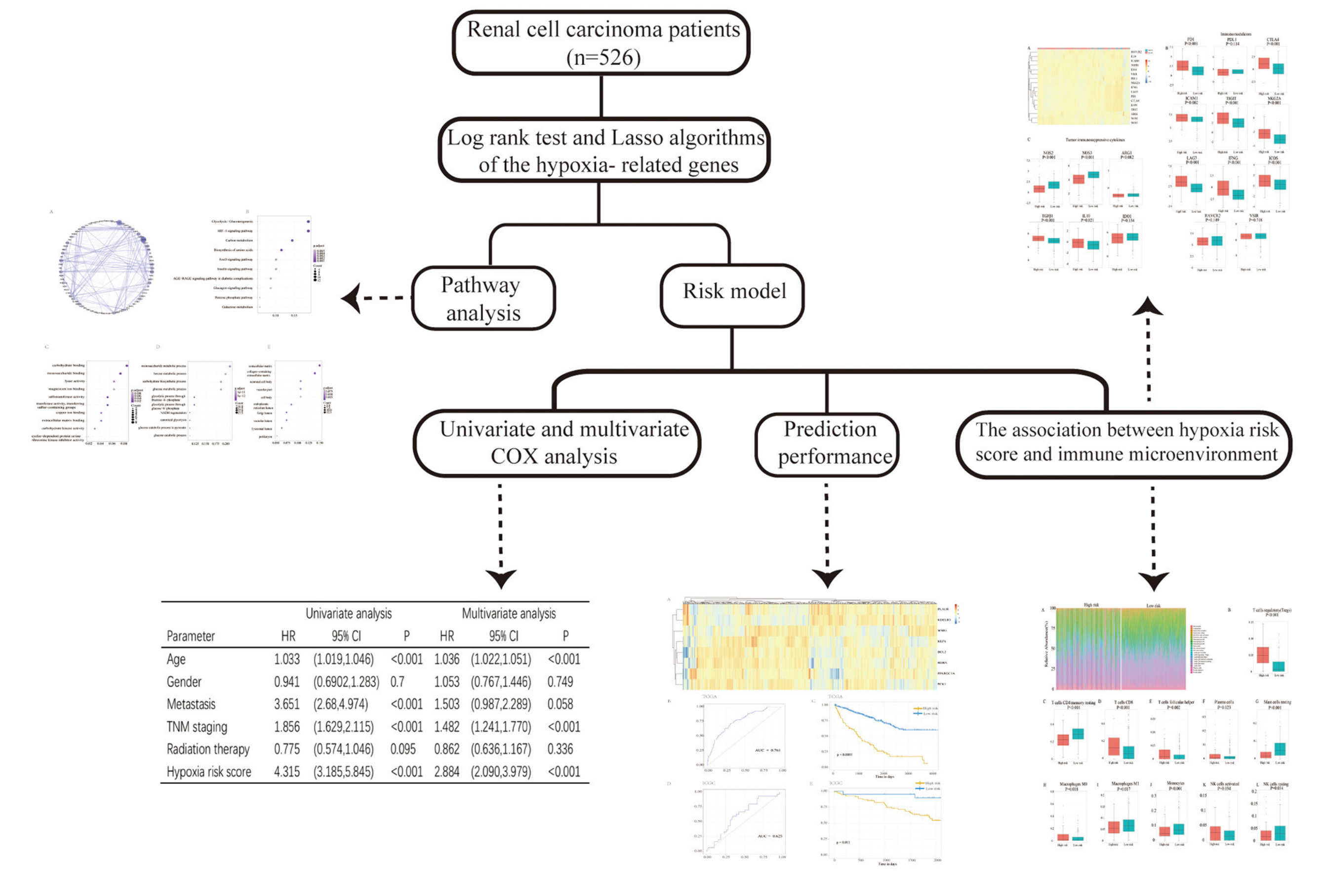
Database

The level 3 gene expression profiles of 526 patients with RCC were downloaded from TCGA database (https://tcga-data.nci.nih.gov/) (June 2020). The patients' clinical characteristics, including age, sex, TNM stage, and survival data, were also obtained from the database. Patients with cancer without pathologic diagnosis or a lack of clinical information were excluded.

Hypoxia-related genes (hallmark- hypoxia genes) were downloaded from GSEA (https://www.gsea-msigdb.org/gsea/index.jsp). The gene expression profiles of 91 patients with RCC determined by the CAGEKID consortium in Europe were downloaded from the International Cancer Genome Consortium (ICGC) database (https://icgc.org/icgc/cgp/65/812/817) and used as the validation cohort to verify the predictive value of the risk model.

Construction of Protein–Protein Interaction Network and Functional Enrichment Analysis

Hypoxia genes were selected using the log-rank test to identify statistically significant prognosis-related genes. The selected hypoxia genes were used to establish a protein–protein interaction (PPI) network and for functional enrichment analysis. The Search Tool for the Retrieval of Interacting Genes (STRING) database was used to generate the PPI network (Szklarczyk et al., 2015). Thereafter, Cytoscape software (version 3.7.0) was used to reconstruct and visualize the PPI network (Shannon et al., 2003). The connectivity degree of each protein node was calculated. The R package clusterprofile was utilized to perform functional enrichment analysis (Yu et al., 2012). Based on Gene Ontology (GO) categories, the genes were identified with different GO terms based on their respective characteristics: molecular functions (MFs), biological processes (BPs), and cellular components (CCs). Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were used for pathway enrichment analysis. The false discovery rate (FDR) was set at 0.05.


Renal cell carcinoma patients (n=526)
Log rank test and Lasso algorithms of the hypoxia- related genes
Pathway analysis
Risk model
Univariate and multivariate COX analysis
Prediction performance
The association between hypoxia risk score and immune microenvironment

Parameter	Univariate analysis HR	95% CI	P	Multivariate analysis HR	95% CI	P
Age	1.033	(1.019,1.046)	<0.001	1.036	(1.022,1.051)	<0.001
Gender	0.941	(0.6902,1.283)	0.7	1.053	(0.767,1.446)	0.749
Metastasis	3.651	(2.68,4.974)	<0.001	1.503	(0.987,2.289)	0.058
TNM staging	1.856	(1.629,2.115)	<0.001	1.482	(1.241,1.770)	<0.001
Radiation therapy	0.775	(0.574,1.046)	0.095	0.862	(0.636,1.167)	0.336
Hypoxia risk score	4.315	(3.185,5.845)	<0.001	2.884	(2.090,3.979)	<0.001

FIGURE 1 | The flowchart of the study.

Construction of a Risk Model

The selected prognosis-related hypoxia genes were applied in the least absolute shrinkage and selection operator (LASSO) using the R package glmnet. The hypoxia risk score formula was established based on gene expression multiplied by a linear combination of the regression coefficient, which was acquired from LASSO. The cases were divided into high- and low-risk groups based on the optimal cutoff point of the risk score with the R package survminer (version 0.4.6). R package survival and ROCR were utilized for Kaplan–Meier analysis and to generate receiver operating characteristic (ROC) curves. To draw heat maps, pheatmap (version 1.64.0) was used in R package. The predictive value of the risk model was verified using data from 91 patients with RCC downloaded from the ICGC database.

Assessment of Immune Cell Type Fractions

Using gene expression data, the analytical method CIBERSORT (https://cibersort.stanford.edu/) can be applied to characterize the cell composition in a mixed cell population (Newman et al., 2015). The leukocyte gene signature matrix containing 547 genes, named LM22 in CIBERSORT, was applied to distinguish 22 immune cell types including CD8 T cells, naive CD4 T cells, resting memory CD4 T cells, activated memory CD4 T cells, naive B cells, memory B cells, plasma cells, follicular helper T cells, T-regulatory cells (Tregs), gamma delta T cells, resting natural killer cells, activated natural killer cells, monocytes, macrophages M0, macrophages M1, M2, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils. We applied CIBERSORT to assess the fractions of these cell types between the low- and high-risk groups.

Expression Profile of Immunomodulators and Immunosuppressive Cytokines

Several key immunomodulators, including lymphocyte activation gene 3 (LAG-3), T cell immunoglobulin and mucin domain containing 3 (TIM-3), cytotoxic T lymphocyte associated protein 4 (CTLA-4), IFN-γ, ICOS inducible T cell costimulator (ICOS), intercellular adhesion molecule 1 (ICAM-1), T cell immunoreceptor with Ig and ITIM domains (TIGIT), PD-1 programmed cell death 1 (PD-1), programmed cell death 1 ligand 1 (PD-L1), natural killer group 2 member A (NKG2A), V-domain immunoglobulin suppressor of T cell activation (VISTA), and immunosuppressive cytokines were quantified. The *t*-test was applied to compare the differences in the expression levels of immunomodulators and immunosuppressive cytokines between the low- and high-risk groups. A two-sided $p < 0.05$ was considered to indicate statistical significance.

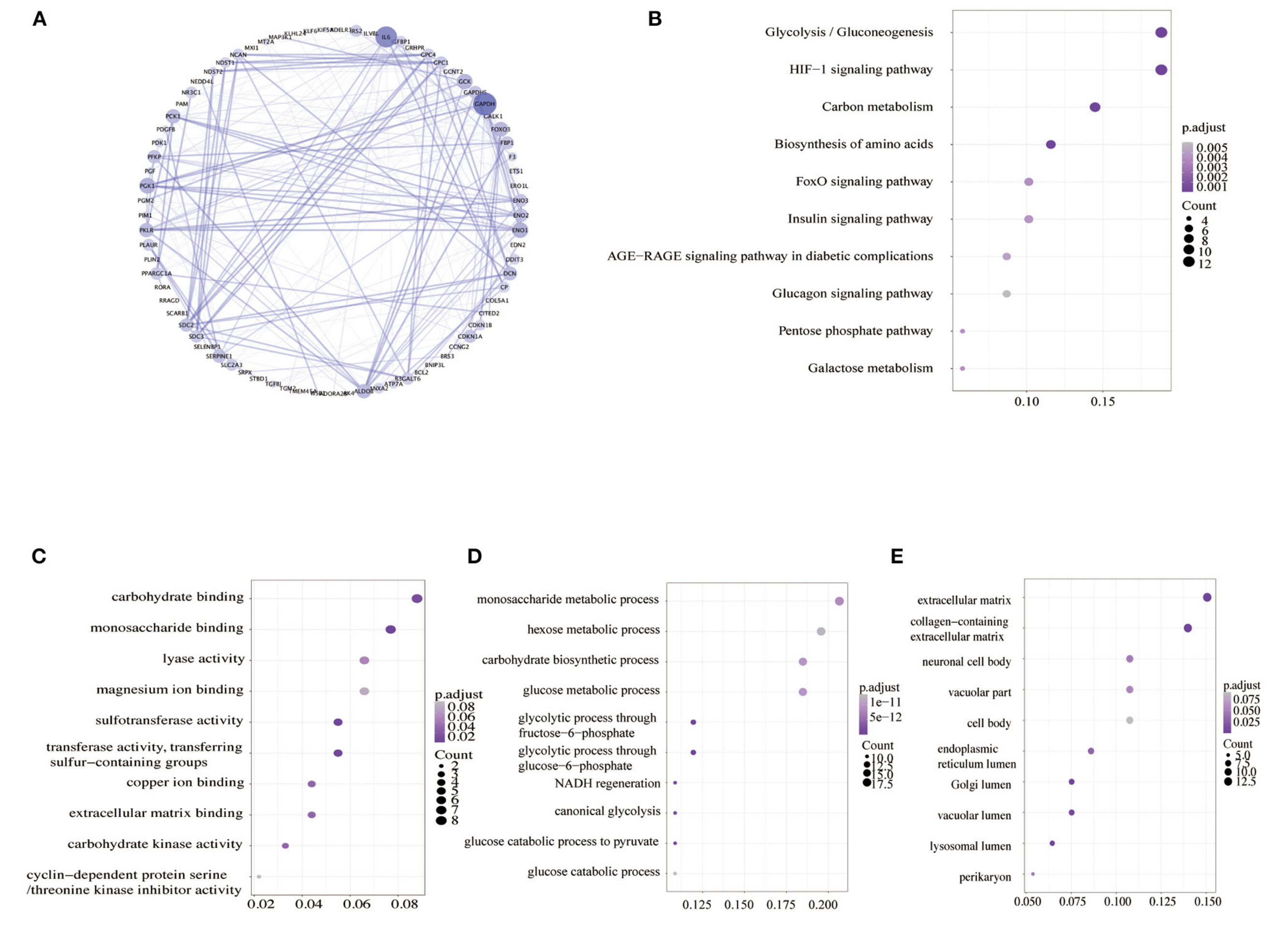


FIGURE 2 | Analysis of hypoxia-related genes. **(A)** PPI networks of hypoxia-related genes. A large node means a higher degree. **(B)** KEGG pathway enrichment analysis of hypoxia-related genes. **(C–E)** GO enrichment analysis of molecular function (MF), biological process (BP), and cellular component (CC). PPI, the protein–protein interaction; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

RESULTS

Characterization of Hypoxia-Related Genes

The hypoxia-related gene (hallmark- hypoxia genes) dataset downloaded from GSEA contained 200 genes. These genes were upregulated following treatment with low oxygen levels. In conjunction with the gene expression profiles of 526 patients with RCC downloaded from TCGA database, the prognostic predictive value of hypoxia-related genes was explored using Kaplan–Meier survival curves. Ninety-three genes were found to be significantly associated with poor overall survival outcomes according to log-rank test ($p < 0.05$; **Supplemental Table 1**). The STRING database and Cytoscape software were used to build the PPI network of these genes (**Figure 2A**). To evaluate the 93 genes, we performed KEGG and GO analyses. KEGG analysis illustrated that the genes primarily participated in 10 pathways (**Figure 2B**), including glycolysis/gluconeogenesis, HIF-1 signaling pathway, carbon metabolism, biosynthesis of amino acids, etc. The 285 GO terms, including 269 biological process terms, eight cellular component terms, and eight molecular function terms, were enriched ($p < 0.05$; **Supplemental Table 2**). The top GO terms, including carbohydrate binding, monosaccharide metabolic process, and extracellular matrix, are shown in **Figures 2C–E**.

Evaluation Prognosis Prediction Power of the Hypoxia-Related Risk Model

LASSO was used to explore the hypoxia-related risk model predicting the prognosis of patients with RCC. The optimal LASSO model was selected that included eight identified genes, *PLAUR*, *BCL2*, *KLF6*, *KDELR3*, *WSB1*, *PPARGC1A*, *PCK1*, and *RORA*. The risk score was calculated using the following formula: risk score = 0.34577 × expression (*PLAUR*) + 0.18588 × expression (*BCL2*) + (−0.45209) × expression (*KLF6*) + 0.22279 × expression (*KDELR3*) + 0.53993 × expression (*WSB1*) + (−0.13366) × expression (*PPARGC1A*) + 0.01587 × expression (*PCK1*) + (−0.55309) × expression (*RORA*). **Figure 3A** shows the heatmap exhibiting the distinct gene expression patterns of the selected genes. Receiver operating characteristic (ROC)

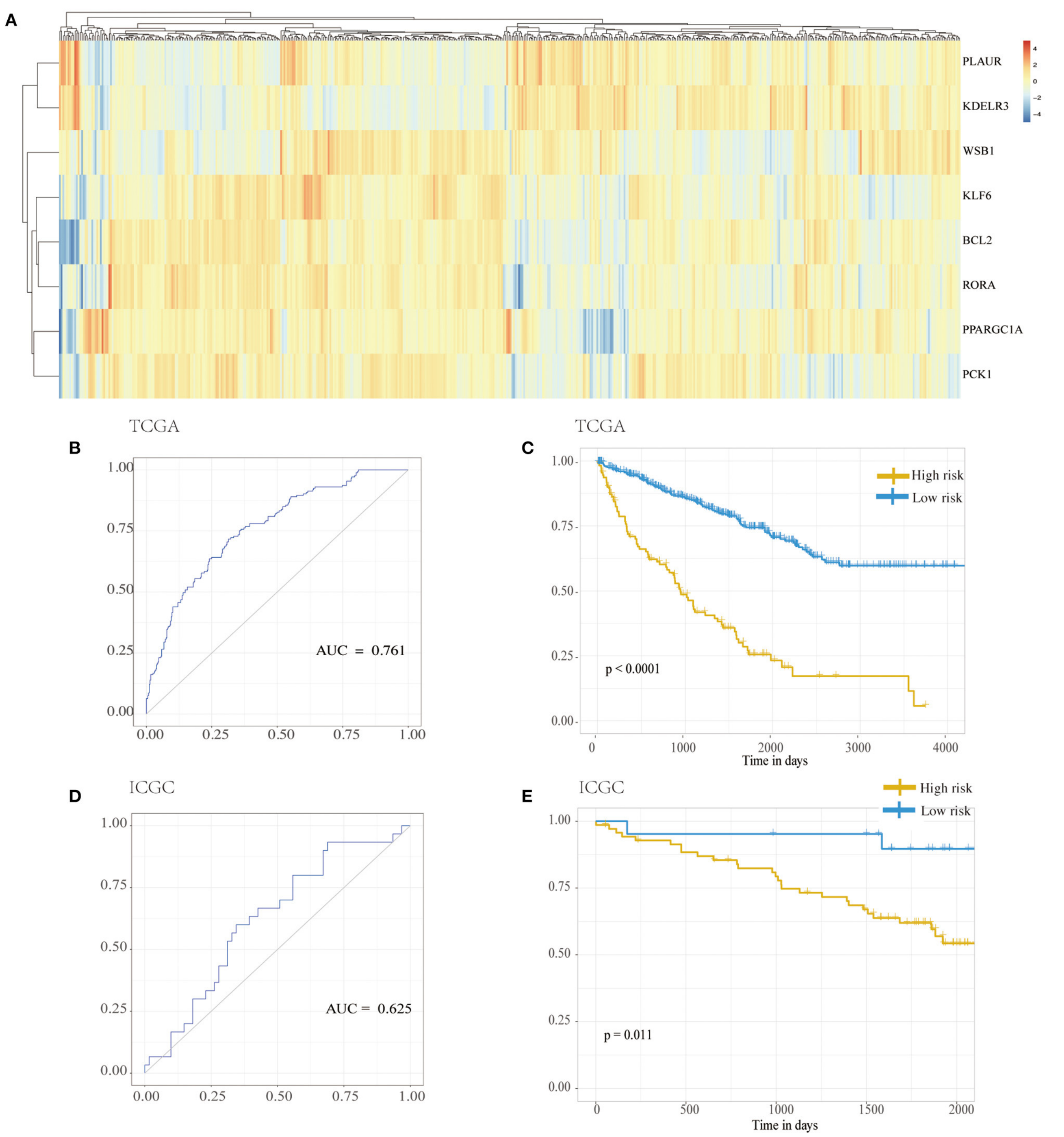


FIGURE 3 | Hypoxia risk model. **(A)** Distribution of genes in the hypoxia risk model. **(B)** ROC analysis for the hypoxia risk model. **(C)** Kaplan–Meier curves for overall survival of risk score in TCGA cohort. **(D)** ROC analysis for the hypoxia risk model in the ICGC cohort. **(E)** Kaplan–Meier curves for overall survival of risk score in ICGC cohort. ICGC, International Cancer Genome Consortium; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas.

curves were used to evaluate the prognosis prediction power of the hypoxia-related risk model shown in **Figure 3B**. The model had an accuracy of 76.1% (95% CI: 0.719–0.804), and its predictive ability was higher than those of any other clinical characteristics (**Table 1**).

Based on the chosen cutoff value of 0.5, the cases were divided into high and low hypoxia risk score group. According to the Kaplan–Meier curve, **Figure 3C** illustrates that the low-risk score group had a better overall survival outcome than the high-risk score group ($p < 0.001$). Adjusting for confounding variables,

TABLE 1 | The predictive accuracy of the hypoxia-related risk model and other clinical characteristics.

	AUC	95% CI
Age	62.9%	0.579–0.680
Gender	49.0%	0.447–0.534
Metastasis	62.9%	0.589–0.669
TNM staging	74.6%	0.701–0.791
Radiation therapy	48.5%	0.439–0.530
Hypoxia risk score	76.1%	0.719–0.804

TABLE 2 | The univariate analysis and multivariate analysis of the hypoxia risk score.

Parameter	Univariate analysis			Multivariate analysis		
	HR	95% CI	*p*	HR	95% CI	*p*
Age	1.033	(1.019–1.046)	<0.001	1.036	(1.022–1.051)	<0.001
Gender	0.941	(0.6902–1.283)	0.7	1.053	(0.767–1.446)	0.749
Metastasis	3.651	(2.68–4.974)	<0.001	1.503	(0.987–2.289)	0.058
TNM staging	1.856	(1.629–2.115)	<0.001	1.482	(1.241–1.770)	<0.001
Radiation therapy	0.775	(0.574–1.046)	0.095	0.862	(0.636–1.167)	0.336
Hypoxia risk score	4.315	(3.185–5.845)	<0.001	2.884	(2.090–3.979)	<0.001

including age, gender, metastasis, TNM staging, and radiation therapy, Cox proportional hazards regression analysis revealed that the hypoxia risk score was an independent predictor of RCC patient prognosis, as shown in **Table 2** [hazard ratio (HR) = 2.884; 95% CI: 2.090–3.979] ($p < 0.001$).

From the ICGC database, data from a cohort of 91 patients with RCC was obtained to verify the results. As shown in **Figure 3D**, the accuracy of the model was 62.5% (95% CI: 0.505–0.745) in the validation samples. Additionally, **Figure 3E** also shows that the low-risk score group had a better overall survival outcome compared to the high-risk score group ($p = 0.011$).

Immune Landscape of High/Low Hypoxia Risk Score Groups

The capability of the hypoxia-related risk model to assess the immune microenvironment of RCC was evaluated. We utilized the CIBERSORT method with the LM22 signature gene file to assess the immune cell fraction between the low- and high-risk groups. A summary of the results based on 526 patients with RCC downloaded from TCGA database is illustrated in **Figure 4A**. The proportions of immunosuppressive cells, such as Tregs, were significantly higher in the high hypoxia risk score groups, as shown in **Figure 4B**. This indicates that patients with high hypoxia risk scores possess an immunosuppressive microenvironment. **Figures 4C–L** show other types of immune cells, which exhibited significantly different proportions between the low- and high-risk groups.

Expression Profile of Immunomodulators and Immunosuppressive Cytokines

The expression of 11 immunomodulators and six immunosuppressive cytokines in 526 patients with RCC downloaded from TCGA database is illustrated in **Figure 5A**. We found that the expression of PD-1, CTLA-4, ICAM-1, TIGIT, NKG2A, LAG-3, IFNG, and ICOS was significantly upregulated in the high hypoxia risk score group, as shown in **Figure 5B**. Immunosuppressive cytokines, such as transforming growth factor (TGF)-β1 and IL-10, were also significantly upregulated in the high hypoxia risk score group, as shown in **Figure 5C**. However, NOS2 and NOS3 were significantly reduced in the high-risk group. As a result, patients with RCC in the high hypoxia risk score group may have an immunosuppressive tumor microenvironment with upregulated immunomodulators and immunosuppressive cytokines. Therefore, targeting hypoxia may benefit immunotherapy in clinical practice.

DISCUSSION

Previous studies demonstrate that hypoxia and hypoxia-related signaling pathways play important roles in the development and progression of RCC (Schödel et al., 2016). Von Hippel–Lindau tumor suppressor (pVHL) and HIFs are critical factors in these pathways. With tumor cell proliferation and growth, RCC results in hypoxia with the activated HIF-α signaling in response to oxygen deprivation (Millet-Boureima et al., 2021). On the other hand, the VHL gene is lost in ∼90% of ccRCC tumors (Linehan and Ricketts, 2019). In normal renal tissue, oxygen-dependent posttranslational modifications on HIF-2α allow pVHL to normally recognize and mediate the proteasomal degradation (Choueiri and Kaelin, 2020). Loss of VHL gene in ccRCC, tumor is under pseudohypoxia state with accumulated HIF-2α and activated HIF-1 to upregulate the expression of hypoxia-inducible genes and increase tumor oxygenation (Haase, 2013). Hypoxia is a phenomenon in other cancers. HIF-2α has been known to regulate tumor proliferation, metabolism, metastasis, and resistance to chemotherapy in digestive system cancers (Zhao et al., 2015). In melanoma, a hypoxia-related signature has been developed to predict prognosis (Shou et al., 2021).

In the present study, we identified 93 hypoxia-related genes significantly associated with the outcomes of patients with RCC. The PPI network of these selected genes significantly included glyceraldehyde-3-phosphate dehydrogenase (GAPDH), IL-6, phosphoglycerate kinase 1 (PGK1), enolase 1 (ENO1), and glucokinase (GCK). GAPDH is a key enzyme involved in glycolysis and is related to cell proliferation in RCC (Vilà et al., 2000). IL-6 has been shown to induce drug resistance in RCC and is associated with poor prognosis (Ishibashi et al., 2018). PGK1 is also a glycolytic enzyme that can be secreted by tumor cells to participate in angiogenesis. ENO1, GCK, PGK1, and GAPDH are involved in tumor energy metabolism. Functional enrichment analysis revealed that these genes were specifically related to the glucometabolic process, hypoxia-related pathway, carbon metabolism, and extracellular matrix. These results suggest that

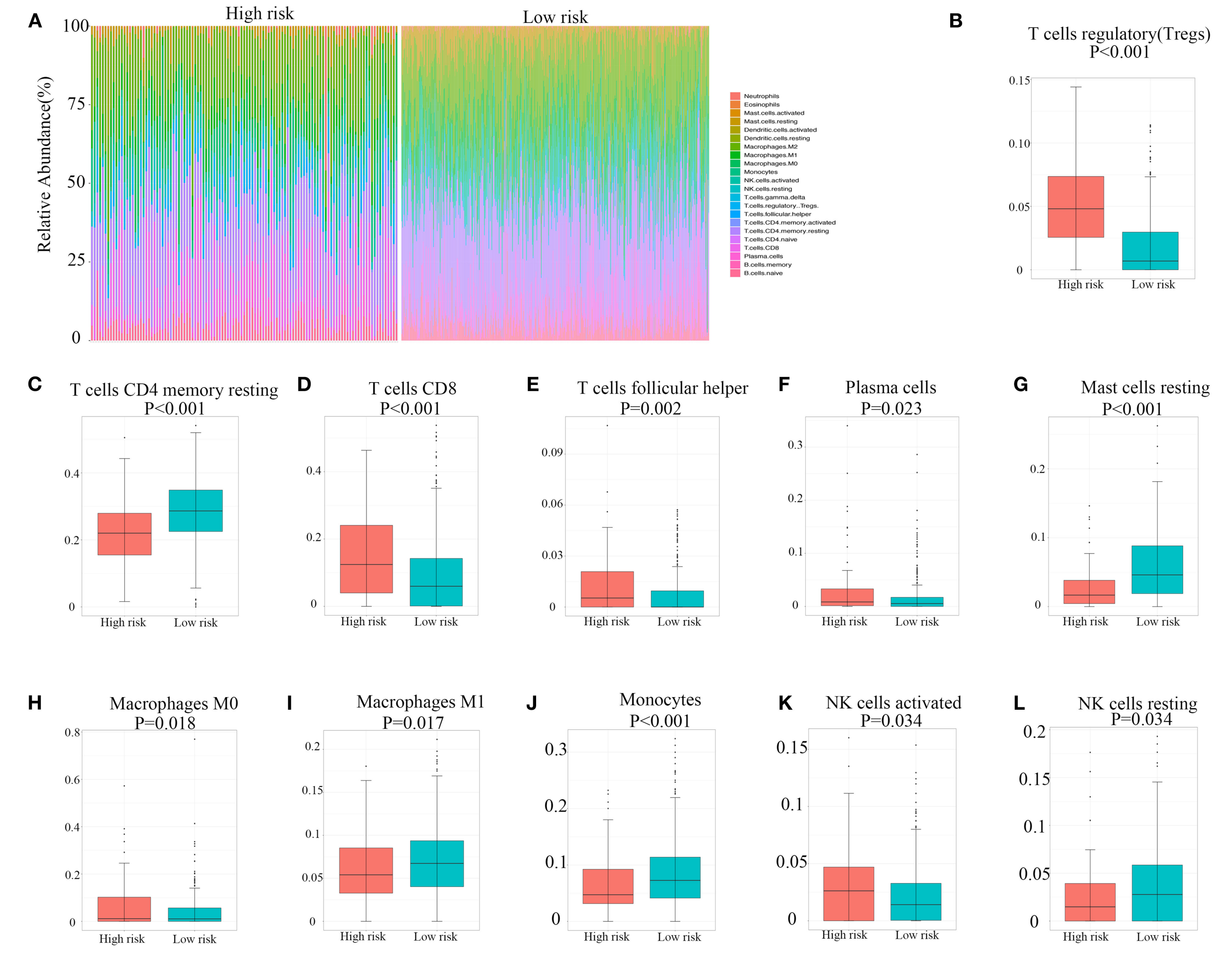


FIGURE 4 | Immune landscape between high and low hypoxia risk score groups of renal cell carcinoma patients. **(A)** The abundance of immune infiltration in high-risk and low-risk groups from TCGA cohort. **(B–L)** The proportions of different immune cells between high-risk and low-risk groups in TCGA cohort. TCGA, The Cancer Genome Atlas.

hypoxia-related energy metabolism is associated with tumor prognosis and the TME condition.

Multiple approaches have been developed to predict the prognosis of RCC, including prognostic models and nomograms. Tumor node metastasis classification remains the most important identified prognostic factor (Klatte et al., 2018). Immunohistochemical staining of Ki-67, p53, and VEGFR-1 was shown to be significantly related to RCC outcomes. Molecular markers have also been applied as prognostic models. The ClearCode34-based model was developed including 34 genes to classify the subtypes of localized ccRCC to predict patient survival outcomes (Brooks et al., 2014). The continuous CLEAR score (continuous linear enhanced assessment of ccRCC) was developed based on an 18-transcript signature to predict patients' disease-specific survival and the response to tyrosine kinase inhibitor (Wei et al., 2017).

We developed a hypoxia-related risk model based on hypoxia-related genes to predict the prognosis of patients with RCC. The model had an accuracy of 76.1% (95% CI: 0.719–0.804) and was found to be an independent predictor in Cox proportional hazards regression analysis. *PLAUR* encodes the receptor for urokinase plasminogen activator. *BCL2* encodes a membrane protein that regulates lymphocyte apoptosis. *KLF6* encodes the zinc finger protein that acts as a tumor suppressor. *KDELR3* encodes a member of the KDEL endoplasmic reticulum protein retention receptor family. *WSB1* encodes a member of the WD-protein subfamily. *PPARGC1A* encodes proteins that regulate energy metabolism. *PCK1* is a critical regulator of gluconeogenesis. *RORA* participates in tumor metastasis regulation. *PLAUR*, *BCL2*, *KLF6*, *WSB1*, *PPARGC1A*, and *PCK1* were identified to be related to the prognosis of ccRCC (Hirata et al., 2009; Syafruddin et al., 2019; Xu et al., 2019; Liu et al., 2020;

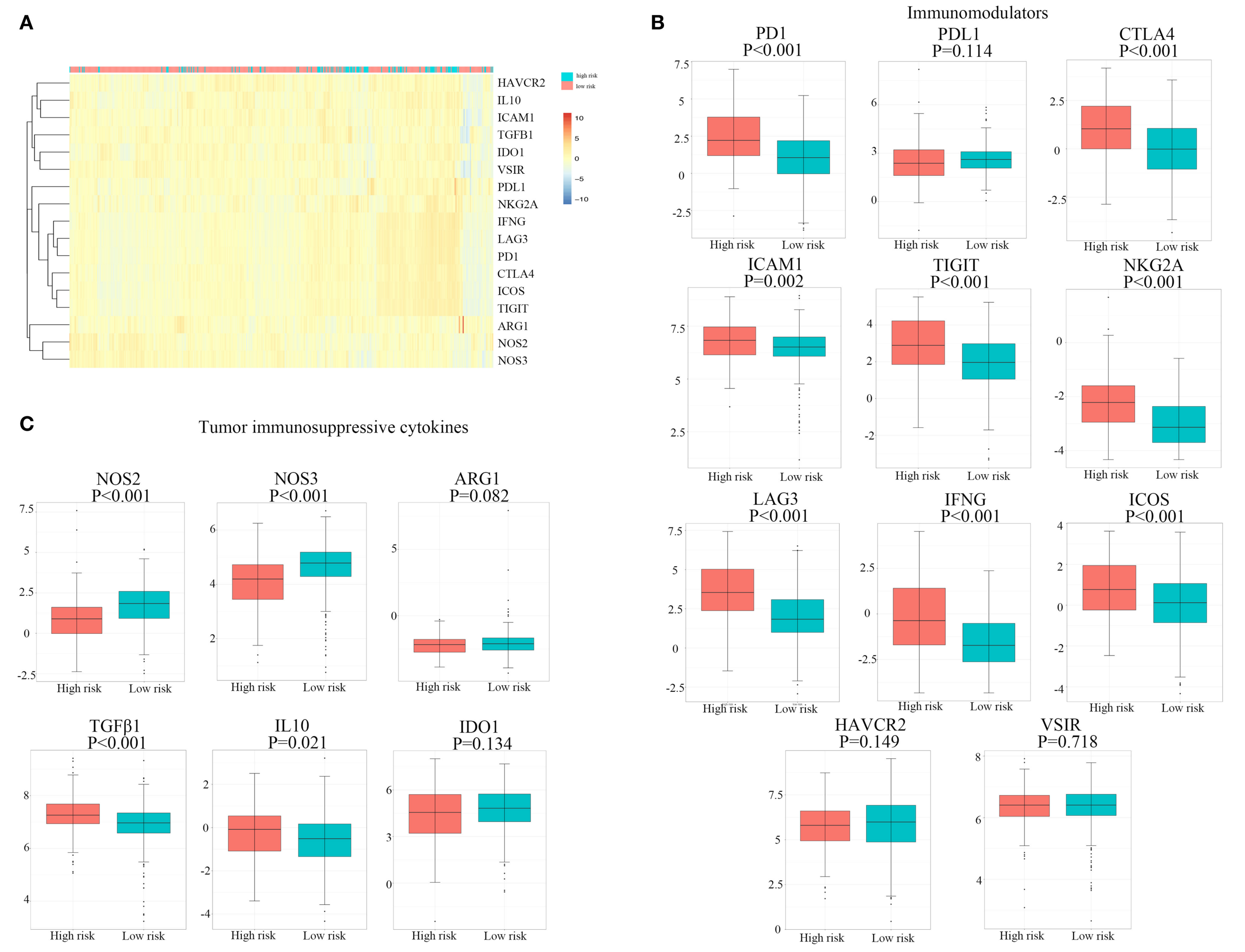


FIGURE 5 | Expression profile of immunomodulators and immunosuppressive cytokines. **(A)** Distribution of genes of immunomodulators and immunosuppressive cytokines. **(B,C)** The proportions of immunomodulators and immunosuppressive cytokines between high and low hypoxia risk groups in TCGA cohort. TCGA, The Cancer Genome Atlas.

Shen et al., 2020; Shi et al., 2020). The functions of KDELR3 and RORA have not been reported in RCC.

Further analysis demonstrated that the hypoxia-related risk model was also related to the immune microenvironment of RCC. Tregs are key players in tumor immune escape and angiogenesis (Facciabene et al., 2012). Monocytes congregate in the TME and differentiate into tumor-associated macrophages (TAMs). Hypoxia has a profound effect on these cells (Lewis and Murdoch, 2005). Tregs were discovered to be significantly higher in the high hypoxia risk score groups, indicating an immunosuppressive microenvironment in these patients. Monocytes and M1 macrophages, which can function as efficient immune effector cells and promote antitumor immune responses, were suppressed in patients with high hypoxia risk scores. Additionally, our results showed that PD-1 and CTLA-4 were significantly upregulated in the high hypoxia risk score groups. However, CD8^{+} T cells and activated natural killer cells were higher in the high hypoxia risk score groups. These results indicate that hypoxia condition has multiple effects on the immune microenvironment. The hypoxia-related risk model may be useful for predicting the immunotherapy response. Improving oxygen deficiency may decrease immunosuppression in the TME of RCC, which may benefit immunotherapy.

It remains difficult to predict or explain the clinical response rate of RCC immunotherapy in practice. However, hypoxia has been reported to lead to immunosuppression and tumor progression (Li et al., 2018). Hypoxia-induced changes in the TME have also been reported as a barrier to immunotherapy in pancreatic adenocarcinoma (Daniel et al., 2019). Furthermore, inhibition of hypoxic stress-relevant pathways can enhance antitumor immunity and improve the response rate of immunotherapy. mTOR inhibitors are applied in current therapies for RCC to target HIF translation. VEGFA inhibitors target the function of HIF-target genes. This may help prevent drug resistance and enhance immunotherapy.

There were some limitations to this study. First, the results are based on data collected from TCGA database. Although the results were verified in the ICGC database with 91 cases, the potential for selection bias cannot be avoided, and it is impossible to collect all clinical information from the patients. Second, the results are descriptive, and *in vitro* or *in vivo* experiments were not performed to clarify the exact immune microenvironment of RCC. Third, further clinical trials are needed to validate the prognostic prediction power of the hypoxia-related risk model. Last but not least, the comparison between different tools for predicting the prognosis of RCC, which may lead to a more objective evaluation of the novel hypoxia-related risk model, is not included in our study. Despite these limitations and lack of further validation in more studies, the presented findings applied a hypoxia-related risk model in RCC prognosis predicting and statistically proved its performance.

We developed a hypoxia-related risk model based on eight identified hypoxia-related genes. The model was validated as an independent predictor of the prognosis of patients with RCC. We hope that the hypoxia-related risk model can be used as a prognostic biomarker in patients with RCC, which may be helpful for underpinning clinical decision-making in the future. Moreover, the hypoxia-related risk model may reflect the immune microenvironment of RCC and help improve the overall clinical response to immunotherapy.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YM and QZ contributed to the design and supervision of the study and contributed to manuscript revision. ZZ and QL contributed to the online data search, acquisition, interpretation, and contributed to manuscript writing. FW and BM contributed to the data extraction. All authors made substantial contributions to the study and have approved the study and are responsible for their own contributions to the study personally.

FUNDING

This work was supported by the Tibetan Natural Science Foundation (Grant No. XZ2017ZR-ZY019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.606816/full#supplementary-material

REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Brooks, S. A., Brannon, A. R., Parker, J. S., Fisher, J. C., Sen, O., Kattan, M. W., et al. (2014). ClearCode34: a prognostic risk predictor for localized clear cell renal cell carcinoma. *Eur. Urol.* 66, 77–84. doi: 10.1016/j.eururo.2014.02.035

Choueiri, T. K., and Kaelin, W. G. Jr. (2020). Targeting the HIF2-VEGF axis in renal cell carcinoma. *Nat. Med.* 26, 1519–1530. doi: 10.1038/s41591-020-1093-z

Daniel, S. K., Sullivan, K. M., Labadie, K. P., and Pillarisetty, V. G. (2019). Hypoxia as a barrier to immunotherapy in pancreatic adenocarcinoma. *Clin. Transl. Med.* 8:10. doi: 10.1186/s40169-019-0226-9

Facciabene, A., Motz, G. T., and Coukos, G. (2012). T-regulatory cells: key players in tumor immune escape and angiogenesis. *Cancer Res.* 72, 2162–2171. doi: 10.1158/0008-5472.CAN-11-3687

Fallah, J., and Rini, B. I. (2019). HIF inhibitors: status of current clinical development. *Curr. Oncol. Rep.* 21:6. doi: 10.1007/s11912-019-0752-z

Guo, L., Wang, C., Qiu, X., Pu, X., and Chang, P. (2020). Colorectal cancer immune infiltrates: significance in patient prognosis and immunotherapeutic efficacy. *Front. Immunol.* 11:1052. doi: 10.3389/fimmu.2020.01052

Haase, V. H. (2013). Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Rev.* 27, 41–53. doi: 10.1016/j.blre.2012.12.003

Hirata, H., Hinoda, Y., Nakajima, K., Kikuno, N., Suehiro, Y., Tabatabai, Z. L., et al. (2009). The bcl2−938CC genotype has poor prognosis and lower survival in renal cancer. *J. Urol.* 182, 721–727. doi: 10.1016/j.juro.2009.03.081

Ishibashi, K., Koguchi, T., Matsuoka, K., Onagi, A., Tanji, R., Takinami-Honda, R., et al. (2018). Interleukin-6 induces drug resistance in renal cell carcinoma. *Fukushima J. Med. Sci.* 64, 103–110. doi: 10.5387/fms.2018-15

Klatte, T., Rossi, S. H., and Stewart, G. D. (2018). Prognostic factors and prognostic models for renal cell carcinoma: a literature review. *World J. Urol.* 36, 1943–1952. doi: 10.1007/s00345-018-2309-4

LaGory, E. L., and Giaccia, A. J. (2016). The ever-expanding role of HIF in tumour and stromal biology. *Nat. Cell Biol.* 18, 356–365. doi: 10.1038/ncb3330

Lewis, C., and Murdoch, C. (2005). Macrophage responses to hypoxia: implications for tumor progression and anti-cancer therapies. *Am. J. Pathol.* 167, 627–635. doi: 10.1016/S0002-9440(10)62038-X

Li, Y., Patel, S. P., Roszik, J., and Qin, Y. (2018). Hypoxia-driven immunosuppressive metabolites in the tumor microenvironment: new approaches for combinational immunotherapy. *Front. Immunol.* 9:1591. doi: 10.3389/fimmu.2018.01591

Linehan, W. M., and Ricketts, C. J. (2019). The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications. *Nat. Rev. Urol.* 16, 539–552. doi: 10.1038/s41585-019-0211-5

Lipworth, L., Morgans, A. K., Edwards, T. L., Barocas, D. A., Chang, S. S., Herrell, S. D., et al. (2016). Renal cell cancer histological subtype distribution differs by race and sex. *BJU Int.* 117, 260–265. doi: 10.1111/bju.12950

Liu, X., Zhang, X., Peng, Z., Li, C., Wang, Z., Wang, C., et al. (2020). Deubiquitylase OTUD6B governs pVHL stability in an enzyme-independent manner and suppresses hepatocellular carcinoma metastasis. *Adv. Sci.* 7:1902040. doi: 10.1002/advs.201902040

Martínez-Sáez, O., Gajate Borau, P., Alonso-Gordoa, T., Molina-Cerrillo, J., and Grande, E. (2017). Targeting HIF-2 α in clear cell renal cell carcinoma: a promising therapeutic strategy. *Crit. Rev. Oncol. Hematol.* 111, 117–123. doi: 10.1016/j.critrevonc.2017.01.013

McAllister, S. S., and Weinberg, R. A. (2014). The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. *Nat. Cell Biol.* 16, 717–727. doi: 10.1038/ncb3015

Millet-Boureima, C., He, S., Le, T. B. U., and Gamberi, C. (2021). Modeling neoplastic growth in renal cell carcinoma and polycystic kidney disease. *Int. J. Mol. Sci.* 22:3918. doi: 10.3390/ijms22083918

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Schödel, J., Grampp, S., Maher, E. R., Moch, H., Ratcliffe, P. J., Russo, P., et al. (2016). Hypoxia, hypoxia-inducible transcription factors, and renal cancer. *Eur. Urol.* 69, 646–657. doi: 10.1016/j.eururo.2015.08.007

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, C., Liu, J., Wang, J., Zhong, X., Dong, D., Yang, X., et al. (2020). Development and validation of a prognostic immune-associated gene signature in clear cell renal cell carcinoma. *Int. Immunopharmacol.* 81:106274. doi: 10.1016/j.intimp.2020.106274

Shi, L., An, S., Liu, Y., Liu, J., and Wang, F. (2020). PCK1 regulates glycolysis and tumor progression in clear cell renal cell carcinoma through LDHA. *Onco Targets Ther.* 13, 2613–2627. doi: 10.2147/OTT.S241717

Shou, Y., Yang, L., Yang, Y., Zhu, X., Li, F., and Xu, J. (2021). Determination of hypoxia signature to predict prognosis and the tumor immune microenvironment in melanoma. *Mol. Omics* 17, 307–316. doi: 10.1039/D0MO00159G

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Straussman, R., Morikawa, T., Shee, K., Barzily-Rokni, M., Qian, Z. R., Du, J., et al. (2012). Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. *Nature* 487, 500–504. doi: 10.1038/nature11183

Syafruddin, S. E., Rodrigues, P., Vojtasova, E., Patel, S. A., Zaini, M. N., Burge, J., et al. (2019). A KLF6-driven transcriptional network links lipid homeostasis and tumour growth in renal carcinoma. *Nat. Commun.* 10:1152. doi: 10.1038/s41467-019-09116-x

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43(Database issue), D447–D452. doi: 10.1093/nar/gku1003

Terry, S., Buart, S., and Chouaib, S. (2017). Hypoxic stress-induced tumor and immune plasticity, suppression, and impact on tumor heterogeneity. *Front. Immunol.* 8:1625. doi: 10.3389/fimmu.2017.01625

van Dijk, N., Funt, S. A., Blank, C. U., Powles, T., Rosenberg, J. E., and van der Heijden, M. S. (2019). The cancer immunogram as a framework for personalized immunotherapy in urothelial cancer. *Eur. Urol.* 75, 435–444. doi: 10.1016/j.eururo.2018.09.022

Vilà, M. R., Nicolás, A., Morote, J., de, I., and Meseguer, A. (2000). Increased glyceraldehyde-3-phosphate dehydrogenase expression in renal cell carcinoma identified by RNA-based, arbitrarily primed polymerase chain reaction. *Cancer* 89, 152–164. doi: 10.1002/1097-0142(20000701)89:1<152::AID-CNCR20>3.0.CO;2-T

Wei, X., Choudhury, Y., Lim, W. K., Anema, J., Kahnoski, R. J., Lane, B., et al. (2017). Recognizing the continuous nature of expression heterogeneity and clinical outcomes in clear cell renal cell carcinoma. *Sci. Rep.* 7:7342. doi: 10.1038/s41598-017-07191-y

Xu, W. H., Xu, Y., Wang, J., Wan, F. N., Wang, H. K., Cao, D. L., et al. (2019). Prognostic value and immune infiltration of novel signatures in clear cell renal cell carcinoma microenvironment. *Aging (Albany NY)* 11, 6999–7020. doi: 10.18632/aging.102233

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhao, J., Du, F., Shen, G., Zheng, F., and Xu, B. (2015). The role of hypoxia-inducible factor-2 in digestive system cancers. *Cell Death Dis.* 6:e1600. doi: 10.1038/cddis.2014.565

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership