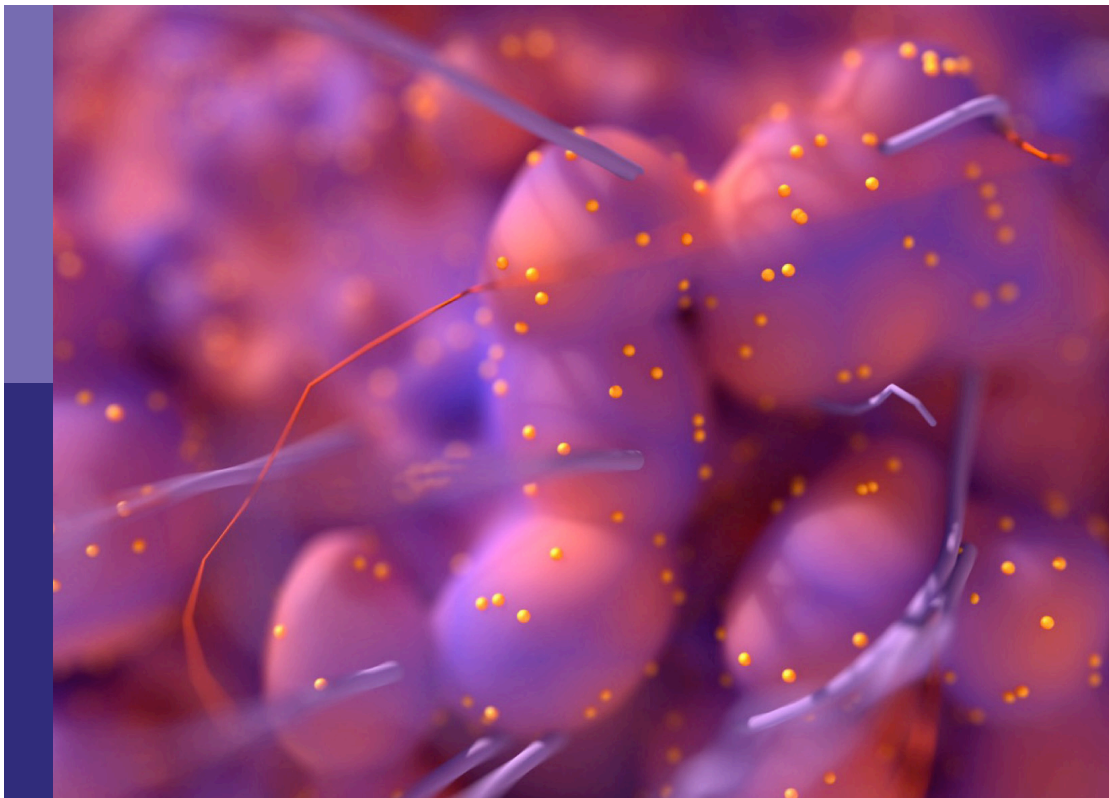
frontiers | Research Topics

Artificial intelligence: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment

Edited by
Mónica Hebe Vazquez-Levin, Jaume Reventos and George Zaki

Published in
Frontiers in Oncology
Frontiers in Genetics
Frontiers in Artificial Intelligence



ISSN 1664-8714
ISBN 978-2-83252-180-9
DOI 10.3389/978-2-83252-180-9

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: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment

Topic editors

Mónica Hebe Vazquez-Levin — National Scientific and Technical Research Council (CONICET), Argentina

Jaume Reventos — Institut d'Investigacio Biomedica de Bellvitge (IDIBELL), Spain

George Zaki — Frederick National Laboratory for Cancer Research, National Cancer Institute at Frederick (NIH), United States

Citation

Vazquez-Levin, M. H., Reventos, J., Zaki, G., eds. (2023). *Artificial intelligence: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-180-9

Table of contents

frontiers | Frontiers in Oncology

TYPE Editorial
PUBLISHED 31 March 2023
DOI 10.3389/fonc.2023.1161118

OPEN ACCESS

EDITED AND REVIEWED BY
Claudio Sette,
Catholic University of the Sacred Heart, Rome, Italy

*CORRESPONDENCE
Mónica Hebe Vazquez-Levin
mhvazl@gmail.com;
mhvazquez@ibyme.conicet.gov.ar

SPECIALTY SECTION
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

RECEIVED 07 February 2023
ACCEPTED 20 February 2023
PUBLISHED 31 March 2023

CITATION
Vazquez-Levin MH, Reventos J and Zaki G (2023) Editorial: Artificial intelligence: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment. *Front. Oncol.* 13:1161118. doi: 10.3389/fonc.2023.1161118

Editorial: Artificial intelligence: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment

Mónica Hebe Vazquez-Levin[1]*, Jaume Reventos[2] and George Zaki[3]

[1]Instituto de Biología y Medicina Experimental (IBYME), Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET) Fundación IBYME (FIBYME), Buenos Aires, Argentina, [2]Institut d'Investigacio Biomedica de Bellvitge (IDIBELL) and Universitat Internacional de Catalunya, Barcelona, Spain, [3]Frederick National Laboratory for Cancer Research, National Cancer Institute at Frederick (NIH), Frederick, MD, United States

KEYWORDS

cancer, artificial intelligence, machine learning, digital twin, precision medicine

Editorial on the Research Topic

Artificial intelligence: A step forward in biomarker discovery and integration towards improved cancer diagnosis and treatment

In cancer, a biomarker refers to a substance or process indicative of the presence of cancer in the body. However, the idea of "one-molecule (or process) marker" indicated by its presence, and the existence of an undergoing transforming cancer process is currently a utopia. During the past decade, there has been a fundamental shift in cancer research and clinical decision-making, moving from qualitative data to quantitative digital data. A large wealth of cancer biomarkers and images has come from research laboratories and clinical institutions worldwide. Moreover, the major bulk of information has arisen from genomics, proteomics, metabolomics, and other omics, but also from oncology clinics, imaging, epidemiology and more. Artificial Intelligence (AI) is a unique technology that is able to combine all the above, and particularly suited to establish novel therapies and predictive models of drug response (1, 2). The combination of several biomarkers, by means of Machine Learning (ML) algorithms, would reach unprecedented conclusions in diagnosis, prediction and general decision making of novel anticancer therapies (3–5). In addition, the multimodal temporal data collected from patients with cancers can feed to initialize and track a Digital Twin to experiment with multiple possible treatments *in silico*.

This Research Topic has gathered 10 selected contributions in the area of ML tools, Deep Learning and Cancer Digital Twin technologies in the field of Precision Oncology, and contains one review, one minireview and eight original contributions.

The review paper by Asada et al. emphasizes the relevance of Precision Oncology and the integration of whole genome sequencing analysis, epigenome analyses and the use of ML, and opens a discussion about future perspectives in the field.

Networks of cellular systems and arrays of biological models described by ordinary and partial differential equations were developed in the last decades towards a better understanding of biological systems. Mertins minireview describes the use of ML algorithms to analyze computational dynamic ordinary differential equation models in combination with omics data, towards the discovery of novel biomarkers and novel molecular targets.

Tumor cell heterogeneity has been for many years a distortion factor in the interpretation of cellular and molecular findings in oncology. Tumor degree of purity is playing an important role in optimizing the correlation of the research findings with therapeutic anticancer strategies. By using Random Forest ML, Yang et al. were able to assess tumor purity in children CNS tumors, which will imply genomic, biological and clinical implications.

More than 60% of cervical cancers are caused by Human Papilloma Virus (HPV) 16 genotype, classified into lineages A, B, C, and D. In their contribution, Asensio-Puig et al. report the development of a Random Forest-based new model to assess HPV16 lineage. Authors highlight that their model is 40 times faster than current assessment done with Maximum Likelihood Tree, which requires a manual annotation and cannot assess poorly sequenced samples.

The work by Chang et al. proposes a novel ML predictive model utilizing a three-Dimensional Convolutional Neural Network (3D-CNN) to predict the presence of lymph node metastasis and the postoperative positive margin status based on preoperative CT scans. Their report provides a proof of concept for the preoperatively use of radiomics and 3D-CNN deep learning framework to improve the prediction of positive resection margins as well as the presence of lymph node metastatic disease.

In the report by Spiller et al., the utility and feasibility of imaging, computer vision and ML to determine patient-derived organoids vital status is reported. By acquiring bright field images at different time points without relying upon vital dyes, authors track the dynamic response of individual organoids to various drugs. In addition, authors report a web-based data visualization tool, called the Organoizer, available for public use.

Abousamra et al. present a Deep Learning workflow that generates Tumor Infiltrating Lymphocytes (TIL) maps to study their abundance and spatial distribution in 23 cancer types. Authors trained three state-of-the-art CNN architectures (namely VGG16, Inception-V4, ResNet-34) with training data from The Cancer Genome Atlas, combining manual annotations from pathologists and computer-generated labels from a first-generation TIL model. It also incorporates automated thresholding to convert model predictions into binary classifications to generate TIL maps.

With the aim to identify putative biomarkers for lung cancer and to elucidate the pathogenesis of this disease, Bahado-Singh et al. combined AI and DNA methylation analysis of circulating cell-free tumor DNA. The study analyzes six AI platforms, including Support Vector ML and Deep Learning, to measure cytosine (CpG) methylation changes across the genome in lung cancer. Training sets and validation sets are generated and 10-fold cross validation performed. To elucidate lung cancer pathogenesis, gene enrichment analysis using g:profiler and GREAT enrichment is done.

Triple-negative breast cancer (TNBC) always requires neoadjuvant chemotherapy (NACT) for a pathological complete response and improved long-term survival. Irajizad et al. previously identified a polyamine biomarker suitable to assess which patient will respond to NACT. In their contribution, Irajizad et al. identified TNBC patients who will be insensitive to NACT, by using ML methods.

Finally, the contribution by Batch et al. aimed to improve the detection of metastatic disease over time from structured radiology reports with the ultimate goal of building and updating a Digital Twin to model long-term prognosis. By exposing prediction models to historical information using Natural Language Processing (NLP), the authors were able to extract and encode relevant features from medical text reports, and use these features to develop, train, and validate models. Over 700 thousand radiology reports were used for model development to predict the presence of metastatic disease. The model uses features from consecutive structured patient text radiology reports. Three models were developed to classify the type of metastatic disease: a simple CNN, a CNN augmented with an attention layer, and Recurrent Neural Network labels. To develop the models, a subset of the reports was curated for ground-truth. Results from the three models were compared (accuracy, precision, recall, and F1-score) to a single-report model previously developed to analyze one report instead of multiple past reports. Results suggest that NLP models can extract cancer progression patterns from multiple consecutive reports and predict the presence of metastatic disease in multiple organs with higher performance when compared with a single-report-based prediction.

In summary, contributions to this special edition highlight how AI will accelerate the advancement of Personalized Medicine and cancer care, by improving patient diagnosis, treatment, and prognosis.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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

References

1. Kehl KL, Xu W, Gusev A, Bakouny Z, Choueiri TK, Riaz IB, et al. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. *Nat Commun* (2021) 12(1):7304. doi: 10.1038/s41467-021-27358-6

2. You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, et al. Artificial intelligence in cancer target identification and drug discovery. *Sig Transduct Target Ther* (2022) 7(1):156. doi: 10.1038/s41392-022-00994-0

3. Koh DM, Papanikolaou N, Bick U, Illing R, Kahn CE Jr, Kalpathi-Cramer J, et al. Artificial intelligence and machine learning in cancer imaging. *Commun Med (Lond)* (2022) 2:133. doi: 10.1038/s43856-022-00199-0

4. Kong J, Ha D, Lee J, Kim I, Park M, Im SH, et al. Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nat Commun* (2022) 13(1):3703. doi: 10.1038/s41467-022-31535-6

5. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat Commun* (2022) 13(1):4013. doi: 10.1038/s41467-022-31666-w

frontiers in Oncology

REVIEW
published: 12 May 2021
doi: 10.3389/fonc.2021.666937

Integrated Analysis of Whole Genome and Epigenome Data Using Machine Learning Technology: Toward the Establishment of Precision Oncology

Ken Asada[1,2*†], *Syuzo Kaneko*[1,2†], *Ken Takasawa*[1,2], *Hidenori Machino*[1,2], *Satoshi Takahashi*[1,2], *Norio Shinkai*[1,2,3], *Ryo Shimoyama*[1,2], *Masaaki Komatsu*[1,2] *and Ryuji Hamamoto*[1,2,3*]

[1] *Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan,* [2] *Division of Medical AI Research and Development, National Cancer Center Research Institute, Tokyo, Japan,* [3] *Department of NCC Cancer Science, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan*

With the completion of the International Human Genome Project, we have entered what is known as the post-genome era, and efforts to apply genomic information to medicine have become more active. In particular, with the announcement of the Precision Medicine Initiative by U.S. President Barack Obama in his State of the Union address at the beginning of 2015, "precision medicine," which aims to divide patients and potential patients into subgroups with respect to disease susceptibility, has become the focus of worldwide attention. The field of oncology is also actively adopting the precision oncology approach, which is based on molecular profiling, such as genomic information, to select the appropriate treatment. However, the current precision oncology is dominated by a method called targeted-gene panel (TGP), which uses next-generation sequencing (NGS) to analyze a limited number of specific cancer-related genes and suggest optimal treatments, but this method causes the problem that the number of patients who benefit from it is limited. In order to steadily develop precision oncology, it is necessary to integrate and analyze more detailed omics data, such as whole genome data and epigenome data. On the other hand, with the advancement of analysis technologies such as NGS, the amount of data obtained by omics analysis has become enormous, and artificial intelligence (AI) technologies, mainly machine learning (ML) technologies, are being actively used to make more efficient and accurate predictions. In this review, we will focus on whole genome sequencing (WGS) analysis and epigenome analysis, introduce the latest results of omics analysis using ML technologies for the development of precision oncology, and discuss the future prospects.

Keywords: artificial intelligence, whole genome analysis, epigenome analysis, machine learning, biomarker discovery, cancer diagnosis and treatment, precision oncology

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Vandhana Krishnan, Stanford University, United States
Francesco Schettini, Institut de Recerca Biomèdica August Pi i Sunyer (IDIBAPS), Spain

****Correspondence:***
Ken Asada
ken.asada@riken.jp
Ryuji Hamamoto
rhamamot@ncc.go.jp

†These authors share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *11 February 2021*
Accepted: *26 April 2021*
Published: *12 May 2021*

Citation:
Asada K, Kaneko S, Takasawa K, Machino H, Takahashi S, Shinkai N, Shimoyama R, Komatsu M and Hamamoto R (2021) Integrated Analysis of Whole Genome and Epigenome Data Using Machine Learning Technology: Toward the Establishment of Precision Oncology. Front. Oncol. 11:666937. doi: 10.3389/fonc.2021.666937

INTRODUCTION

The structure of DNA was first reported by Watson and Crick in 1953 (1). Following this, the first sequencing technique known as the Sanger sequencing method was developed in 1977 (2). In 1987, the first automatic sequencing machine (AB370) was introduced by Applied Biosystems, which uses capillary electrophoresis without the need for a gel, which enabled the sequencing process to be more convenient in terms of accuracy and time (3). This technology truly accelerated the completion of the International Human Genome Project, which was aimed at decoding three billion human nucleotide base pairs (4). With the completion of the International Human Genome Project, the era known as the post-genome era began, and attempts to apply genomic information to medicine began to be actively pursued. Consequently, the concept of personalized medicine has also come to attract attention (5–7). Under such circumstances, the advent of a new analysis method called next-generation sequencing (NGS) technology has rapidly accelerated the speed of nucleotide sequence analysis and dramatically lowered the cost of performing whole genome analysis (8, 9). As a result, genome-wide analysis can now be performed routinely. In addition to DNA sequence analysis, various analysis methods using NGS technology have emerged, such as RNA sequencing (RNA-seq) for gene expression analysis, chromatin immunoprecipitation sequencing (ChIP-seq) for histone modification analysis and identification of transcription factor binding sites, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) and Hi-C for chromatin structure analysis (10, 11) (**Figure 1**). Along with technological innovation, there have also been attempts to apply genomic information to actual clinical practice. Targeted-gene panels (TGPs), which use NGS to examine the mutation status of a limited number of cancer-related genes, are actively being used to select the optimal treatment (12–14). On the other hand, one of the major problems in promoting precision oncology using the TGP method is that the number of patients who will benefit from the information obtained by the TGP method alone is limited (15–17). In order to increase the number of patients who will benefit from the promotion of precision oncology in the future, it is necessary to add more detailed omics data, such as whole genome analysis data and epigenome data, for integrated analysis. In recent years, it has been reported that epigenomic abnormalities play an important role in the development and progression of cancer (10, 18–25), and it is important to take into account information on epigenomic abnormalities when genomic mutations alone cannot elucidate the molecular mechanisms. In fact, the concept of epigenetic driver (epi-driver) is currently being used to describe the phenomenon of cancer development and progression based on epigenomic abnormalities (26, 27).

Another important issue is that the amount of data that researchers have to deal with has become enormous due to the emergence of various new methods with NGS analysis at their core as a result of technological innovation. For example, the amount of data generated by a single NGS run can be up to a million times larger than the data generated by a single Sanger sequencing run (28). In addition, there is a growing need for multimodal analysis, such as integrated analysis of genomic and epigenomic data, not just data from one modality. This kind of advanced analysis using a large amount of data is difficult to perform using conventional statistical methods, but nowadays, by proactively introducing artificial intelligence (AI) with machine learning (ML) and deep learning (DL) technologies at its core, good results can be obtained (29–31). In our view, there are four properties of ML and DL that are of particular importance. First, multimodal learning, which allows us to integrate multiple omics data as input (32–35). Second, multitask learning, which allows us to learn multiple different tasks simultaneously by sharing parts of the model (36, 37). Third, representation learning and semi-supervised learning, which allows us to acquire representations of data from large amounts of unlabeled data and thereby obtain small amounts of labels (38–41). The fourth is the ability to automatically acquire

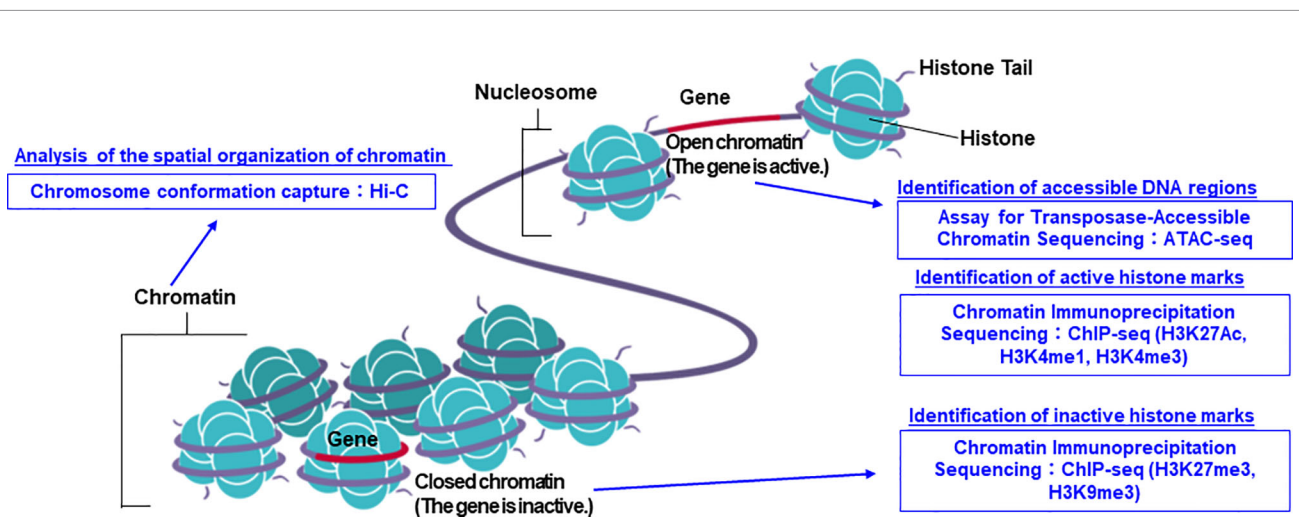


FIGURE 1 | The summarized figure of chromatin structure and epigenomic analysis methods. ChIP-seq, ATAC-seq, and Hi-C methods can be used to predict the state of transcriptional activation or inactivation, and chromatin structure. Image credit: Shutterstock.com/ellepigrafica.

hierarchical features to capture higher-order correlations in the input (10, 42). More importantly, AI has already become one of the key technologies in the medical field, with a number of AI-powered medical devices approved by the US FDA (43). Under these circumstances, the active introduction of AI in the field of precision oncology seems to be an inevitable trend in the future.

Therefore, this review introduces the current status of efforts to establish precision oncology, focusing on whole genome sequencing (WGS) analysis and epigenome analysis, with particular emphasis on the results obtained through the use of ML and DL technologies.

WHOLE GENOME ANALYSIS

In this section, we introduce the recently published up to date WGS analyses using ML and DL. The cost of WGS dropped from 100 million US dollars in 2001 to 1,000 dollars in 2020 (NIH National Human Genome Research Institute; https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost; Cost per genome data - 2020). In 2020, an international collaboration to identify common mutation patterns in more than 2,600 cancer whole genomes was performed by the Cancer Genome Atlas Research Network as The Cancer Genome Atlas Pan-Cancer Analysis of Whole Genomes (PCAWG) project (44). The results described in the flagship paper were accompanied with related papers that focused on specific analysis, such as peak calls, structural variations (SV), and non-coding variants.

As summarized in **Table 1**, we categorized WGS analyses into five groups based on the purpose of their use. The first type of analysis considered is peak calling. Finding an accurate peak calling is one of the most important and difficult parts of WGS analysis. Aligning several hundred bps to the whole genome (three billion bps in length) while considering sequencing errors is technically challenging (65, 66). Thus, reports comparing the benchmarks and new pipelines, particularly deep neural networks (DNNs), have been published for both peak calling and the identification of variants (45–51) in **Table 1**. In general, DNN models were first trained with publicly available datasets followed by the evaluation of their performance with the test dataset. Validation is performed with the validation dataset either using publicly available data or their in-house dataset. For example, the WGS dataset obtained from the PCAWG was used for training and testing the model. To independently validate the DNN model, the authors assembled several datasets outside the PCAWG (67).

The second analysis type is a genome graph or graph-based genome alignment. This approach has been recently reported and summarized (68). The advantage of using genome graphs is that they can accurately map (genotype) the polymorphisms of genomes with a good visualization, as well as perform fast and memory-efficient alignments (52–55) in **Table 1**. There is increasing recognition that a single, linear, monoploid reference genome is not always the best reference structure for human genetics, because they represent only a small fraction of existing human variations, particularly when they span SV breakpoints.

Third, heterogeneity in samples can be analyzed. Cancers are often observed to have various morphologies. These types of results are inconsistent with peak calls because they reflect where tissue samples are dissected. However, it is also true that tumors are composed of subpopulations of cells, and some cancer cells can migrate to other tissues. This heterogeneity results in a variety of features that can affect cancer phenotypes. To handle this, some published papers specifically focused on and investigated these phenotypes (56–58) in **Table 1**.

The fourth category is mutational signatures. The patterns of mutation or substitution signatures in cancer genome are discernible. Therefore, to categorize them, mutational signatures have been reported. Mutational signature analysis algorithms produce a decomposition matrix by using ML, a non-negative matrix factorization (NMF) approach, to extract mutational signatures (69–72). Additionally, other pipelines have been reported to perform mutational signature analyses to classify the samples (59–61) in **Table 1**.

The last is ML in a genome-wide association study (GWAS). GWAS has been used to discover genetic variants that are associated with diseases (73). To improve the analysis of GWAS, a combination of ML and DL analyses was reported (62) in **Table1**. However, how to improve mapping of regulatory variants (non-coding regions) identified by GWAS is still on going. Therefore, Arloth et al. developed DL-based approach and showed SNPs identified by DL were nominally significant in classical univariate GWAS analysis (63) in **Table 1**. They also identified disease/trait-relevant transcriptionally active genomic loci by integrating gene expression and DNA methylation quantitative trait loci (eQTL and meQTL) information of multiple resources and tissues. Although this is not a cancer research, another ML- and DL-based approach using GWAS data showed a good classification of amyotrophic lateral sclerosis (ALS) patient, and this approach can identify potentially ALS-associated promoter regions (64) in **Table 1**.

By integrating other omics data and analyzing single nucleotide variants (SNVs), indels, SV, and copy number alterations in non-coding regions, researchers can address the question of how pan-negative cancers developed, which we introduce in the following sections.

DNA METHYLATION

DNA methylation is an epigenetic modification that can discriminate specific patterns between in normal tissue cells and in cancer cells (74, 75). These epigenetic alterations affect gene expression, and thus, cell-specific DNA methylation patterns are used in the diagnosis and treatment selection of cancer by identifying cancer-specific DNA methylation patterns in biopsy specimens and blood samples (76, 77). A few diagnostic measures utilizing cancer-specific DNA methylation patterns have already received FDA approval (78, 79). Moreover, ML and DL analyses have been increasingly used to identify novel disease-specific DNA methylation patterns; they have also been used in research that aims to utilize the DNA methylation data

TABLE 1 | Overview of whole genome analysis using machine learning.

Features	Pipeline name	Brief summary	Reference
Peak calling, mutational signature, or *de novo* assembly	HipSTR (Haplotype inference and phasing for short tandem repeat)	This method identifies *de novo* STRs; genotyping 1.6 million STRs in the human genome using HipSTR can be done in an average of 10 CPU hours per sample.	Nat. Methods (2017) (45)
	BayesTyper	This method performs genotyping of all types of variation (including SNPs, indels and complex structural variants) based on an input set of variants and read k-mer counts.	Nature (2017) (46)
	Genomiser	This method identifies pathogenic regulatory variants in non-coding regions.	Am. J. Hum. Genet (2016) (47).
	DeepVariant	This is a universal SNP and small-indel variant caller using deep neural networks, highlighting the benefits of using automated and generalizable techniques for variant calling.	Nat. Biotechnol (2018) (48).
	ARC (Artifact Removal by Classifier)	This is a supervised random forest model designed to distinguish true rare *de novo* variants (RDNVs) from genetic aberrations specific to lymphoblastoid cell lines (LCLs) or other types of artifacts, such as sequencing and mapping errors.	Cell (2019) (49)
	N/A	This method addresses the challenge of detecting the contribution of non-coding variants to disease using a deep learning-based framework that predicts the specific regulatory and detrimental effects of genetic variants.	Nat. Genet (2019) (50).
	NeuroSomatic	This is a convolutional neural network for somatic mutation detection.	Nat.Commun (2019) (51).
Genome graph	Graphtyper	This is an algorithm and software for discovering and genotyping sequence variation, which rearranges short read sequence data into a pan-genome and creates a graph structure that takes into account the mutations that encode sequence variation in a population by representing possible haplotypes as graph paths.	Nat. Genet (2017) (52).
	N/A	The results of the missing mutations are added to a structure that can be described as a mathematical graph, the genome graph. Compared to the existing reference genome map (GRCh38), the genome graph can significantly improve the percentage of reads that map uniquely and completely.	bioRxiv (2017) (53)
	GenGraph	This provides a set of tools for generating graph-based representations of sets of sequences.	BMC Bioinformatics (2019) (54)
	N/A	This is a SV caller that uses genome graphs, which is used to analyze cancer somatic DNA rearrangements and revealed three novel complex rearrangement phenomena.	Cell (2020) (55)
Heterogeneity	PyClone	This is a Bayesian clustering method for grouping sets of deeply sequenced somatic mutations into putative clonal clusters while estimating their cellular prevalences and accounting for allelic imbalances introduced by segmental copy-number changes and normal-cell contamination.	Nat. Methods (2014) (56)
	MOBSTER	This is an approach for model-based tumor subclonal reconstructions. Cancer genomic data are generated from bulk samples composed of mixtures of cancer subpopulations, as well as normal cells. Subclonal reconstruction methods based on machine learning aim to separate those subpopulations in a sample and infer their evolutionary history.	Nat. Genet (2020) (57).
	DigiPico/MutLX	This method is a powerful framework for the identification of clone-specific variants with high accuracy.	ELife (2020) (58)
Mutational signature	SigMA (signature multivariant analysis)	This provides an accurate identification of mutational signatures with a likelihood approach, even when the mutation count is very small.	Nat. Genet (2019) (59).
	DeepMS (deep learning of mutational signature)	This is a regression-based model to estimate the correlation between signatures and clinical and demographical phenotypes in order to identify mutational signatures.	Oncogenes (2020) (60)
	SigLASSO	This method performs efficient cancer mutation signature analysis by accounting for sampling uncertainty, and also improves performance by allowing knowledge transfer through cooperative fitting of linear mixtures and maximizing sampling likelihood.	Nat. Commun (2020) (61).
GWAS	COMBI	This is a two-step algorithm that trains a support vector machine to determine candidate SNPs and then performs hypothesis testing on these SNPs.	Sci Rep (2016) (62).
	DeepWAS	This integrates regulatory effects predictions of single variants into a multivariate GWAS setting and provide evidence that DeepWAS results directly identify disease/trait-associated SNPs with a common effect on a specific chromatin feature.	PLoS Comput. Biol (2019) (63).
	Promoter-CNN + ALS-Net	This is a DL-based approach for genotype-phenotype association studies to predict the occurrence of ALS from individual genotype data. A two step-approach employs (1); promoter regions that are likely associated to ALS are identified and (2) individuals are classified based on their genotype in the selected genomic regions.	Bioinformatics (2019) (64)

from cancer patients for diagnosis, staging, and prognosis predictions (80–83).

Cell-free DNA (cfDNA) is circulating DNA found in plasma, and is known to be elevated in cancer patients (84). The clinical significance of analyzing cfDNA is that (1) it is noninvasive (2), it can be applied for monitoring, and (3) it can detect a more global signature compared to the data obtained from a biopsy on a single metastatic site. Therefore, ML can be applied for DNA methylation analyses using cfDNA. The DNA methylation levels of plasma cfDNA in renal cell carcinoma (RCC) patients have been assessed by cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq), and RCC

detection was performed using the elastic net regularized generalized linear model method (80). In this aforementioned study, DNA methylation data obtained from blood and urine samples were used for validation, and the area under the receiver operating characteristic (AUROC) curve was found to be of 0.99 for blood samples and 0.86 for urine samples, respectively. In another study, cfDNA methylation data from blood samples of patients with intracranial tumors were obtained with cfMeDIP-seq and successfully used to generate a cancer detection model using the Random Forest algorithm (81). This model was also shown to have high discriminative capacity among the five tumor types (isocitrate dehydrogenase (IDH) wild-type glioma, IDH mutant glioma, low-grade glial-neuronal, hemangiopericytoma, and meningioma).

Next, we review DNA methylation analyses that use solid tumor samples. First, to distinguish metastatic head and neck squamous cell carcinoma (HNSC) from primary squamous cell carcinoma of the lung (LUSC), DNA methylation data were extracted from surgical specimens of lung cancer patients and artificial neural networks (NN), and a support vector machine (SVM) and a random forest (RF) classifier was constructed because current diagnostics show no possibility to distinguish metastatic HNSC from primary LUSC. Authors developed models that classified 96.4% of the cases by NN, 95.7% by SVM, and 87.8% by RF (82). The DL-based approach is also used to detect DNA methylation patterns related to breast cancer metastases and predict recurrence by conducting feature selection using an autoencoder with a single hidden layer followed by ML techniques for classification, or enrichment analysis for finding a biological relevance, genomic context, and functional annotation of best genes (83).

CANCER EPIGENETICS WITH A FOCUS ON ENHANCER FUNCTION

As mentioned earlier, since the advent of NGS technology and analyses based on ML, remarkable progress has been made in understanding the genetic basis of cancer. These studies have mainly defined genetic alterations as either causal (driver mutations), which confer a selective advantage to cancer cells, or consequential (passenger mutations, not directly causal), which do not have a selective advantage (26). Furthermore, genomic sequencing of tumor samples has revealed that different patients share a unique combination of one or two strong driver mutations such as gain-of-function EGFR and loss-of-function TP53 mutations typically detected in lung cancer and less frequent driver mutations (85, 86). On the other hand, the genetic component of the general disease risk is distributed mainly in the non-coding regions, which seem to be particularly rich in enhancers specific to the cell types associated with the disease (87, 88). Therefore, this has led to a growing interest in the annotation and understanding of human enhancers.

Measurable genome-wide biochemical annotations for enhancer regions include ChIP-seq or cleavage under targeted and release using nucleases (CUT&RUN) assays (89) for histone modifications or transcription factor (TF) binding, DNase I hypersensitivity sequencing (DNase-seq) for open chromatin (90), and ATAC-seq (91). On the other hand, it has long been hypothesized that enhancers loop in 3D space to access their target promoters. In recent years, the more powerful chromosome conformation capture (3C) method has yielded a series of high-resolution 3D conformation maps of the human genome in several cell types. In the 3C method, genomic DNA fragments are ligated to other genomic DNA fragments in physical proximity in the nucleus (92). These results have led to the identification of large compartments related to genomic organization, including enhancer-promoter loops (93), topologically associating domains (TADs) (94), and A/B compartments (92). In addition, 3C methods have been integrated with biochemical assays to annotate potentially functional interactions. For example, paired-end tag sequencing (ChIA-PET) (95), HiChIP (96), and proximity ligation-assisted ChIP-seq (PLAC-seq) (97) provide an overview of genome structures with a focus on proteins. Despite the development of various epigenomic methods as described above, and the obvious importance of human enhancers in both basic and disease biology, we still do not understand the repertoire of enhancers, including where they reside, how they act, and through which genes they mediate their effects.

In addition, it has recently been reported that super-enhancers are involved in abnormal gene expression in cancer cells (98). A super-enhancer is a region of the mammalian genome consisting of multiple enhancers, which are joined by a sequence of transcription factor proteins to drive the transcription of genes involved in cell identity (**Figure 2**) (99). An interesting finding is that disease-associated genetic mutations are particularly prevalent in super-enhancers of disease-associated cell types (100). Furthermore, cancer cells have been found to produce super-enhancers for oncogenes and other genes important in cancer development, suggesting that super-enhancers play an important role in human cell health and disease identity (100, 101). Importantly, super-enhancers are enriched in active chromatin marks such as H3K27ac and H3K4me3, while they are depleted in posed marks such as H3K27me3 (102). Therefore, epigenetic dysregulation may be involved in the production of super-enhancers in cancer cells. Since many disease-specific genetic variants are observed in super-enhancers, it seems to be pretty important to combine the information on genetic variants in non-coding regions obtained by WGS with the information on super enhancers based on epigenome data and analyze them in an integrated manner. As an example of super-enhancer analysis using ML, Gong et al. used two-dimensional lasso to improve the reproducibility of the Hi-C contact matrix and then classified the TAD boundaries based on the insulation score (103). The results showed that a higher TAD boundary insulation score was associated with higher CTCF levels, which may vary by cell type. They also showed that strong TAD boundaries and super-enhancer elements frequently overlap in cancer patients, suggesting that super-enhancer insulated by strong TAD boundaries may be used by cancer cells as a functional unit to promote tumorigenesis (103). Furthermore, Bu et al. proposed a new computational method, DEEPSEN, for super-enhancer

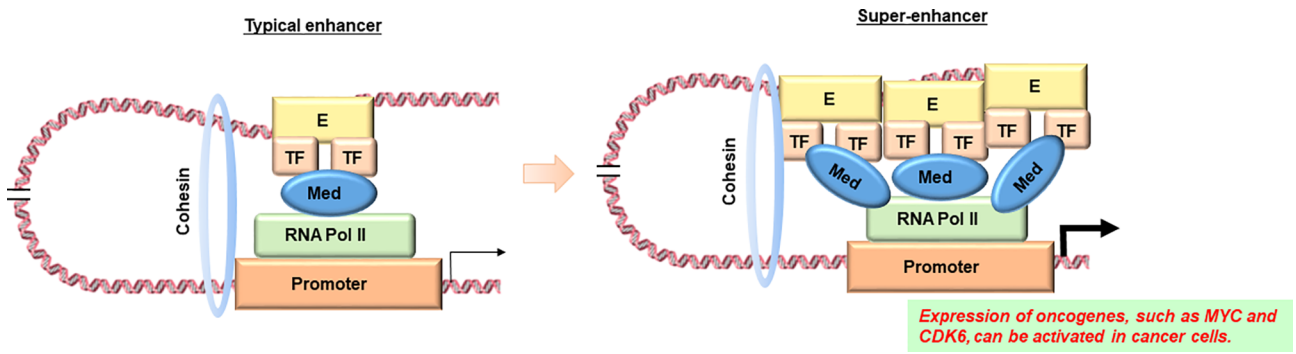


FIGURE 2 | Diagram of comparison between a typical enhancer and a super-enhancer. According to reference 87, super enhancers are observed in the transcriptional regulatory regions of oncogenes such as *MYC* in cancer cells, but not in their counterparts in normal tissues. E, enhancer; TF, transcription factor; Med, Mediator complex; RNA pol II, RNA polymerase II. Image credit: Shutterstock.com/ellepigrafica.

prediction using a convolutional neural network, which is a DL algorithm (104). The proposed method integrates 36 different features and shows that it is capable of genome-wide prediction of super enhancers compared to existing methods.

In transcriptome and epigenome profiling, one of the conservative ML approaches of cluster analysis often yields reproducible regulatory subtypes. In this way, somatic mutations in cancer, although chaotic, often converge in a regulatory manner. These events suggest that cancer cells follow the same rules of transcriptional regulation as normal cells, despite the presence of aberrant combinations of transcription factors and genomic enhancers (105). Furthermore, a major unresolved question is how primary cancer cells metastasize and what the molecular events underlying this process are. However, extensive sequencing studies have shown that mutations may not be the causative factors in the transition from primary to metastasis (106). On the other hand, epigenetic changes are dynamic in nature and may play an important role in determining the metastatic phenotype, and research in this area is only beginning to be evaluated (107, 108). Unlike genetic studies, the current limitations in studying epigenetic events in cancer metastasis are the lack of conceptual understanding and the lack of an analytical framework to identify the putative driver and passenger epigenetic changes. We would therefore like to introduce an ML analysis that has the potential to address these issues.

CHALLENGES THAT MACHINE LEARNING CAN OVERCOME

Genomic and epigenetic data-driven science operates by comprehensively exploring genome-wide data to discover new properties, rather than testing existing models and hypotheses (109). These data-driven approaches include finding relationships between genotypes and phenotypes, searching for biomarkers for personalized medicine, discovering driver genes and predicting their functions, and tracking genomic regions with biochemical activities such as transcriptional enhancers, as mentioned in the previous section. Due to the large scale and complexity of genomic and epigenetic data, it is often not sufficient to check pairwise correlations to make predictions. Therefore, analytical tools are needed to support the discovery of new relationships, the derivation of new hypotheses and models, and to make predictions. ML is designed to automatically detect patterns in data, unlike algorithms that have predetermined assumptions and expertise. Therefore, ML is well suited for data-driven science, especially genomics and epigenomics (110). However, the performance of ML is highly dependent on how the data are represented and how each variable or a feature is extracted. Epigenetic information and various modalities are known to be interrelated events, which are thought to interact with each other to change gene activity patterns. Based on these hypotheses, Wang et al. predicted the DNA methylation state of a specific region using a deterministic ML model [stacked denoising autoencoders (SdAs)] based on the 3D genome topology and DNA sequence obtained from Hi-C experiments (111). Against the backdrop of the high cost and difficulty of experimental techniques, which is the bottleneck of Hi-C data acquisition, inference from 1D information such as ChIP-seq, ATAC-seq, and RNA-seq to 3D genome topology structure has been actively attempted using various ML methods (**Table 2**). However, the prediction accuracy may not be improved due to inaccurate extraction of the essential structures within the epigenetic dataset, such as the still unelucidated mechanism of gene transcription regulation by high-dimensional interactions between enhancer and promoter regions. To solve these issues, an integrated approach that combines not only the acquisition of multi-layered omics data over time but also the generation and selection of phenotypic features and ML, is necessary.

INTEGRATED ANALYSIS OF WHOLE GENOME SEQUENCING AND EPIGENOME DATASETS

For decades, cancer genome research has made significant progresses in the identification of driver gene mutations, largely owing to the wide application of WES. However, we are

TABLE 2 | Epigenetic analysis typically focusing on regulatory regions.

Features	Pipeline name	Brief summary	Reference
Epigenomic Atlas (chromatin marks/ chromatin states, DHSs, active enhancers)	N/A	Mapping nine chromatin marks across nine cell types. Systematically characterizes regulatory elements, cell-type specificities, and functional interactions. Defining multicell activity profiles for chromatin state, gene expression, regulatory motif enrichment, and regulator expression. Assigning candidate regulatory functions to disease-associated variants from GWAS.	Nature (2011) (112)
	N/A	Presenting extensive map of human DNase I hypersensitive site (DHSs) to identify through genome-wide profiling in 125 diverse cells and tissue types. The map shows relationships between chromatin accessibility, transcription, DNA methylation, and mutation rate in regulatory DNA.	Nature (2012) (113)
	N/A	The bidirectional capped RNAs measured by cap analysis of gene expression (CAGE) are robust predictors of enhancer activity. Enhancers share properties with CpG-poor messenger RNA promoters but produce bidirectional, exosome-sensitive, relatively short unspliced RNAs. The generation of RNA is strongly related to enhancer activity.	Nature (2014) (114)
Regulatory sequence/ Network identify (enhancer/ promoter/EPI, *etc.*)	ELMER (Enhancer Linking by Methylation/ Expression Relationships)	This uses methylation and expression data to identify cancer-specific regulatory transcription factors, detect enhancer-gene promoter pairs, and correlate enhancer status with expression of neighboring genes.	Genome Biol (2015) (115).
	JEME (joint effect of multiple enhancers)	This method is an inference of enhancer-target networks, and consists of two steps: identifying enhancers that regulate transcription start sites (TSSs) across all samples, and detecting enhancers that regulate TSSs in a particular sample, to determine the target genes of transcriptional enhancers in a particular cell or tissue.	Nat. Genet (2017) (116).
	FOCS (FDR-corrected OLS with Cross-validation and Shrinkage)	This method estimates the link between enhancers and promoters based on the correlation of activity patterns between samples and implements a leave-cell-type-out cross-validation (LCTO CV) procedure to avoid overfitting of the regression model to the training samples. The cross-validation scheme consists of learning training set of samples and evaluation left-out samples from other cell types. This also provides extensive enhancer–promoter maps from ENCODE, Roadmap Epigenomics, FANTOM5, and a new compendium of GRO-seq samples. FOCS suggests repressor–promoter links.	Genome Biol (2018) (117).
	SPEID (Sequence-based Promoter-Enhancer Interaction with Deep learning; pronounced "speed")	This method predicts enhancer-promoter interactions using DL models from genomic sequences, using only the location of enhancers and promoters in specific cell types. Using the melanoma dataset, this shows that there is potential to identify somatic non-coding mutations that reduce or interrupt important enhancer-promoter interactions (EPIs).	Quant. Biol (2019) (118).
	EP2vec	This method uses natural language processing to predict enhancer-promoter interactions, and also extracts sequence-embedded features (fixed-length vector representations) using an unsupervised DL model, the paragraph vector. The extracted features are used to train a classifier to predict the interaction using supervised learning. This can also merge sequence embedded features with experimental features for more accurate prediction.	BMC Genomics (2018) (119)
Inference of the 3D structure of chromatin	Transcriptional decomposition	This separates RNA expression into positionally dependent (PD) component and positionally independent (PI) effects by transcriptional decomposition method to show the predictability of fine-scale chromatin interactions, chromosomal positioning, and three-dimensional chromatin architecture.	Nat. Commun (2018) (120).
	CHINN (Chromatin Interaction Neural Network)	This predicts chromatin interactions between open chromatin regions using DNA sequence and distance using convolutional neural network. This also extracts sequence features and feed into classifiers.	bioRxiv (2019) (121)
	HiC-Reg	This method uses one-dimensional regulatory signals (chromatin marks, architecture, transcription factor proteins, and chromatin accessibility) and the published Hi-C dataset as training count data to predict cell line-specific contact counts. A random forest regression model is used as the main prediction algorithm.	Nat. Commun (2019) (122).

now realizing that druggable gene mutations are limited, and the majority of cancer patients are left with unmet medical needs. Therefore, academic interest has gradually shifted to the analysis of mutations in non-coding genomes based on WGS analysis and the search for "epi-drivers", which are mechanisms of cancer development and progression caused by epigenomic abnormalities. For this purpose, WGS and epigenetic sequence technologies such as ChIP-seq, ATAC-seq, and Hi-C are effective tools because they offer comprehensive information about the genome, epigenome, and crosstalk between these (**Figure 1**).

Integrated analysis of genome and epigenetic data can be applied to predict the functional significance of single nucleotide polymorphisms (SNPs) and germline/somatic mutations. In order to analyze the function of DNA mutations in non-coding genomes, it is important to focus on eQTLs, which are genomic sites involved in the variation of expression levels of target genes. It is known that most functionally active SNPs and mutations fall within the open chromatin region, especially at inferred transcription factor binding sites. Indeed, approximately 55% of eQTLs SNPs are reported to coincide with those of open chromatin-associated SNPs and mutations (123). An impressive study on integrated analyses of WGS, ATAC-seq, and RNA-seq datasets has been posted (124). In a case of bladder cancer, they found that a single base mutation in enhancer region of the

FGD4 gene generated a putative *de novo* binding site for an NKX transcription factor, associated with an increase in chromatin accessibility and *FGD4* gene expression (124). Since high expression of the *FGD4* gene correlates with worse clinical outcomes in bladder cancer patients, this non-coding mutation might contribute to the malignant transformation of the cells by altering chromatin structure, thereby upregulating *FGD4* gene expression.

However, it should be noted that the majority of non-coding mutations might not exert an active function. In general, the regional mutation rates of human cancer cells tend to be higher in repressive chromatin states than in active chromatin states, which may reflect differing efficiencies of DNA repair signals or mutagen exposure (125). Thus, from a probabilistic view, most of mutations in the heterochromatin region occur only because of their closed chromatin states; that is, they are less likely to have any selective advantages or active functions. Intriguingly, this tendency toward higher mutational occurrences in heterochromatin states offers potentially useful information. By applying the ML model, genome-wide mutation data can be utilized to infer the cell-of-origin of cancer cells. For example, the mutational landscape of melanoma is best correlated with the epigenetic profile of skin melanocytes than skin fibroblasts or skin keratinocytes, suggesting the true cell-of-origin of melanoma (126). This approach can be clinically applicable to predict the cell-of-origin for cancer of unknown primary origin and may yield a better phenotypic understanding of them. WGS can resolve non-coding SVs and CNVs. RNA-seq detects the expression levels of driver genes and aberrantly expressed genes caused by alternative promoter usage and exon skipping (127–130). The utility of an integrative, comprehensive approach, with WGS, RNA-seq, and DNA methylation, independently and in combination, has been reported (130). Comprehensive molecular tumor profiling comprising WGS, RNA-seq, and DNA methylation analyses identified pathogenic variants and provided therapy recommendations, which could accelerate the development of precision medications.

Overall, the genomic and epigenetic data of non-coding regions contain enormous, complex and interdependent information, and we believe that integrated analysis, effectively utilizing ML and DL technologies, is important to discover new drivers of human cancer.

DISCUSSION

The genetic variants or SNPs were refined by the international haplotype map (HapMap) project to create a haplotype map of genes and genetic variants that affect health and disease (131–133). This project was attempted to genotype one common SNP in every 5,000 bps. At that time, it was believed that more than 99.9% of DNA sequences between any two people were identical, suggesting that only less than 0.1% of the genetic variants affect health and disease (https://www.genome.gov/11511175/about-the-international-hapmap-project-fact-sheet). Nowadays, analyzing WGS data has identified a considerable number of the genomic variants. The international consortium embarked on the 1000 Genomes Project to find common human genetic variations by applying WGS to a diverse set of individuals from multiple populations. High-throughput sequencing technologies do facilitate WGS in terms of accuracy, cost, and time. Almost two decades after the completion of the Human Genome Project, we have already entered a new era of sequencing, which led to individual genomic information becoming analyzable data. In practical terms, WGS analysis is becoming cost-effective. In addition, there is a trend to apply WGS routinely in both basic sciences and clinical cancer care to help us better understand and identify potential therapeutic targets or predictive biomarkers.

Epigenetics analyses were also drastically and positively affected by NGS. Chromatin conformations analyzed by ChIP-seq, ATAC-seq, or Hi-C are known to be related to cancer phenotypes (124, 134). Epigenetic alterations of DNA methylation at promoter and enhancer regions that induce chromatin dysregulation are found in cancer (135, 136). NGS analysis can help resolve both genetic and epigenetic alterations, and we expect to reveal the mechanism of pan-negative cancers using these data. From this point of view, we further introduced enhancers as an important concept in precision oncology. The current understanding is that enhancers bind to cell type-specific transcription factors, associate with regions of open chromatin, and are flanked by histones with H3K27ac and/or H3K4me1 modifications. These enhancers interact with promoters in 3D space and are either potentially primed or activated. Despite their obvious importance in both basic biology and disease biology, much remains to be learned about the relationship between enhancers and chromatin higher-order structure, including the identification of enhancer regions, how enhancers work, and through which genes they mediate their effects. In the future, we hope that multimodal analysis of multidimensional omics data by effective use of ML and DL techniques may contribute to precision oncology by providing an integrated understanding of more detailed molecular mechanisms.

CONCLUDING REMARKS

In this review, we first summarized the importance of genomic and epigenetic data and introduced the importance of omics data of interest in each section. Cancer is one of the leading causes of death worldwide, and molecular mechanisms remain unknown in certain cancers, which are categorized as pan-negative cancers. Multi-omics analyses by simply integrating omics data may encounter difficulties in identifying the mechanism causing cancer because none of the methodologies can address the comprehensive understanding underlying pan-negative cancers. Therefore, as we reviewed here, integrating multi-omics analysis with the assistance of ML is required for future cancer studies because each omics data is tightly linked to each other, and all omics data are associated with patient outcomes. Currently, there are high expectations for the development of medical AI, and it is expected that AI technology will be actively introduced in actual clinical practice in the future. On the other hand, medical AI research for clinical applications is currently focused on medical image analysis (137–144), and research on the introduction of AI

to omics analysis such as whole genome analysis and epigenome analysis, as well as its clinical application, has not progressed sufficiently yet. In this regard, one of the problems associated with the widespread adoption of AI-based methodologies in omics analysis is that even though sequencing technology and other advanced analytics are increasingly being used in research and clinical practice, there is still a lot of confusion about the best protocols to adopt for analysis. For example, the RNA-seq pipeline is not sufficiently standardized, and the methodology relies heavily on the expertise and experience of a single research group/bioinformatics. As a result, in areas where uncertainty remains, the spread of AI-specific technologies may be delayed. We hope that this review will trigger the interest of more researchers in this field, and that the standardization of omics analysis will actively promote the adoption of AI and contribute to the establishment of the field of precision oncology in the future.

AUTHOR CONTRIBUTIONS

KA and RH contributed to the study concept, design, and are guarantor of integrity of the entire study. KA, SK, KT, HM, ST, NS, RS, MK, and RH all contributed to literature search, manuscript preparation, and manuscript editing. All authors contributed to the article and submitted version.

FUNDING

This work was supported by JST CREST (Grant Number JPMJCR1689), JST AIP-PRISM (Grant Number JPMJCR18Y4), JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number JP18H04908), and JSPS KAKENHI (Grant Number JP20K17982).

REFERENCES

1. Watson JD, Crick FH. Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid. *Nature* (1953) 171(4356):737–8. doi: 10.1038/171737a0
2. Sanger F, Nicklen S, Coulson AR. DNA Sequencing With Chain-Terminating Inhibitors. *Proc Natl Acad Sci USA* (1977) 74(12):5463–7. doi: 10.1073/pnas.74.12.5463
3. Watson JD, Cook-Deegan RM. Origins of the Human Genome Project. *FASEB J* (1991) 5(1):8–11. doi: 10.1096/fasebj.5.1.1991595
4. Collins FS, Morgan M, Patrinos A. the Human Genome Project: Lessons From Large-Scale Biology. *Science* (2003) 300(5617):286–90. doi: 10.1126/science.1084564
5. Katsnelson A. Momentum Grows to Make 'Personalized' Medicine More 'Precise'. *Nat Med* (2013) 19(3):249. doi: 10.1038/nm0313-249
6. Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AM, et al. Cancer Genomics: Technology, Discovery, and Translation. *J Clin Oncol* (2012) 30(6):647–60. doi: 10.1200/JCO.2011.39.2316
7. Roychowdhury S, Chinnaiyan AM. Translating Genomics for Precision Cancer Medicine. *Annu Rev Genomics Hum Genet* (2014) 15:395–415. doi: 10.1146/annurev-genom-090413-025552
8. Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet* (2016) 15:95–115. doi: 10.1146/annurev-genom-083115-022413
9. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol* (2018) 122(1):e59. doi: 10.1002/cpmb.59
10. Hamamoto R, Komatsu M, Takasawa K, Asada K, Kaneko S. Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial Intelligence in the Era of Precision Medicine. *Biomolecules* (2020) 10(1):62. doi: 10.3390/biom10010062
11. Wang Z, Gerstein M, Snyder M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nat Rev Genet* (2009) 10(1):57–63. doi: 10.1038/nrg2484
12. Cimmino F, Lasorsa VA, Vetrella S, Iolascon A, Capasso MA. Targeted Gene Panel for Circulating Tumor DNA Sequencing in Neuroblastoma. *Front Oncol* (2020) 10:596191. doi: 10.3389/fonc.2020.596191
13. Fernandes MGO, Jacob M, Martins N, Moura CS, Guimaraes S, Reis JP, et al. Targeted Gene Next-Generation Sequencing Panel in Patients With Advanced Lung Adenocarcinoma: Paving the Way for Clinical Implementation. *Cancers (Basel)* (2019) 11(9):1229. doi: 10.3390/cancers11091229
14. Surrey LF, MacFarland SP, Chang F, Cao K, Rathi KS, Akgumus GT, et al. Clinical Utility of Custom-Designed NGS Panel Testing in Pediatric Tumors. *Genome Med* (2019) 11(1):32. doi: 10.1186/s13073-019-0644-8
15. Zhang X, Yang H, Zhang R. Challenges and Future of Precision Medicine Strategies for Breast Cancer Based on a Database on Drug Reactions. *Biosci Rep* (2019) 39(9):BSR20190230. doi: 10.1042/BSR20190230
16. Prasad V. Perspective: The Precision-Oncology Illusion. *Nature* (2016) 537(7619):S63. doi: 10.1038/537S63a
17. Meric-Bernstam F, Brusco L, Shaw K, Horombe C, Kopetz S, Davies MA, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. *J Clin Oncol* (2015) 33(25):2753–62. doi: 10.1200/JCO.2014.60.4165
18. Oki S, Sone K, Oda K, Hamamoto R, Ikemura M, Maeda D, et al. Oncogenic Histone Methyltransferase EZH2: A Novel Prognostic Marker With Therapeutic Potential in Endometrial Cancer. *Oncotarget* (2017) 8(25):40402–11. doi: 10.18632/oncotarget.16316
19. Kogure M, Takawa M, Saloura V, Sone K, Piao L, Ueda K, et al. The Oncogenic Polycomb Histone Methyltransferase EZH2 Methylates Lysine 120 on Histone H2B and Competes Ubiquitination. *Neoplasia* (2013) 15(11):1251–61.
20. Asada K, Bolatkan A, Takasawa K, Komatsu M, Kaneko S, Hamamoto R. Critical Roles of N(6)-Methyladenosine (M(6)a) in Cancer and Virus Infection. *Biomolecules* (2020) 10(7):1071. doi: 10.3390/biom10071071
21. Hayami S, Kelly JD, Cho HS, Yoshimatsu M, Unoki M, Tsunoda T, et al. Overexpression of LSD1 Contributes to Human Carcinogenesis Through Chromatin Regulation in Various Cancers. *Int J Cancer* (2011) 128(3):574–86. doi: 10.1002/ijc.25349
22. Kim S, Bolatkan A, Kaneko S, Ikawa N, Asada K, Komatsu M, et al. Deregulation of the Histone Lysine-Specific Demethylase 1 is Involved in Human Hepatocellular Carcinoma. *Biomolecules* (2019) 9(12):810. doi: 10.3390/biom9120810
23. Sone K, Piao L, Nakakido M, Ueda K, Jenuwein T, Nakamura Y, et al. Critical Role of Lysine 134 Methylation on Histone H2AX for Gamma-H2AX Production and DNA Repair. *Nat Commun* (2014) 5:5691. doi: 10.1038/ncomms6691
24. Saloura V, Cho HS, Kyiotani K, Alachkar H, Zuo Z, Nakakido M, et al. Whsc1 Promotes Oncogenesis Through Regulation of Nima-Related-Kinase-7 in Squamous Cell Carcinoma of the Head and Neck. *Mol Cancer Res* (2015) 13(2):293–304. doi: 10.1158/1541-7786.MCR-14-0292-T
25. Wada M, Kukita A, Sone K, Hamamoto R, Kaneko S, Komatsu M, et al. Epigenetic Modifier SETD8 as a Therapeutic Target for High-Grade Serous Ovarian Cancer. *Biomolecules* (2020) 10(12):1686. doi: 10.3390/biom10121686
26. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA,Jr., Kinzler KW. Cancer Genome Landscapes. *Science* (2013) 339(6127):1546–58. doi: 10.1126/science.1235122
27. Chatterjee A, Rodger EJ, Eccles MR. Epigenetic Drivers of Tumourigenesis and Cancer Metastasis. *Semin Cancer Biol* (2018) 51:149–59. doi: 10.1016/j.semcancer.2017.08.004
28. Eliseev A, Gibson KM, Avdeyev P, Novik D, Bendall ML, Perez-Losada M, et al. Evaluation of Haplotype Callers for Next-Generation Sequencing of Viruses. *Infect Genet Evol* (2020) 82:104277. doi: 10.1016/j.meegid.2020.104277

29. Asada K, Kobayashi K, Joutard S, Tubaki M, Takahashi S, Takasawa K, et al. Uncovering Prognosis-Related Genes and Pathways by Multi-Omics Analysis in Lung Cancer. *Biomolecules* (2020) 10(4):524. doi: 10.3390/biom10040524
30. Kobayashi K, Bolatkan A, Shiina S, Hamamoto R. Fully-Connected Neural Networks With Reduced Parameterization for Predicting Histological Types of Lung Cancer From Somatic Mutations. *Biomolecules* (2020) 10(9):1249. doi: 10.3390/biom10091249
31. Takahashi S, Asada K, Takasawa K, Shimoyama R, Sakai A, Bolatkan A, et al. Predicting Deep Learning Based Multi-Omics Parallel Integration Survival Subtypes in Lung Cancer Using Reverse Phase Protein Array Data. *Biomolecules* (2020) 10(10):1460.
32. Srivastava N, Salakhutdinov R. Multimodal Learning With Deep Boltzmann Machines. *J Mach Learn Res* (2014) 15:2949–80.
33. Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, et al. Integrating Clinical and Multiple Omics Data for Prognostic Assessment Across Human Cancers. *Sci Rep* (2017) 7(1):16954. doi: 10.1038/s41598-017-17031-8
34. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* (2018) 24(6):1248–59. doi: 10.1158/1078-0432.CCR-17-0853
35. Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A Machine Learning Approach to Integrate Big Data for Precision Medicine in Acute Myeloid Leukemia. *Nat Commun* (2018) 9(1):42. doi: 10.1038/s41467-017-02465-5
36. Gonen M, Margolin AA. Drug Susceptibility Prediction Against a Panel of Drugs Using Kernelized Bayesian Multitask Learning. *Bioinformatics* (2014) 30(17):i556–63. doi: 10.1093/bioinformatics/btu464
37. Yuan H, Paskov I, Paskov H, Gonzalez AJ, Leslie CS. Multitask Learning Improves Prediction of Cancer Drug Sensitivity. *Sci Rep* (2016) 6:31619. doi: 10.1038/srep31619
38. Xiao Y, Wu J, Lin Z, Zhao X. a Semi-Supervised Deep Learning Method Based on Stacked Sparse Auto-Encoder for Cancer Prediction Using RNA-Seq Data. *Comput Methods Programs BioMed* (2018) 166:99–105. doi: 10.1016/j.cmpb.2018.10.004
39. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. *IEEE Trans Pattern Anal Mach Intell* (2013) 35(8):1798–828. doi: 10.1109/TPAMI.2013.50
40. Shi M, Zhang B. Semi-Supervised Learning Improves Gene Expression-Based Prediction of Cancer Recurrence. *Bioinformatics* (2011) 27(21):3017–23. doi: 10.1093/bioinformatics/btr502
41. Chapelle O, Sindhwani V, Keerthi SS. Optimization Techniques for Semi-Supervised Support Vector Machines. *J Mach Learn Res* (2008) 9:203–33.
42. Bengio Y. Learning Deep Architectures for AI. *Foundations Trends® Mach Learn* (2009) 2(1):p1–p127.
43. Hamamoto R, Suvarna K, Yamada M, Kobayashi K, Shinkai N, Miyake M, et al. Application of Artificial Intelligence Technology in Oncology: Towards the Establishment of Precision Medicine. *Cancers (Basel)* (2020) 12 (12):3532. doi: 10.3390/cancers12123532
44. Consortium, I.T.P.-C.A.o.W.G. Pan-Cancer Analysis of Whole Genomes. *Nature* (2020) 578(7793):82–93. doi: 10.1038/s41586-020-1969-6
45. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-Wide Profiling of Heritable and De Novo STR Variations. *Nat Methods* (2017) 14(6):590–2. doi: 10.1038/nmeth.4267
46. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and De Novo Assembly of 150 Genomes From Denmark as a Population Reference. *Nature* (2017) 548(7665):87–91. doi: 10.1038/nature23264
47. Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet* (2016) 99(3):595–606. doi: 10.1016/j.ajhg.2016.07.005
48. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks. *Nat Biotechnol* (2018) 36(10):983–7. doi: 10.1038/nbt.4235
49. Ruzzo EK, Perez-Cano L, Jung JY, Wang LK, Kashef-Haghighi D, Hartl C, et al. Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* (2019) 178(4):850–66 e26. doi: 10.1016/j.cell.2019.07.015
50. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, et al. Whole-Genome Deep-Learning Analysis Identifies Contribution of Noncoding Mutations to Autism Risk. *Nat Genet* (2019) 51(6):973–80. doi: 10.1038/s41588-019-0420-0
51. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep Convolutional Neural Networks for Accurate Somatic Mutation Detection. *Nat Commun* (2019) 10(1):1041. doi: 10.1038/s41467-019-09027-x
52. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, et al. Graphtyper Enables Population-Scale Genotyping Using Pangenome Graphs. *Nat Genet* (2017) 49(11):1654–60. doi: 10.1038/ng.3964
53. Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, et al. Genome Graphs. *bioRxiv* (2017). doi: 10.1101/101378
54. Ambler JM, Mulaudzi S, Mulder N. Gengraph: A Python Module for the Simple Generation and Manipulation of Genome Graphs. *BMC Bioinf* (2019) 20(1):519. doi: 10.1186/s12859-019-3115-8
55. Hadi K, Yao X, Behr JM, Deshpande A, Xanthopoulakis C, Tian H, et al. Distinct Classes of Complex Structural Variation Uncovered Across Thousands of Cancer Genome Graphs. *Cell* (2020) 183(1):197–210.e32. doi: 10.1016/j.cell.2020.08.006
56. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. Pyclone: Statistical Inference of Clonal Population Structure in Cancer. *Nat Methods* (2014) 11 (4):396–8. doi: 10.1038/nmeth.2883
57. Caravagna G, Heide T, Williams MJ, Zapata L, Nichol D, Chkhaidze K, et al. Subclonal Reconstruction of Tumors by Using Machine Learning and Population Genetics. *Nat Genet* (2020) 52(9):898–907. doi: 10.1038/s41588-020-0675-5
58. KaramiNejadRanjbar M, Sharifzadeh S, Wietek NC, Artibani M, El-Sahhar S, Sauka-Spengler T, et al. A Highly Accurate Platform for Clone-Specific Mutation Discovery Enables the Study of Active Mutational Processes. *Elife* (2020) 9:e55207. doi: 10.7554/eLife.55207
59. Gulhan DC, Lee JJ, Melloni GEM, Cortes-Ciriano I, Park PJ. Detecting the Mutational Signature of Homologous Recombination Deficiency in Clinical Samples. *Nat Genet* (2019) 51(5):912–9. doi: 10.1038/s41588-019-0390-2
60. Pei G, Hu R, Dai Y, Zhao Z, Jia P. Decoding Whole-Genome Mutational Signatures in 37 Human Pan-Cancers by Denoising Sparse Autoencoder Neural Network. *Oncogene* (2020) 39(27):5031–41. doi: 10.1038/s41388-020-1343-z
61. Li S, Crawford FW, Gerstein MB. Using Siglasso to Optimize Cancer Mutation Signatures Jointly With Sampling Likelihood. *Nat Commun* (2020) 11(1):3575. doi: 10.1038/s41467-020-17388-x
62. Mieth B, Kloft M, Rodriguez JA, Sonnenburg S, Vobruba R, Morcillo-Suarez C, et al. Combining Multiple Hypothesis Testing With Machine Learning Increases the Statistical Power of Genome-Wide Association Studies. *Sci Rep* (2016) 6:36671. doi: 10.1038/srep36671
63. Arloth J, Eraslan G, Andlauer TFM, Martins J, Iurato S, Kuhnel B, et al. Deepwas: Multivariate Genotype-Phenotype Associations by Directly Integrating Regulatory Information Using Deep Learning. *PloS Comput Biol* (2020) 16(2):e1007616. doi: 10.1371/journal.pcbi.1007616
64. Yin B, Balvert M, van der Spek RAA, Dutilh BE, Bohte S, Veldink J, et al. Using the Structure of Genome Data in the Design of Deep Neural Networks for Predicting Amyotrophic Lateral Sclerosis From Genotype. *Bioinformatics* (2019) 35(14):i538–i47. doi: 10.1093/bioinformatics/btz369
65. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-Specific Error Profile of Illumina Sequencers. *Nucleic Acids Res* (2011) 39(13):e90. doi: 10.1093/nar/gkr344
66. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, et al. An Empirical Bayesian Framework for Somatic Mutation Detection From Cancer Genome Sequencing Data. *Nucleic Acids Res* (2013) 41(7):e89. doi: 10.1093/nar/gkt126
67. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Subtypes PT, et al. A Deep Learning System Accurately Classifies Primary and Metastatic Cancers Using Passenger Mutation Patterns. *Nat Commun* (2020) 11(1):728. doi: 10.1038/s41467-019-13825-8
68. Paten B, Novak AM, Eizenga JM, Garrison E. Genome Graphs and the Evolution of Genome Inference. *Genome Res* (2017) 27(5):665–76. doi: 10.1101/gr.214155.116
69. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* (2012) 149(5):979–93. doi: 10.1016/j.cell.2012.04.024
70. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of Mutational Processes in Human Cancer. *Nature* (2013) 500(7463):415–21. doi: 10.1038/nature12477

71. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep* (2013) 3(1):246–59. doi: 10.1016/j.celrep.2012.12.008
72. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. a Practical Guide for Mutational Signature Analysis in Hematological Malignancies. *Nat Commun* (2019) 10(1):2969. doi: 10.1038/s41467-019-11037-8
73. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional Snps in the Lymphotoxin-Alpha Gene That are Associated With Susceptibility to Myocardial Infarction. *Nat Genet* (2002) 32(4):650–4. doi: 10.1038/ng1047
74. Baylin SB. DNA Methylation and Gene Silencing in Cancer. *Nat Clin Pract Oncol* (2005) 2 Suppl:1(S4–11. doi: 10.1038/ncponc0354
75. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, et al. 5' Cpg Island Methylation is Associated With Transcriptional Silencing of the Tumour Suppressor P16/CDKN2/MTS1 in Human Cancers. *Nat Med* (1995) 1(7):686–92. doi: 10.1038/nm0795-686
76. Stewart GD, Van Neste L, Delvenne P, Delree P, Delga A, McNeill SA, et al. Clinical Utility of an Epigenetic Assay to Detect Occult Prostate Cancer in Histopathologically Negative Biopsies: Results of the Matloc Study. *J Urol* (2013) 189(3):1110–6. doi: 10.1016/j.juro.2012.08.219
77. Gilbert MR, Dignam JJ, Armstrong TS, Wefel JS, Blumenthal DT, Vogelbaum MA, et al. a Randomized Trial of Bevacizumab for Newly Diagnosed Glioblastoma. *N Engl J Med* (2014) 370(8):699–708. doi: 10.1056/NEJMoa1308573
78. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *N Engl J Med* (2014) 370(14):1287–97. doi: 10.1056/NEJMoa1311194
79. Lamb YN, Dhillon S. Epi Procolon((R)) 2.0 CE: A Blood-Based Screening Test for Colorectal Cancer. *Mol Diagn Ther* (2017) 21(2):225–32. doi: 10.1007/s40291-017-0259-y
80. Nuzzo PV, Berchuck JE, Korthauer K, Spisak S, Nassar AH, Abou Alaiwi S, et al. Detection of Renal Cell Carcinoma Using Plasma and Urine Cell-Free Dna Methylomes. *Nat Med* (2020) 26(7):1041–3. doi: 10.1038/s41591-020-0933-1
81. Nassiri F, Chakravarthy A, Feng S, Shen SY, Nejad R, Zuccato JA, et al. Detection and Discrimination of Intracranial Tumors Using Plasma Cell-Free Dna Methylomes. *Nat Med* (2020) 26(7):1044–7. doi: 10.1038/s41591-020-0932-2
82. Jurmeister P, Bockmayr M, Seegerer P, Bockmayr T, Treue D, Montavon G, et al. Machine Learning Analysis of Dna Methylation Profiles Distinguishes Primary Lung Squamous Cell Carcinomas From Head and Neck Metastases. *Sci Transl Med* (2019) 11(509):eaaw8513. doi: 10.1126/scitranslmed.aaw8513
83. Macias-Garcia L, Martinez-Ballesteros M, Luna-Romera JM, Garcia-Heredia JM, Garcia-Gutierrez J, Riquelme-Santos JC. Autoencoded DNA Methylation Data to Predict Breast Cancer Recurrence: Machine Learning Models and Gene-Weight Significance. *Artif Intell Med* (2020) 110:101976. doi: 10.1016/j.artmed.2020.101976
84. Volik S, Alcaide M, Morin RD, Collins C. Cell-Free DNA (Cfdna): Clinical Significance and Utility in Cancer Shaped by Emerging Technologies. *Mol Cancer Res* (2016) 14(10):898–908. doi: 10.1158/1541-7786.MCR-16-0044
85. Cancer Genome Atlas Research. N. Comprehensive Molecular Profiling of Lung Adenocarcinoma. *Nature* (2014) 511(7511):543–50. doi: 10.1038/nature13385
86. Saito M, Shiraishi K, Kunitoh H, Takenoshita S, Yokota J, Kohno T. Gene Aberrations for Precision Medicine Against Lung Adenocarcinoma. *Cancer Sci* (2016) 107(6):713–20. doi: 10.1111/cas.12941
87. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, et al. Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. *Science* (2012) 336(6082):736–9. doi: 10.1126/science.1217277
88. Cohen AJ, Saiakhova A, Corradin O, Luppino JM, Lovrenert K, Bartels CF, et al. Hotspots of Aberrant Enhancer Activity Punctuate the Colorectal Cancer Epigenome. *Nat Commun* (2017) 8:14400. doi: 10.1038/ncomms14400
89. Skene PJ, Henikoff S. An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites. *Elife* (2017) 6:e21856. doi: 10.7554/eLife.21856
90. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* (2008) 132(2):311–22. doi: 10.1016/j.cell.2007.12.014
91. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nat Methods* (2013) 10(12):1213–8. doi: 10.1038/nmeth.2688
92. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* (2009) 326 (5950):289–93. doi: 10.1126/science.1181369
93. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome At Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* (2014) 159(7):1665–80. doi: 10.1016/j.cell.2014.11.021
94. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* (2012) 485(7398):376–80. doi: 10.1038/nature11082
95. Fullwood MJ, Ruan Y. Chip-Based Methods for the Identification of Long-Range Chromatin Interactions. *J Cell Biochem* (2009) 107(1):30–9. doi: 10.1002/jcb.22116
96. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. Hichip: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture. *Nat Methods* (2016) 13(11):919–22. doi: 10.1038/nmeth.3999
97. Fang R, Yu M, Li G, Chee S, Liu T, Schmitt AD, et al. Mapping of Long-Range Chromatin Interactions by Proximity Ligation-Assisted Chip-Seq. *Cell Res* (2016) 26(12):1345–8. doi: 10.1038/cr.2016.137
98. Zhang J, Liu W, Zou C, Zhao Z, Lai Y, Shi Z, et al. Targeting Super-Enhancer-Associated Oncogenes in Osteosarcoma With Thz2, a Covalent Cdk7 Inhibitor. *Clin Cancer Res* (2020) 26(11):2681–92. doi: 10.1158/1078-0432.CCR-19-1418
99. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master Transcription Factors and Mediator Establish Super-Enhancers At Key Cell Identity Genes. *Cell* (2013) 153(2):307–19. doi: 10.1016/j.cell.2013.03.035
100. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell* (2013) 155 (4):934–47. doi: 10.1016/j.cell.2013.09.053
101. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* (2013) 153(2):320–34. doi: 10.1016/j.cell.2013.03.036
102. Khan A, Mathelier A, Zhang X. Super-Enhancers are Transcriptionally More Active and Cell Type-Specific Than Stretch Enhancers. *Epigenetics* (2018) 13 (9):910–22. doi: 10.1080/15592294.2018.1514231
103. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, et al. Stratification of TAD Boundaries Reveals Preferential Insulation of Super-Enhancers by Strong Boundaries. *Nat Commun* (2018) 9(1):542. doi: 10.1038/s41467-018-03017-1
104. Bu H, Hao J, Gan Y, Zhou S, Guan J. DEEPSEN: A Convolutional Neural Network Based Method for Super-Enhancer Prediction. *BMC Bioinf* (2019) 20(Suppl 15):598. doi: 10.1186/s12859-019-3180-z
105. Atkins M, Potier D, Romanelli L, Jacobs J, Mach J, Hamaratoglu F, et al. an Ectopic Network of Transcription Factors Regulated by Hippo Signaling Drives Growth and Invasion of a Malignant Tumor Model. *Curr Biol* (2016) 26(16):2101–13. doi: 10.1016/j.cub.2016.06.035
106. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell* (2011) 144(5):646–74. doi: 10.1016/j.cell.2011.02.013
107. Cheng PF, Shakhova O, Widmer DS, Eichhoff OM, Zingg D, Frommel SC, et al. Methylation-Dependent Sox9 Expression Mediates Invasion in Human Melanoma Cells and is a Negative Prognostic Factor in Advanced Melanoma. *Genome Biol* (2015) 16:42. doi: 10.1186/s13059-015-0594-4
108. Vizoso M, Ferreira HJ, Lopez-Serra P, Carmona FJ, Martinez-Cardus A, Girotti MR, et al. Epigenetic Activation of a Cryptic Tbc1d16 Transcript Enhances Melanoma Progression by Targeting Egfr. *Nat Med* (2015) 21 (7):741–50. doi: 10.1038/nm.3863
109. Brown PO, Botstein D. Exploring the New World of the Genome With DNA Microarrays. *Nat Genet* (1999) 21(1 Suppl):33–7. doi: 10.1038/4462
110. Libbrecht MW, Noble WS. Machine Learning Applications in Genetics and Genomics. *Nat Rev Genet* (2015) 16(6):321–32. doi: 10.1038/nrg3920
111. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo YY, et al. Predicting DNA Methylation State of Cpg Dinucleotide Using Genome Topological Features and Deep Networks. *Sci Rep* (2016) 6:19598. doi: 10.1038/srep19598

112. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types. *Nature* (2011) 473(7345):43–9. doi: 10.1038/nature09906
113. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The Accessible Chromatin Landscape of the Human Genome. *Nature* (2012) 489(7414):75–82. doi: 10.1038/nature11232
114. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. an Atlas of Active Enhancers Across Human Cell Types and Tissues. *Nature* (2014) 507(7493):455–61. doi: 10.1038/nature12787
115. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring Regulatory Element Landscapes and Transcription Factor Networks From Cancer Methylomes. *Genome Biol* (2015) 16(1):105. doi: 10.1186/s13059-015-0668-3
116. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of Enhancer-Target Networks in 935 Samples of Human Primary Cells, Tissues and Cell Lines. *Nat Genet* (2017) 49(10):1428–36. doi: 10.1038/ng.3950
117. Hait TA, Amar D, Shamir R, Elkon R. FOCS: A Novel Method for Analyzing Enhancer and Gene Activity Patterns Infers an Extensive Enhancer-Promoter Map. *Genome Biol* (2018) 19(1):56. doi: 10.1186/s13059-018-1432-2
118. Singh S, Yang Y, Póczos B, Ma J. Predicting Enhancer-Promoter Interaction From Genomic Sequence With Deep Neural Networks. *Quantitative Biol* (2019) 7(2):122–37. doi: 10.1007/s40484-019-0154-0
119. Zeng W, Wu M, Jiang R. Prediction of Enhancer-Promoter Interactions Via Natural Language Processing. *BMC Genomics* (2018) 19(Suppl 2):84. doi: 10.1186/s12864-018-4459-6
120. Rennie S, Dalby M, van Duin L, Andersson R. Transcriptional Decomposition Reveals Active Chromatin Architectures and Cell Specific Regulatory Interactions. *Nat Commun* (2018) 9(1):487. doi: 10.1038/s41467-017-02798-1
121. Cao F, Zhang Y, Loh YP, Cai Y, Fullwood MJ. (2019). doi: 10.1101/720748
122. Zhang S, Chasman D, Knaack S, Roy S. In Silico Prediction of High-Resolution Hi-C Interaction Matrices. *Nat Commun* (2019) 10(1):5449. doi: 10.1038/s41467-019-13423-8
123. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, et al. Dnase I Sensitivity Qtls are a Major Determinant of Human Expression Variation. *Nature* (2012) 482(7385):390–4. doi: 10.1038/nature10808
124. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The Chromatin Accessibility Landscape of Primary Human Cancers. *Science* (2018) 362(6413):eaav1898. doi: 10.1126/science.aav1898
125. Schuster-Bockler B, Lehner B. Chromatin Organization is a Major Influence on Regional Mutation Rates in Human Cancer Cells. *Nature* (2012) 488 (7412):504–7. doi: 10.1038/nature11273
126. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer. *Nature* (2015) 518(7539):360–4. doi: 10.1038/nature14221
127. Maqbool MA, Pioger L, El Aabidine AZ, Karasu N, Molitor AM, Dao LTM, et al. Alternative Enhancer Usage and Targeted Polycomb Marking Hallmark Promoter Choice During T Cell Differentiation. *Cell Rep* (2020) 32(7):108048. doi: 10.1016/j.celrep.2020.108048
128. Demircioglu D, Cukuroglu E, Kindermans M, Nandi T, Calabrese C, Fonseca NA, et al. a Pan-Cancer Transcriptome Analysis Reveals Pervasive Regulation Through Alternative Promoters. *Cell* (2019) 178(6):1465–77 e17. doi: 10.1016/j.cell.2019.08.018
129. Reimer KA, Mimoso CA, Adelman K, Neugebauer KM. Co-Transcriptional Splicing Regulates 3' End Cleavage During Mammalian Erythropoiesis. *Mol Cell* (2021) 81(5):998–1012. doi: 10.1016/j.molcel.2020.12.018
130. Wong M, Mayoh C, Lau LMS, Khuong-Quang DA, Pinese M, Kumar A, et al. Whole Genome, Transcriptome and Methylome Profiling Enhances Actionable Target Discovery in High-Risk Pediatric Cancer. *Nat Med* (2020) 26(11):1742–53. doi: 10.1038/s41591-020-1072-4
131. International HapMap C. A Haplotype Map of the Human Genome. *Nature* (2005) 437(7063):1299–320. doi: 10.1038/nature04226
132. International HapMap, C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. a Second Generation Human Haplotype Map of Over 3.1 Million Snps. *Nature* (2007) 449(7164):851–61. doi: 10.1038/nature06258
133. International HapMap, C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* (2010) 467(7311):52–8. doi: 10.1038/nature09298
134. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, et al. Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods. *Science* (2016) 351(6280):1454–8. doi: 10.1126/science.aad9024
135. Ando M, Saito Y, Xu G, Bui NQ, Medetgul-Ernar K, Pu M, et al. Chromatin Dysregulation and Dna Methylation At Transcription Start Sites Associated With Transcriptional Repression in Cancers. *Nat Commun* (2019) 10 (1):2188. doi: 10.1038/s41467-019-09937-w
136. Bell RE, Golan T, Sheinboim D, Malcov H, Amar D, Salamon A, et al. Enhancer Methylation Dynamics Contribute to Cancer Plasticity and Patient Mortality. *Genome Res* (2016) 26(5):601–11. doi: 10.1101/gr.197194.115
137. Dozen A, Komatsu M, Sakai A, Komatsu R, Shozu K, Machino H, et al. Image Segmentation of the Ventricular Septum in Fetal Cardiac Ultrasound Videos Based on Deep Learning Using Time-Series Information. *Biomolecules* (2020) 10(11):1526. doi: 10.3390/biom10111526
138. Jinnai S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R. the Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. *Biomolecules* (2020) 10(8):1123. doi: 10.3390/biom10081123
139. Komatsu M, Sakai A, Komatsu R, Matsuoka R, Yasutomi S, Shozu K, et al. Detection of Cardiac Structural Abnormalities in Fetal Ultrasound Videos Using Deep Learning. *Appl Sci* (2021) 11(1):371.
140. Shozu K, Komatsu M, Sakai A, Komatsu R, Dozen A, Machino H, et al. Model-Agnostic Method for Thoracic Wall Segmentation in Fetal Ultrasound Videos. *Biomolecules* (2020) 10(12):1691. doi: 10.3390/biom10121691
141. Yamada M, Saito Y, Imaoka H, Saiko M, Yamada S, Kondo H, et al. Development of a Real-Time Endoscopic Image Diagnosis Support System Using Deep Learning Technology in Colonoscopy. *Sci Rep* (2019) 9 (1):14465. doi: 10.1038/s41598-019-50567-5
142. Yasutomi S, Arakaki T, Matsuoka R, Sakai A, Komatsu R, Shozu K, et al. Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows. *Appl Sci* (2021) 11(3):1127. doi: 10.3390/app11031127
143. Hamamoto R. Application of Artificial Intelligence for Medical Research. *Biomolecules* (2021) 11(1):90. doi: 10.3390/biom11010090
144. Takahashi S, Takahashi M, Kinoshita M, Miyake M, Kawaguchi R, Shinojima N, et al. Fine-Tuning Approach for Segmentation of Gliomas in Brain Magnetic Resonance Images With a Machine Learning Method to Normalize Image Differences Among Facilities. *Cancers (Basel)* (2021) 13 (6):1415. doi: 10.3390/cancers13061415

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: 03 December 2021
doi: 10.3389/fgene.2021.707802

An Integrated Analysis of Tumor Purity of Common Central Nervous System Tumors in Children Based on Machine Learning Methods

Jian Yang†, Jiajia Wang†, Shuaiwei Tian†, Qinhua Wang, Yang Zhao, Baocheng Wang, Liangliang Cao, Zhuangzhuang Liang, Heng Zhao, Hao Lian and Jie Ma**

Department of Pediatric Neurosurgery, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

OPEN ACCESS

Edited by:
Jaume Reventos, Institut d'Investigacio Biomedica de Bellvitge (IDIBELL), Spain

Reviewed by:
Maurizio Polano, Aviano Oncology Reference Center (IRCCS), Italy
Dimitar Vassilev, Sofia University, Bulgaria

***Correspondence:**
Hao Lian
sdwfys1@126.com
Jie Ma
majie3004@xinhuamed.com.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *10 May 2021*
Accepted: *10 November 2021*
Published: *03 December 2021*

Citation:
Yang J, Wang J, Tian S, Wang Q, Zhao Y, Wang B, Cao L, Liang Z, Zhao H, Lian H and Ma J (2021) An Integrated Analysis of Tumor Purity of Common Central Nervous System Tumors in Children Based on Machine Learning Methods. Front. Genet. 12:707802. doi: 10.3389/fgene.2021.707802

Background: Tumor purity is defined as the proportion of cancer cells in the tumor tissue, and its effects on molecular genetics, the immune microenvironment, and the prognosis of children's central nervous system (CNS) tumors are under-researched.

Methods: We applied random forest machine learning, the InfiniumPurify algorithm, and the ESTIMATE algorithm to estimate the tumor purity of every child's CNS tumor sample in several published pediatric CNS tumor sample datasets from Gene Expression Omnibus (GEO), aiming to perform an integrated analysis on the tumor purity of children's CNS tumors.

Results: Only the purity of CNS tumors in children based on the random forest (RF) machine learning method was normally distributed. In addition, the children's CNS tumor purity was associated with primary clinical pathological and molecular indicators. Enrichment analysis of biological pathways related to the purity of medulloblastoma (MB) revealed some classical signaling pathways associated with MB biology and development-related pathways. According to the correlation analysis between MB purity and the immune microenvironment, three immune-related genes, namely, CD8A, CXCR2, and TNFRSF14, were negatively related to MB purity. In contrast, no significant correlation was detected between immunotherapy-associated markers, such as PD-1, PD-L1, and CTLA4; most infiltrating immune cells; and MB purity. In the tumor purity–related survival analysis of MB, ependymoma (EPN), and children's high-grade glioma, we discovered a minor effect of tumor purity on the survival of the aforementioned pediatric patients with CNS tumors.

Conclusion: Our purity pediatric pan-CNS tumor analysis provides a deeper understanding and helps with the clinical management of pediatric CNS tumors.

Keywords: pediatric, central nervous system tumors, medulloblastoma, tumor purity, machine learning

Abbreviations: CNS, Central nervous system; DEGs, Differentially expressed genes; EPN, Ependymoma; GEO, Gene Expression Omnibus; GO, Gene ontology; GSEA, Gene set enrichment analysis; GSVA, Gene set variation analysis; K-M, Kaplan–Meier; MB, Medulloblastoma; RF, Random forest; ssGSEA, Single-sample gene set enrichment analysis; TME, Tumor microenvironment.

INTRODUCTION

As the most frequent solid tumors in children, pediatric tumors of the central nervous system (CNS) represent an array of molecularly and clinically diverse entities. The tumor microenvironment (TME) is a complicated milieu comprising many factors that promote and inhibit tumor growth, nutrients, chemokines, and the spectrum of non-tumor cells (e.g., immunocytes, fibroblasts, and endotheliocytes). Increasing evidence has revealed that the TME plays a pivotal role in tumorigenesis, tumor progression, and the response to therapy (Schreiber et al., 2011).

For the past few years, high-throughput techniques have been increasingly applied in the field of pediatric CNS tumors (Kumar et al., 2018). These techniques offer some new means for the clinical diagnosis, prognostic prediction, and precise classification of pediatric CNS tumors. Nevertheless, the surgically acquired tumor tissues used for high-throughput techniques are a mixture of both tumor cells and non-tumor tissues. The DNA and RNA extracted from such a mixture are from all of the cells involved, so the measurement result is a kind of mixed signal (Zheng et al., 2017). Such a sample mixture may bias the downstream analyses and thus could mask true biologically meaningful signals.

Tumor purity is defined as the proportion of tumor cells in tumor tissue. Some recent studies have reported the confounding effect of tumor purity on gene clustering, coexpression networks, molecular taxonomy, and tumor prognosis and microenvironment (Aran et al., 2015; Rhee et al., 2018). Currently, there are three main methods available for tumor purity estimation. The first is to estimate the tumor purity based on the pathological images of the tumor tissue by histopathological researchers and clinical pathologists. However, these results are subject to the observer's proficiency and the pathological sensitivity of the tumor tissue (Zhang et al., 2017). The second way determines tumor purity by virtue of cell sorting–based techniques such as magnetic-activated cell sorting (Schmitz et al., 1994) and fluorescent-activated cell sorting (Basu et al., 2010). However, these methods demand high inputs of time, effort, and money and are therefore difficult to apply in large-scale studies.

More recently, with the development of high-throughput techniques and improved bioinformatics approaches, many purity estimation methods by computational methods have been developed, and they are based on transcriptome data, copy number variation data, DNA methylation data, or genetic mutation data. These methods include the random forest (RF) algorithm based on DNA methylation data (Capper et al., 2018), ESTIMATE based on gene expression data (Yoshihara et al., 2013), ABSOLUTE based on somatic copy number data (Carter et al., 2012), and InfiniumPurify based on DNA methylation data (Zheng et al., 2017).

The existing studies on tumor purity are limited to adult samples from the Cancer Genome Atlas, and little is known regarding the relationship between tumor purity and the clinicopathologic or genomic features in pediatric CNS tumors. In addition, the association between the purity and microenvironment of pediatric CNS tumors remains unclear. In this study, we used these major means of tumor purity estimation to infer tumor purity and sought to evaluate the impact of purity on pediatric CNS tumor prognosis, genetic profiling, and the immune microenvironment, which may deepen our understanding of pediatric CNS tumor biology and provide new insights into the clinical management of pediatric CNS tumors.

MATERIALS AND METHODS

Data Collection

The data of children's CNS tumors (e.g., medulloblastoma (MB), ependymoma (EPN), pilocytic astrocytoma, diffuse midline glioma, atypical teratoma/rhomboid tumor, and embryonal tumor with multilayered rosettes) used in this study were from Gene Expression Omnibus (GEO) and ArrayExpress. **Supplementary Table S1** lists the general information about the datasets involved.

Selection of an Adequate Algorithm for Purity Estimation of Common Pediatric CNS Tumors

Random forest (RF), InfiniumPurify, and ESTIMATE algorithms were used to estimate tumor purity. The RF model was established by training the DNA methylation data extracted from the panglioma dataset (795 samples of glioma) (Ceccarelli et al., 2016) in TCGA based on the ABSOLUTE algorithm (a direct purity estimation method) (Capper et al., 2018). We selected the optimal algorithm from the aforementioned three algorithms according to the distribution of purity in different datasets of common pediatric CNS tumors.

Exploration of Biological Functions Related to Common Pediatric CNS Tumor Purity

We screened the genes that correlated with tumor purity by Pearson correlation analysis (Pearson $|R| > 0.3$). In total, 1,051 genes were eligible for Gene Ontology (GO) enrichment analysis and gene set enrichment analysis (GSEA) (Subramanian et al., 2005). Both GO analysis and GSEA were performed utilizing the R package "clusterProfiler." In addition, the cases were split into high- and low-purity groups based on the median purity. By utilizing the R package "GSVA," we performed gene set variation analysis (GSVA) of hallmark pathways between the high- and low-purity samples (Hänzelmann et al., 2013).

Evaluation of the Relationship Between the Purity of Common Pediatric CNS Tumors and the Tumor Microenvironment

By applying CIBERSORT(Gentles et al., 2015), we scored 22 immune cell types for their relative abundance in pediatric CNS tumor samples. For any given sample, we computed the relationships between tumor purity and the relative proportions of the individual immune cell types. In addition, we also computed the associations between tumor purity and the relative fractions of 24 immune cell types by using single-sample gene set enrichment analysis (ssGSEA) (Bindea et al., 2013), as

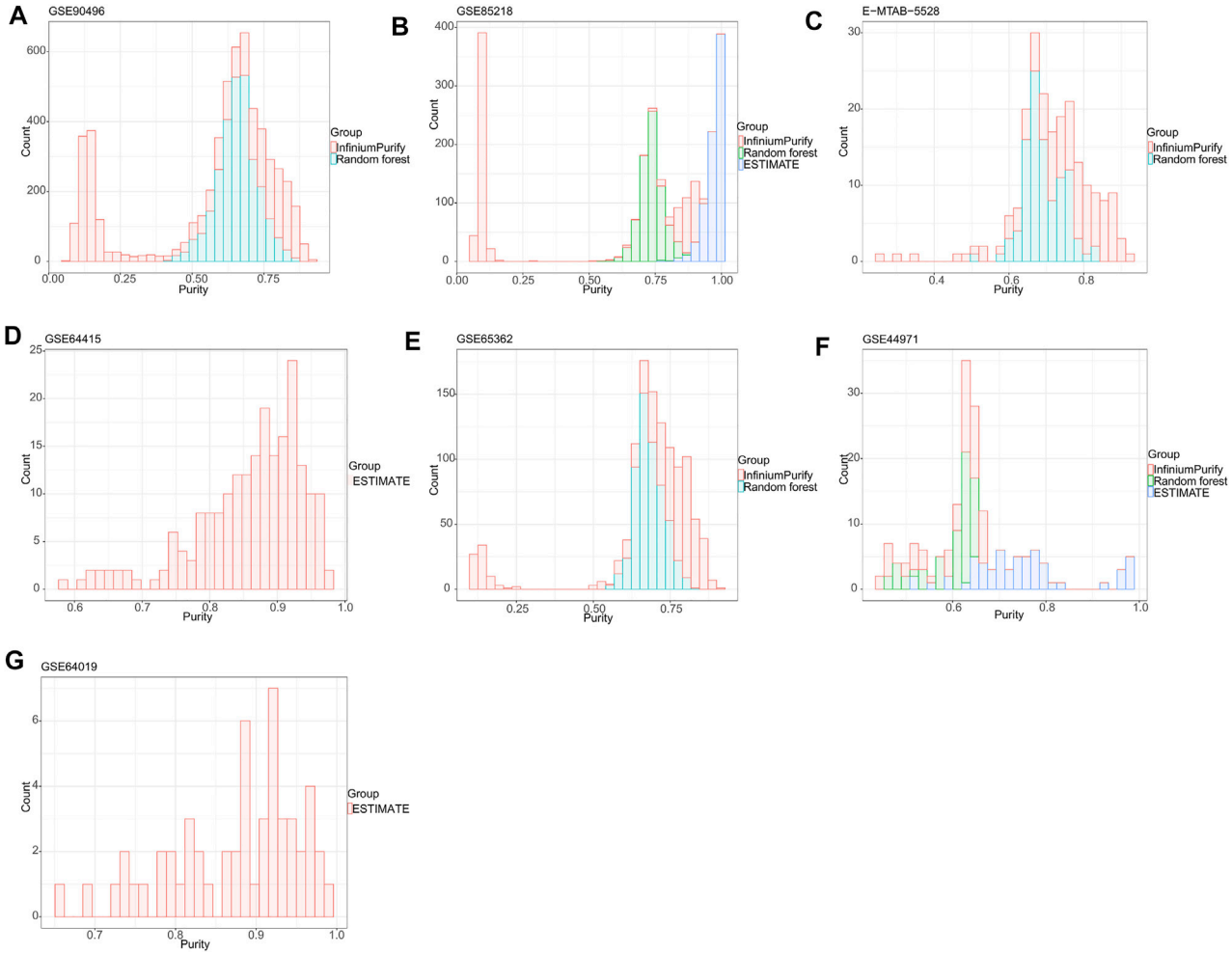


FIGURE 1 | Tumor purity distribution of seven common pediatric CNS tumor datasets based on three methods of tumor purity estimation. **(A)** Tumor purity distribution of the GSE90496 dataset based on InfiniumPurify and random forest (RF) algorithms; **(B)** tumor purity distribution of the GSE85218 dataset based on the InfiniumPurify, ESTIMATE, and RF algorithms; **(C)** tumor purity distribution of the E-MTAB-5528 dataset based on the InfiniumPurify and RF algorithms; **(D)** tumor purity distribution of the GSE64415 dataset based on the ESTIMATE algorithm; **(E)** tumor purity distribution of the GSE65362 dataset based on the InfiniumPurify and RF algorithms; **(F)** tumor purity distribution of the GSE44971 dataset based on the InfiniumPurify, ESTIMATE, and RF algorithms; **(G)** tumor purity distribution of the GSE64019 dataset based on the ESTIMATE algorithm.

implemented in the R package "GSVA." Finally, we determined the correlations between tumor purity and 14 immune-related genes (GZMA, PRF1, CD8A, PD-1, PD-L1, CTLA4, IDO1, CXCR2, TNFRSF14, TNFRSF18, CD247, LAG3, BTLA, and HAVCR2).

Survival Analysis

For each type of pediatric CNS tumors, we divided the samples into high- and low-purity groups based on the optimal cutoff value generated by using the R package "survMisc." Kaplan–Meier (K-M) curves were used to estimate the overall survival distribution.

Statistical Analysis

R software version 3.4.4 was employed for all statistical analyses. p values for the associations between tumor purity and the immune microenvironment were computed utilizing Pearson correlation analyses, followed by multiple testing utilizing the Benjamini–Hochberg method. For all statistical analyses, $p < 0.05$ was considered statistically significant.

RESULTS

Selection of the Most Adequate Algorithm for Estimating the Purity of Common Pediatric CNS Tumors

To establish a general understanding of the purity distribution of common pediatric CNS tumors, we estimated the tumor purity of samples in the GSE90496 datasets containing MB,

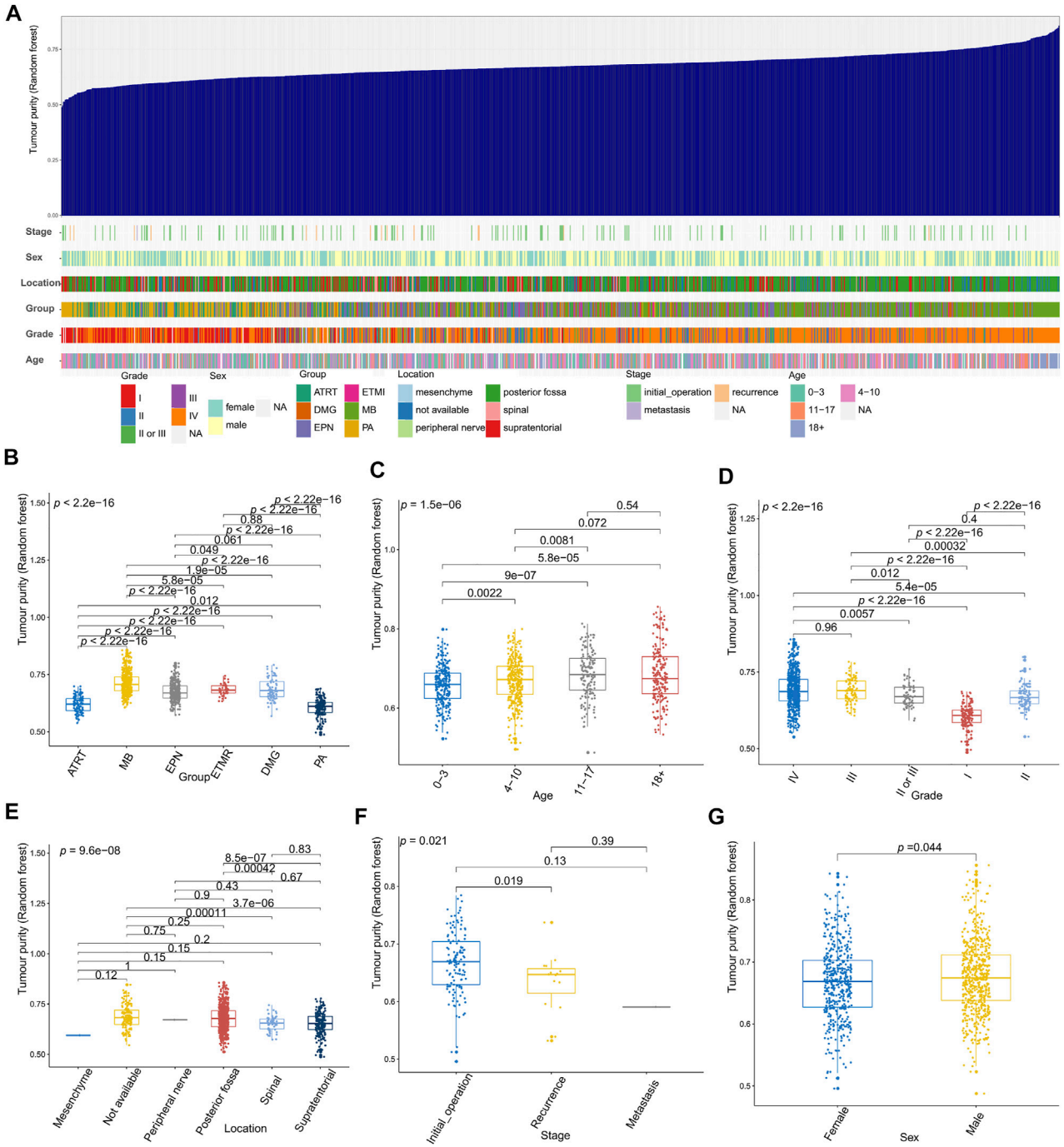


FIGURE 2 | Relationship between tumor purity and the patients' clinical features in the GSE90496 dataset. **(A)** An overview of the correlation between the clinical features and tumor purity in the GSE90496 dataset; **(B)** box plot of the tumor purity by tumor histology; **(C)** box plot of the tumor purity by age at diagnosis; **(D)** box plot of the tumor purity by the tumor grade; **(E)** Box plot of the tumor purity by the tumor location; **(F)** box plot of the tumor purity by the tumor grade; **(G)** box plot of the tumor purity by the patient gender.

EPN, pilocytic astrocytoma, diffuse midline glioma, atypical teratoma/rhomboid tumor, and embryonal tumor with multilayered rosettes. As shown in **Figure 1A**, the tumor purity distribution resulting from the InfiniumPurify algorithm had a bimodal pattern, with an average tumor purity of 49.8 ± 29.3%, while that from the RF algorithm was normal, with an average tumor purity of 65.9 ± 7.1%. Regarding the tumor purity distribution of the GSE85218 dataset (MB) (**Figure 1B**), the tumor purity based on the InfiniumPurify algorithm was bimodal (average tumor

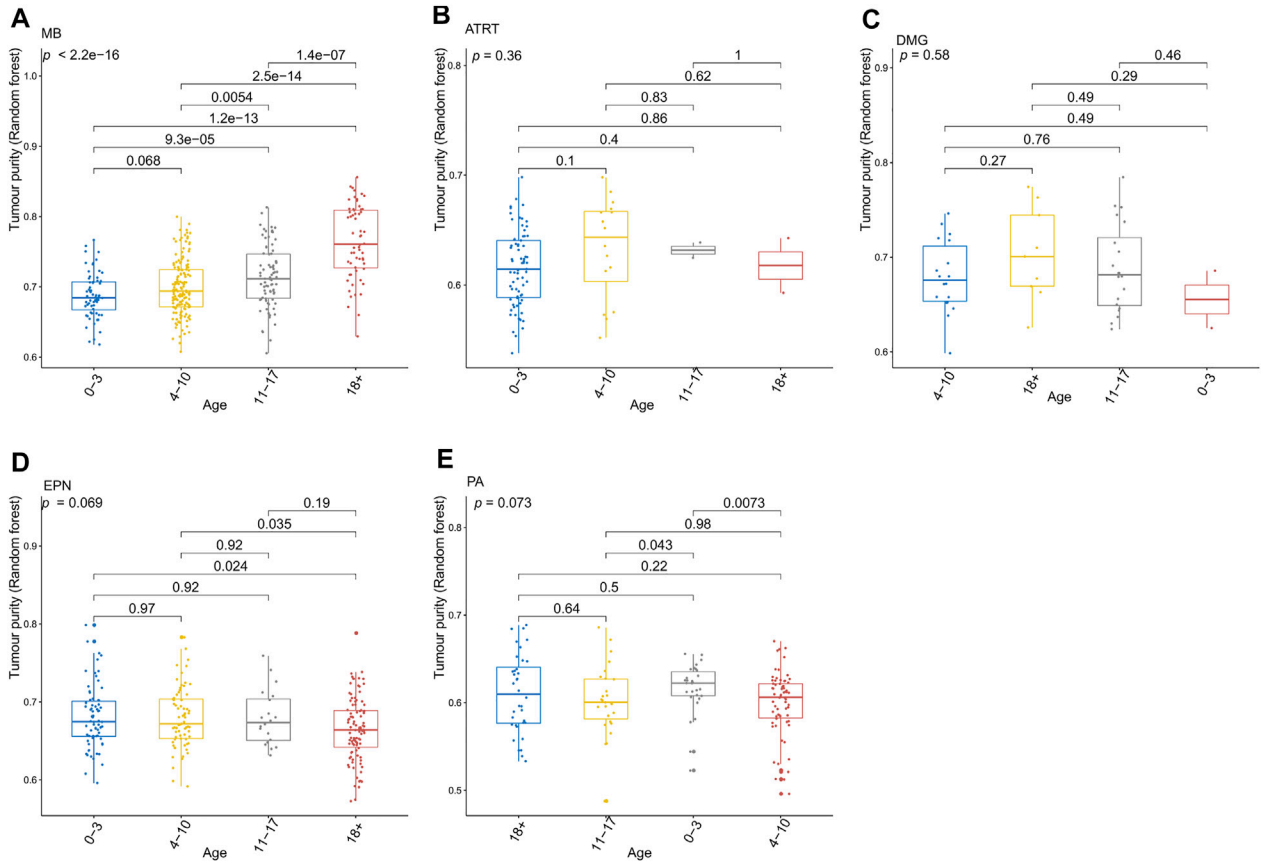


FIGURE 3 | Relationship between the tumor purity of each pediatric CNS tumor and the patient age at diagnosis in the GSE90496 dataset. **(A)** Box plot of the tumor purity in medulloblastoma (MB) samples by the age at diagnosis; **(B)** box plot of the tumor purity in atypical teratoma/rhabdoid tumor samples by the age at diagnosis; **(C)** box plot of the tumor purity in diffuse midline glioma samples by the age at diagnosis; **(D)** box plot of the tumor purity in ependymoma (EPN) samples by the age at diagnosis; **(E)** box plot of the tumor purity in pilocytic astrocytoma samples by the age at diagnosis.

purity: 39.8 ± 37.6%), while that based on the ESTIMATE algorithm was skewed and focused on 80% or more of the total area (with an average tumor purity of 96.99 ± 3.3%), but the tumor purity resulting from the RF algorithm was normal, with an average tumor purity of 73.7 ± 4.5%. When applied to the E-MTAB-5528 dataset (diffuse midline glioma) (**Figure 1C**), the InfiniumPurify algorithm determined the tumor purity to be skewed and the average tumor purity to be 74.04 ± 12.4%, while the RF algorithm generated normal tumor purity, with an average value of 69.5 ± 5.5%. For the GSE64415 and GSE65362 datasets (EPN) (**Figures 1D,E**), the tumor purity based on the ESTIMATE algorithm was skewed, with an average value of 85.95 ± 8.01%, and that based on InfiniumPurify was also skewed, with an average value of 67.1 ± 22.4%, but that based on the RF algorithm was normal, with an average value of 68.4 ± 4.6%. For the GSE44971 dataset (pilocytic astrocytoma) (**Figure 1F**), the average tumor purities generated were 59.4 ± 6.9, 74.8 ± 11.8, and 59.9 ± 5.5% for InfiniumPurify, ESTIMATE, and RF, respectively, but they were all skewed. For the GSE64019 dataset (atypical teratoma/rhomboid tumor), the tumor purity distributed according to the ESTIMATE algorithm was skewed, with an average tumor purity of 87.4 ± 8.1%.

Judging from these results, the distribution of pediatric CNS tumors resulting from the ESTIMATE algorithm was skewed and focused on the part with over 70% of the total area, and the tumor purity distributions based on InfiniumPurify and RF were skewed and normal, respectively. The ESTIMATE method estimates purity indirectly by measuring stromal and immune counterparts in the tumor sample (Yoshihara et al., 2013). Therefore, the presence of non-stromal and immune cells in a cancer sample, such as contaminating adjacent normal cells, could affect ESTIMATE-based tumor purity estimation. In addition, the InfiniumPurify method estimates purity indirectly by identifying differentially methylated regions between cancer and normal samples (Zheng et al., 2017). However, paired normal controls were lacking in our pediatric pan–central nervous system tumor analysis. Although the InfiniumPurify method has a control-free variant, this is only applicable for tumor entities that are included in the TCGA datasets and not suitable for entities from the pediatric spectrum that we have used here. In contrast to the ESTIMATE and InfiniumPurify purity estimates,

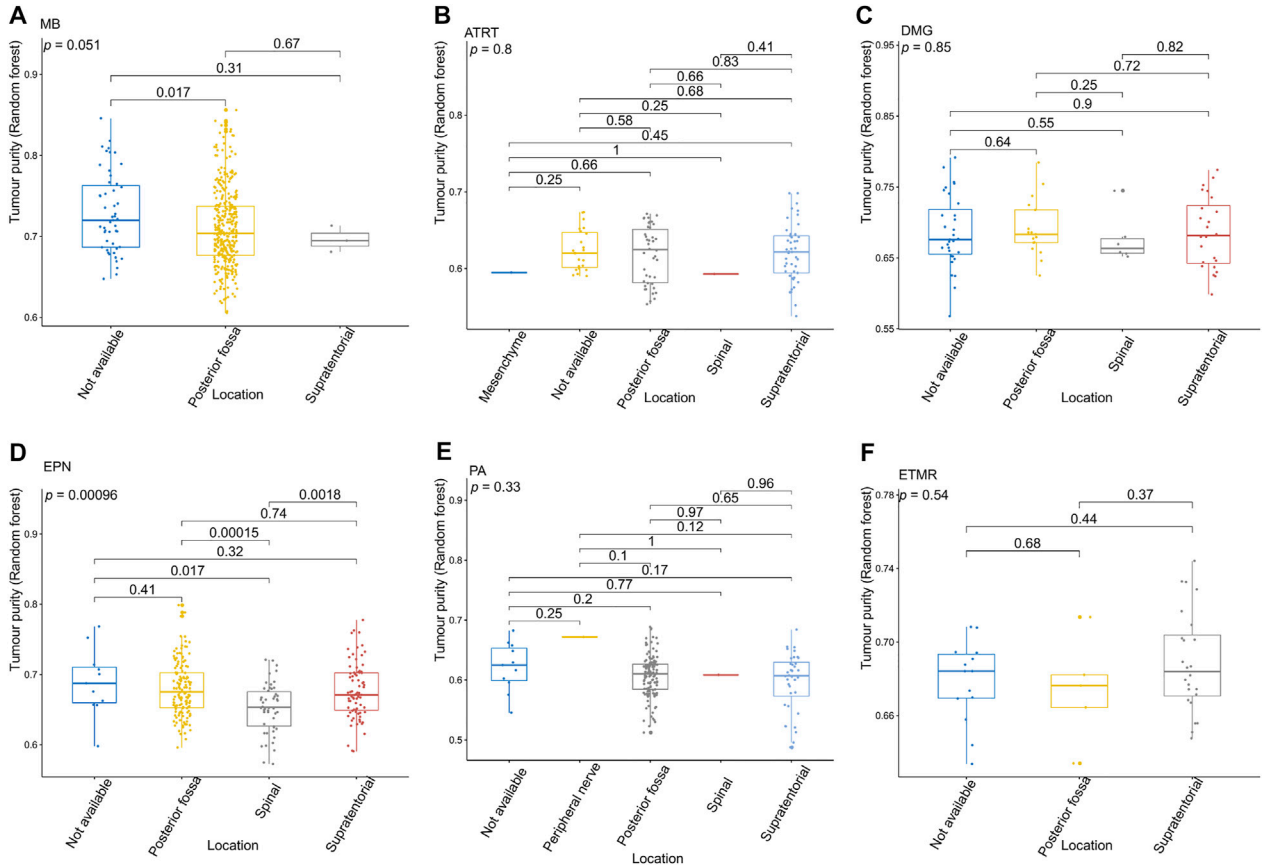


FIGURE 4 | Relationship between the tumor purity of each pediatric CNS tumor and the tumor location in the GSE90496 dataset. **(A)** Box plot of the tumor purity in medulloblastoma (MB) samples by the tumor location; **(B)** box plot of the tumor purity in atypical teratoma/rhabdoid tumor samples by the tumor location; **(C)** box plot of the tumor purity in diffuse midline glioma samples by the tumor location; **(D)** box plot of the tumor purity in ependymoma (EPN) samples by the tumor location; **(E)** box plot of the tumor purity in pilocytic astrocytoma samples by the tumor location; **(F)** box plot of the tumor purity in embryonic tumors of multilayered rosettes by the tumor location.

ABSOLUTE is a direct measure of the cancer cells in a sample (Carter et al., 2012). Taken together, we selected the ABSOLUTE-based RF method for the purity estimation of pediatric CNS tumors in this study, and all subsequent studies were based on the RF algorithm.

Tumor Purity and Molecular and Clinicopathologic Features

Figures 2–4 illustrate the relationships between tumor purity and the patients' clinical features in the GSE90496 dataset. For the tumor histology (**Figures 2A,B**), we observed that MB had the highest purity, whereas pilocytic astrocytoma and atypical teratoma/rhabdoid tumors had the lowest purity (p < 2.2e−16). For the age at diagnosis (**Figures 2A,C**), we found that the patients aged 0–3 years had the lowest tumor purity, while those older than 11 years had the highest purity (p = 1.5e−06). For the tumor grade (**Figures 2A,D**), the purity of Grade I tumor was the lowest, while that of Grade IV was the highest (p < 2.2e−16). For the tumor location (**Figures 2A,E**), we found that the purity of tumors located in the posterior cranial fossa was higher than that in the supratentorial parts (p = 8.5e−07). Regarding the tumor stage (**Figures 2A,F**), compared with primary tumors, recurrent tumors had lower purity (p = 0.019). For patient sex (**Figures 2A,G**), we observed higher tumor purity in male patients (p = 0.044) than in female patients.

Figure 3 presents the relationship between tumor purity and the age at diagnosis in each type of pediatric CNS tumor in the GSE90496 dataset. We found a positive correlation between tumor purity and the age at diagnosis in MB (p < 2.2e−16, **Figure 3A**) but not in other pediatric CNS tumors (including atypical teratoma/rhabdoid tumor, diffuse midline glioma, EPN, and pilocytic astrocytoma) (**Figures 3B–E**). As shown in **Figures 4A–F**, among six pediatric CNS tumors, no significant difference was detected between tumors located in the posterior cranial fossa and those in supratentorial sites in terms of tumor purity. The relationships between MB purity and clinicopathologic features in the GSE85218 dataset are shown in **Figure 5**. The four molecular subgroups of MB

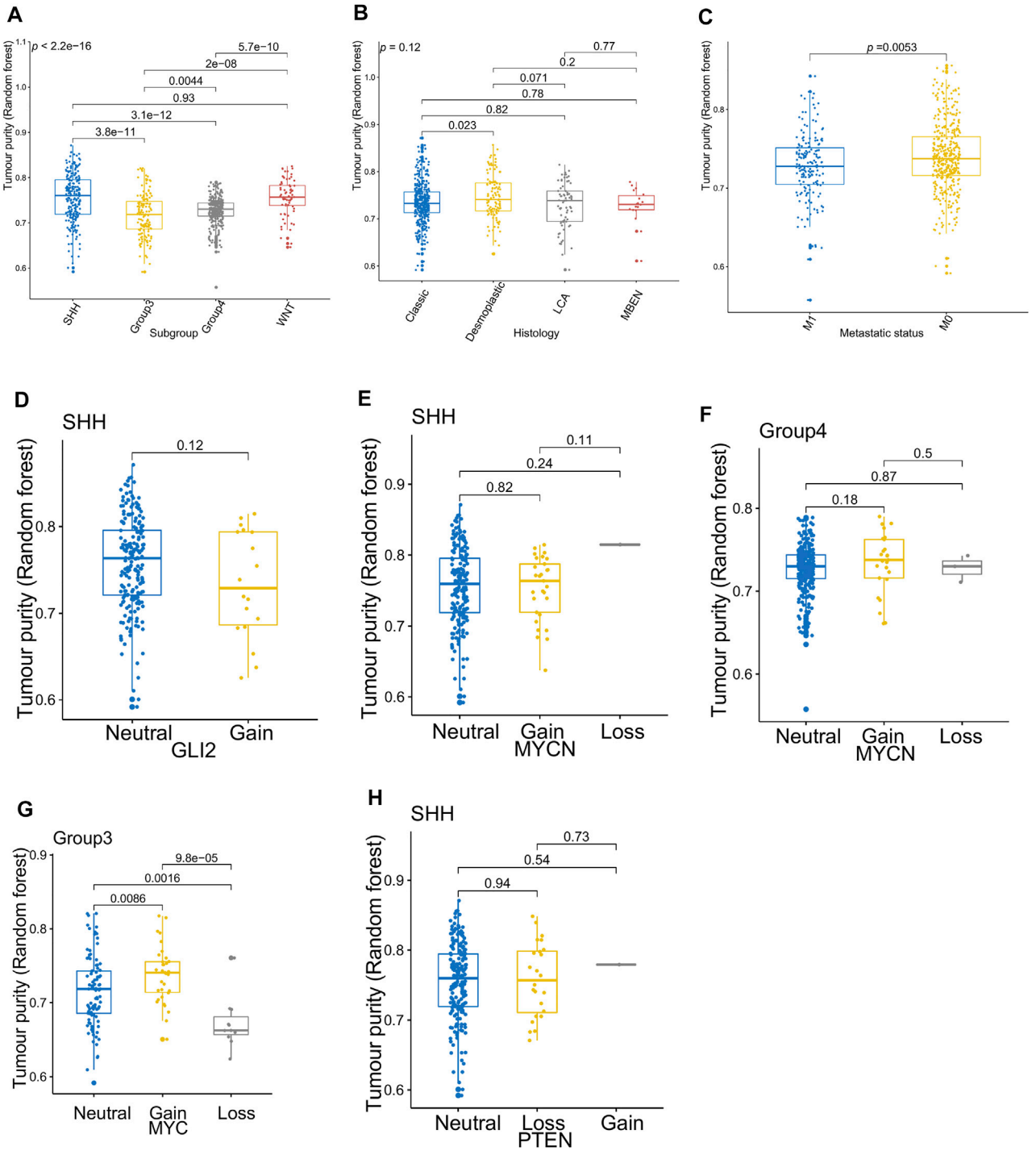


FIGURE 5 | Relationship between the tumor purity of the medulloblastoma (MB) and its clinical molecular features in the GSE85218 dataset. **(A)** Box plot of the tumor purity by molecular subgroup; **(B)** box plot of the tumor purity by histopathology; **(C)** box plot of the tumor purity by metastatic status; **(D)** box plot of the tumor purity of the SHH subgroup by GLI amplification; **(E)** box plot of the tumor purity of the SHH subgroup by MYCN amplification; **(F)** box plot of the tumor purity of the Group 4 subgroup by MYCN amplification; **(G)** box plot of the tumor purity of the Group 3 subgroup by MYC amplification; **(H)** box plot of the tumor purity of the SHH subgroup by PTEN deletion.

(**Figure 5A**) differed greatly from each other in terms of tumor purity (p < 2.2e−16). Compared with the non-WNT/SHH (Groups 3 and 4) subgroups of MB with an inferior prognosis, the WNT and SHH subgroups with a superior prognosis had a higher tumor purity. For the metastatic status of MB patients (**Figure 5C**), non-metastatic patients

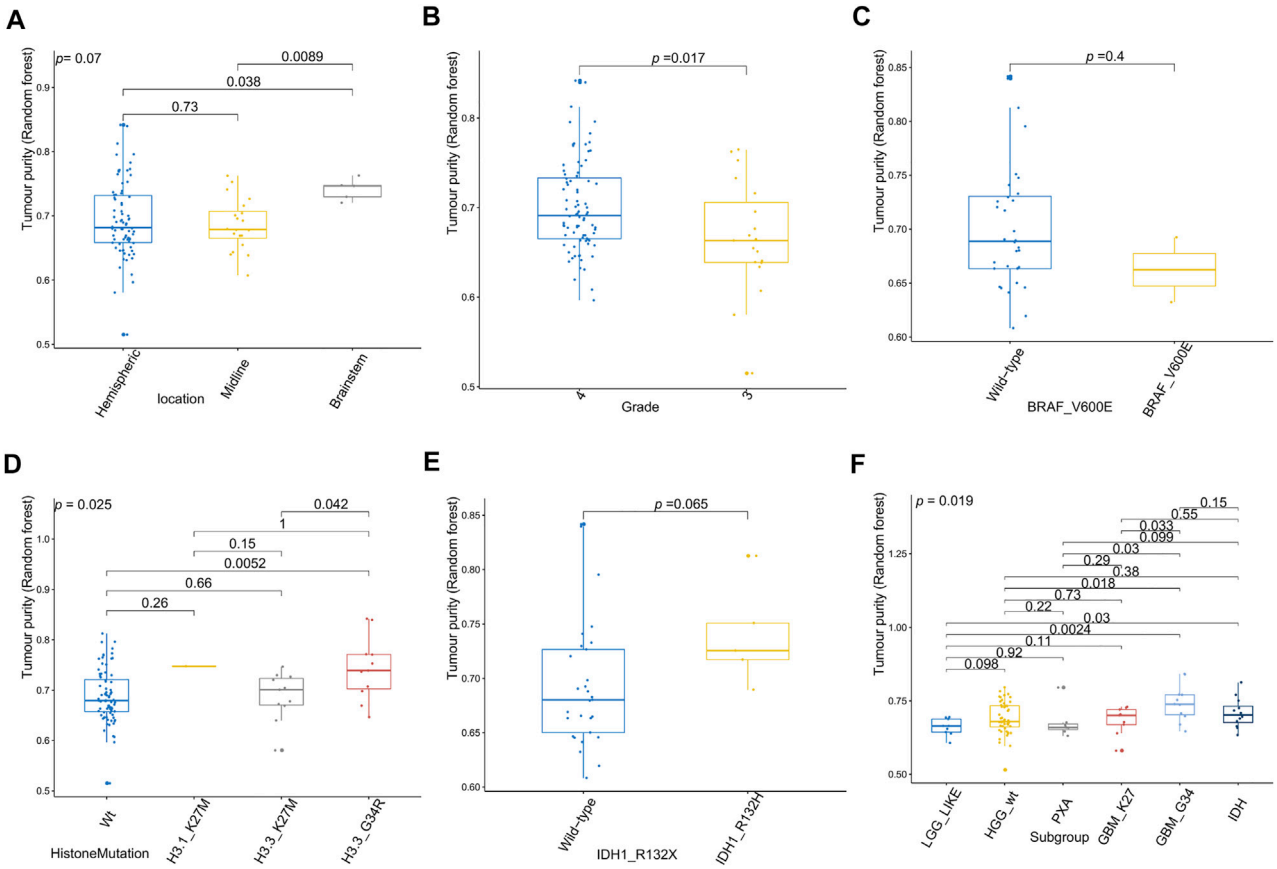


FIGURE 6 | Relationship between tumor purity of high-grade glioma (HGG) and the patients' clinical molecular features in the E-MTAB-5528 dataset. **(A)** Box plot of the tumor purity by tumor location; **(B)** box plot of the tumor purity by tumor grade; **(C)** box plot of the tumor purity by BRAF_V600E mutation status; **(D)** box plot of the tumor purity by histone mutation status; **(E)** box plot of the tumor purity by IDH1 mutation status; **(F)** box plot of the tumor purity by molecular subgroup.

had higher tumor purity than metastatic patients ($p = 0.0053$). For the MYC amplifications of Group 3 MB patients (**Figure 5G**), the tumor purity of Group 3 MB with MYC amplifications was significantly different from that of Group 3 MB without MYC amplifications (MYC amplifications vs. MYC balance, $p = 0.0086$; MYC amplifications vs. MYC deletion, $p = 9.8e\text{-}05$). However, no significant difference was detected among all of the groups in tumor purity when other clinical and molecular features of MB were taken into account (**Figures 5B,D–F,H**).

Figure 6 shows the relationships between high-grade glioma tumor purity and the other clinicopathologic and molecular features in the E-MTAB-5528 dataset. However, for the tumor location (**Figure 6A**), none of the groups were significantly different from each other in tumor purity. For the tumor grade (**Figure 6B**), we found that the tumor purity of Grade IV patients was higher than that of Grade III patients ($p = 0.017$). Regarding BRAF_V600E mutation status (**Figure 6C**), no evident difference was found between the wild-type BRAF patients and mutant-type BRAF patients in tumor purity. For histone mutation status (**Figure 6D**), the tumor purity of subgroups divided by histone H3 mutation differed significantly ($p = 0.025$). For IDH1 mutation status (**Figure 6H**), patients with wild-type IDH1 were not significantly different from those with mutant-type IDH1 in tumor purity. Regarding the molecular subgroup (**Figure 6F**), a significant difference was detected between all of the molecular subgroups of high-grade glioma in tumor purity ($p = 0.019$).

Functional Annotation of Transcriptomic Analysis in Tumor Purity

Since only the MB samples in the GSE85218 dataset came with gene expression and DNA methylation data as well as complete clinical information, we performed an analysis of tumor purity–related biological functions in this dataset. GO analysis revealed that many development-associated pathways were related to tumor purity (**Figure 7A**). Gene set enrichment analysis determined the top three biological pathways, including the MYC signaling pathway, DNA repair pathway, and E2F targets signaling pathway (**Figure 7B**). According to GSVA, the MYC signaling, DNA repair,

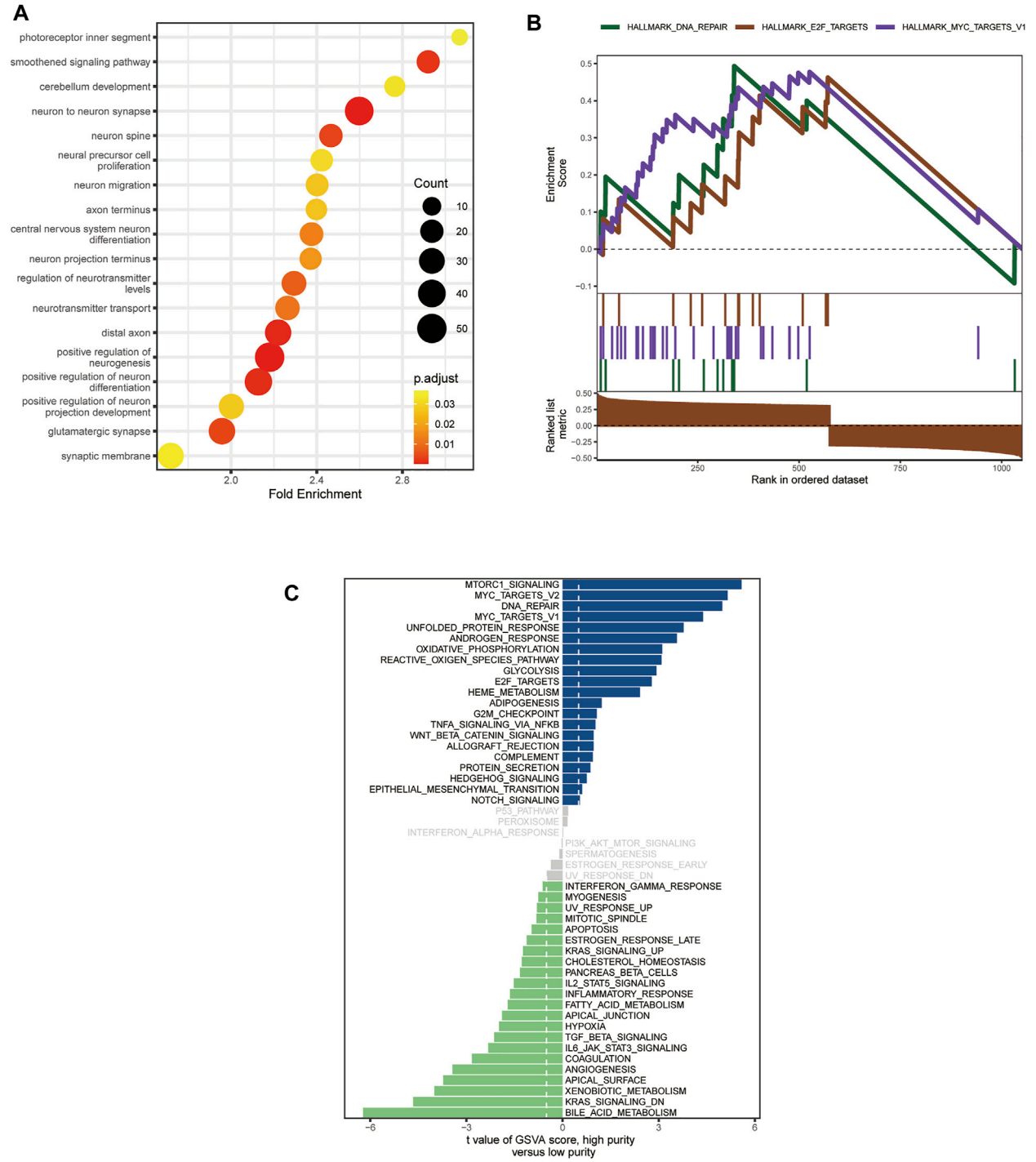


FIGURE 7 | Biological functions related to the purity of medulloblastoma (MB) in the GSE85218 dataset. **(A)** Biological pathways related to the MB purity as revealed by Gene Ontology (GO) analysis; **(B)** biological pathways related to Mb purity as revealed by gene set enrichment analysis (GSEA); **(C)** biological pathways related to MB purity as revealed by gene set variation analysis (GSVA).

glycolysis, WNT signaling, Hedgehog signaling, mTORC1 signaling, and oxidative phosphorylation pathways were positively related to tumor purity, whereas the KRAS signaling, IL2-STAT5 signaling, inflammatory response, and angiogenesis pathways were negatively related to tumor purity (**Figure 7C**).

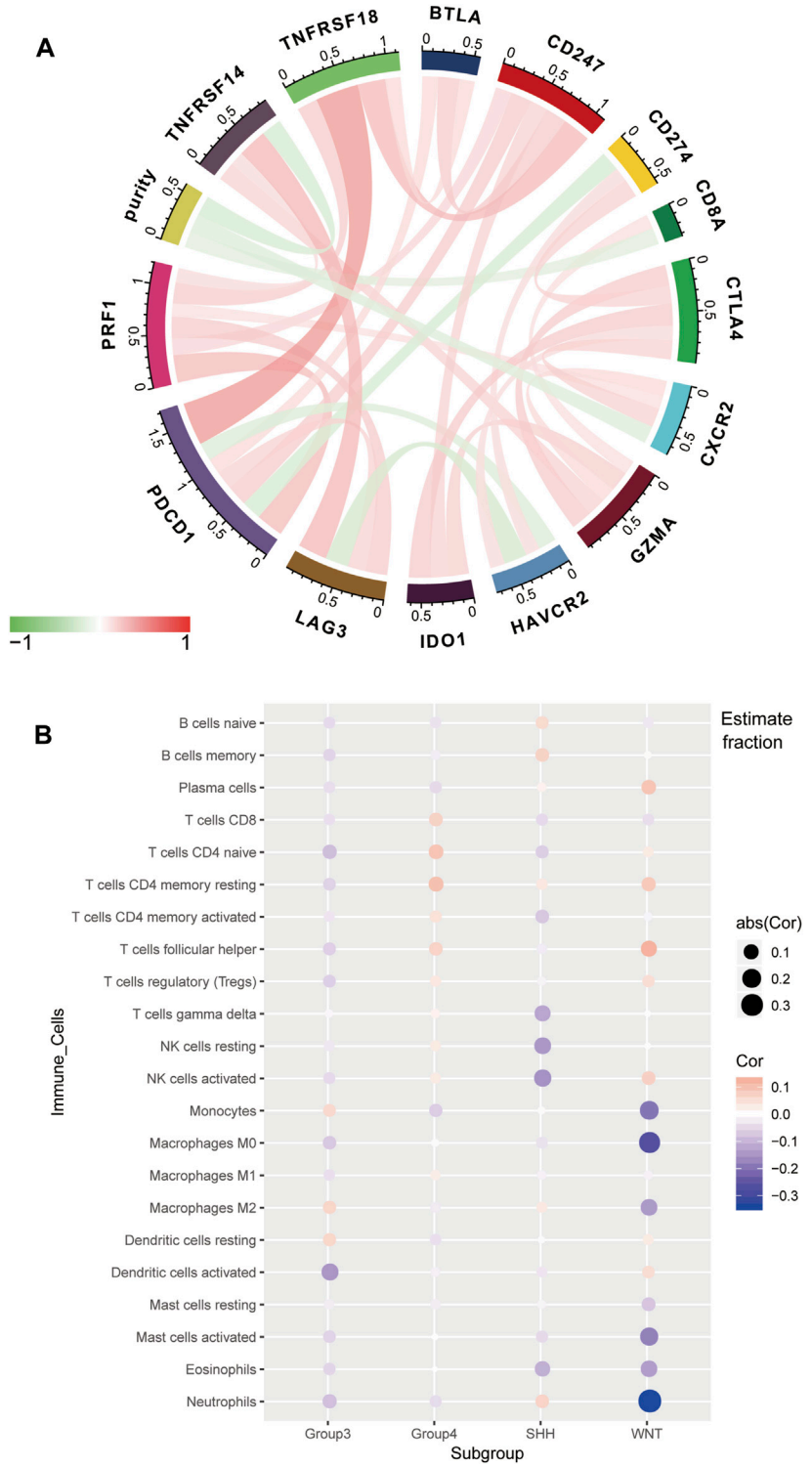


FIGURE 8 | Relationship between the purity of medulloblastoma (MB) and the immune microenvironment in the GSE85218 dataset. **(A)** Relationship between the MB purity and immune-related genes; **(B)** correlation between the purity of each MB subgroup and the CIBERSORT-based infiltrating immunocyte proportions.

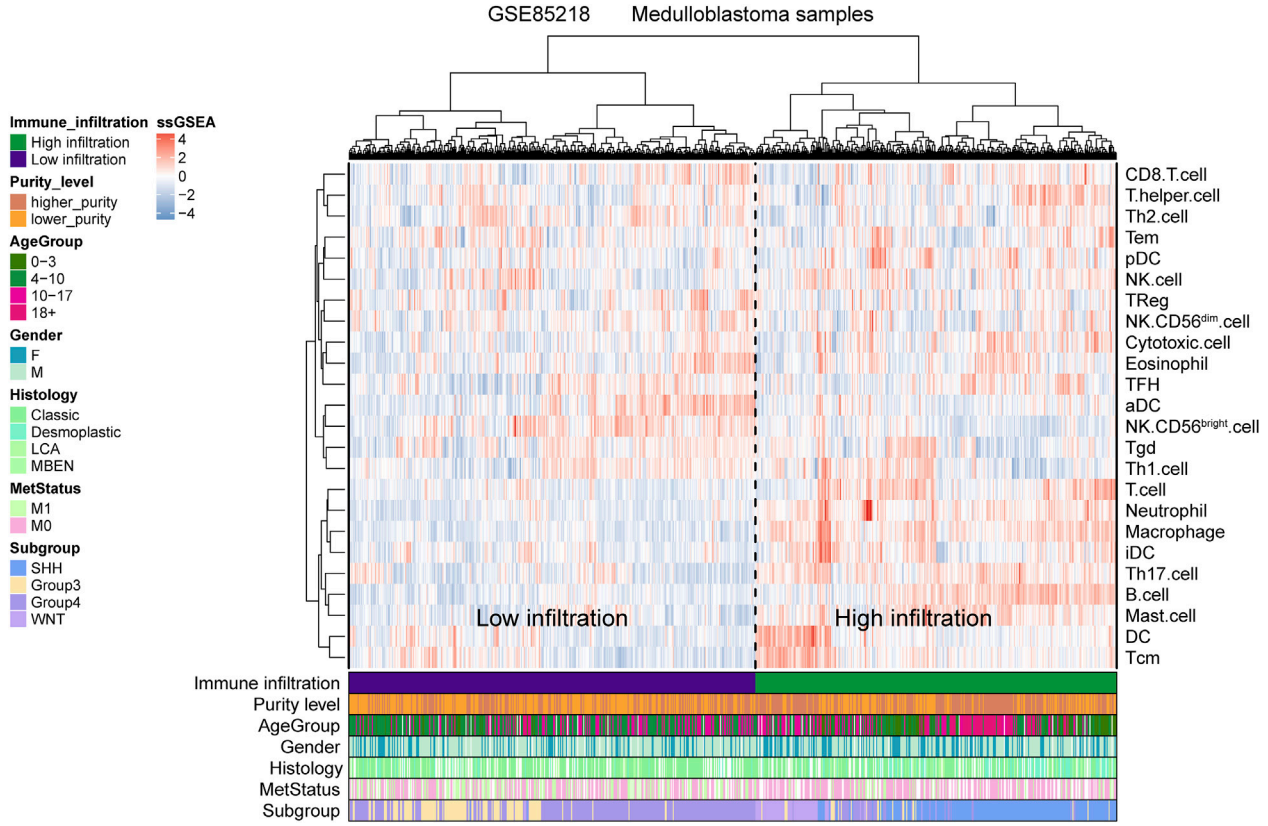


FIGURE 9 | Relationship between the medulloblastoma (MB) purity and the clinical features and immune microenvironment in the GSE85218 dataset as revealed by the single-sample gene set enrichment analysis (ssGSEA).

Tumor Immune Microenvironment and Tumor Purity

For the GSE85218 dataset, we also identified the relationship between tumor purity and the immune microenvironment. As indicated in **Figure 8A**, we found that tumor purity was only negatively related to three immune genes, namely, CD8A (R = −0.18, p = 1.06 e-06), CXCR2 (R = −0.18, p = 2.90 e-07), and TNFRSF14 (R = −0.21, p = 2.58 e-09), but not to other immune-related genes, including the well-known PD1, PD-L1, and CTLA4. **Figure 8B** reveals the correlation between the tumor purity of each subgroup of MB and CIBERSORT-based proportions of infiltrating immunocytes. In WNT MB, only neutrophils were significantly negatively related to tumor purity (R = 0.34, p = 0.004). For SHH MB, only natural killer cells were significantly negatively related to tumor purity (resting, R = −0.14, p = 0.03; activated, R = −0.15, p = 0.02). However, no statistical correlation was detected between the tumor purity and infiltrating immunocyte proportions in Groups 3 and 4 MB. As shown in **Figure 9**, WNT and SHH MBs were significantly enriched in the high–immunocyte infiltration group, whereas Groups 3 and 4 MBs were more enriched in the low–immunocyte infiltration group.

The Prognostic Role of Tumor Purity

Since only the GSE85218, GSE117130, and E-MTAB-5528 datasets included clinical outcome data, they were used to assess the relationship between tumor purity and clinical outcome. For each type of pediatric CNS tumor, we divided the patients into a high-purity group and a low-purity group. As shown in **Figures 10A–H**, the two groups did not differ much in terms of survival rate in all of the CNS tumor datasets. The aforementioned findings suggest that among all pediatric CNS tumors, the association between tumor purity and patient prognosis may be weak.

DISCUSSION

With the development of high-throughput techniques, many novel computation methods based on bioinformatics could be employed to infer tumor purity. In contrast to those based on histopathology, bioinformatics algorithms elicit more highly concordant and objective results. In this study, we performed a comprehensive purity analysis of pediatric CNS tumors with DNA methylation data and gene expression data from several CNS tumor–related large sample datasets on the basis of three tumor purity calculation methods (namely, RF, InfiniumPurify, and ESTIMATE). We found that only the RF estimation approach could produce normally distributed tumor purity.

These results suggest that 1) to prevent bias arising from the introduction of other tumor molecular data, we should employ

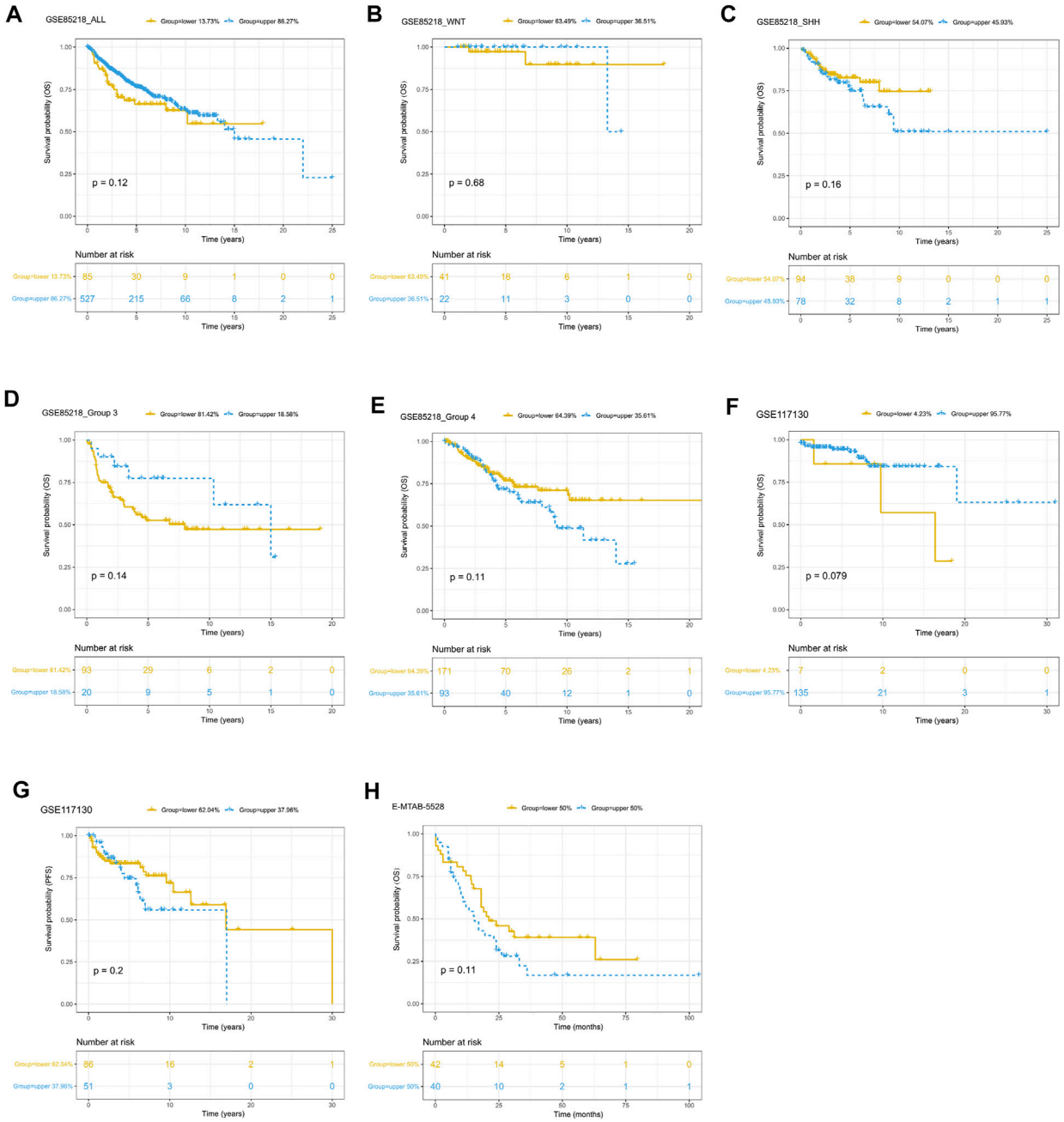


FIGURE 10 | Prognostic role of tumor purity in three pediatric CNS tumor datasets. **(A)** Kaplan–Meier (K-M) curves for overall survival according to tumor purity in the GSE85218 dataset; **(B)** K-M curves for overall survival according to tumor purity of the WNT subgroup medulloblastoma (MB) in the GSE85218 dataset; **(C)** K-M curves for overall survival according to the tumor purity of the SHH subgroup MB in the GSE85218 dataset; **(D)** K-M curves for overall survival according to the tumor purity of Group 3 subgroup MB in the GSE85218 dataset; **(E)** K-M curves for overall survival according to the tumor purity of Group 4 subgroup MB in the GSE85218 dataset; **(F)** K-M curves for overall survival according to the tumor purity in the GSE117130 dataset; **(G)** K-M curves for progression-free survival according to the tumor purity in the GSE117130 dataset; **(H)** K-M curves for overall survival according to the tumor purity in the E-MTAB-5528 dataset.

high-throughput data of the same tumor type (glioma in this study) to construct a prediction model for estimating tumor purity; and 2) given that the presence of non-immune and stromal cells in CNS tumor tissues may affect the purity estimation results of indirect algorithms such as ESTIMATE, it is more reasonable to choose direct methods of tumor purity estimation. We found that there was some relationship between pediatric CNS tumor purity and the molecular and clinicopathologic features. These findings suggested that tumor purity may be an intrinsic characteristic of pediatric CNS tumors. When analyzing the purity of MB in a systematic way, we discovered that tumor purity was lower in Groups 3 and 4

MBs with a worse prognosis than in WNT and SHH MBs. This is consistent with previous studies with regard to glioma purity (Zhang et al., 2017). A possible reason for this is that Groups 3 and 4 MBs are more inclined to undergo metastasis and tumor cell spreading and have difficulty forming dense solid bulks.

An enrichment analysis of MB purity–related biological pathways unveiled some classical signaling pathways related to the biology of MB, including MYC, WNT, and Hedgehog pathways (Northcott et al., 2011). For instance, the WNT pathway is enriched in WNT MB, and the sonic Hedgehog pathway is enriched in SHH MB (Northcott et al., 2011; Ramaswamy and Taylor, 2017; Wang et al., 2018). Moreover, amplification of the MYC oncogene is the most common genetic alteration of Group 3 MB (Ramaswamy and Taylor, 2017; Wang et al., 2018). In addition, we found that some development-associated pathways were associated with tumor purity; thus, abnormalities in such pathways may lead to the occurrence of MB. In the correlation analysis of MB purity and the immune microenvironment, three genes related to immunity, namely, CD8A, CXCR2, and TNFRSF14, were negatively related to tumor purity. These findings suggested that such immune-related genes may be potential targets for immune microenvironment–specific MB therapies. On the other hand, genes related to classical immunosuppression checkpoints, such as PD-1, PD-L1, and CTLA4, were not significantly associated with MB purity. This finding indicates that the efficacy of immunotherapies with PD-1, PD-L1, and CTLA4 inhibitors may be limited to MB. In addition, most infiltrating immunocytes were unrelated to MB purity, indicating that immunocyte-based therapies may also be limited to MB.

While exploring the tumor purity–related survival analyses of MB, EPN, and pediatric high-grade glioma, we confirmed that the effect of tumor purity was insignificant for the survival of patients. These results are inconsistent with previous studies on tumor purity (Aran et al., 2015; Zhang et al., 2017). Cancer cells are capable of recruiting immune infiltrating cells to the glioma microenvironment (Silver et al., 2016), which could influence the prognosis of glioma patients (Zhang et al., 2017). However, childhood brain tumors are considered to be relatively immunologically "cold" due to the lack of genetic mutations (Gröbner et al., 2018). Furthermore, Bockmayr et al. did not observe associations between intratumoral immune infiltrates and MB survival, and they attributed their results to the overall very low immune infiltration (Bockmayr et al., 2018). The hypothesis that the ability of pediatric CNS tumors to recruit immune infiltrating cells is relatively weak may provide a direction for why tumor purity does not influence the overall survival of pediatric CNS tumor patients. In addition, these results may indirectly confirm the difference between children's CNS tumors and adults' brain tumors in terms of clinical and molecular features.

Nevertheless, the present work has some limitations. First, our findings require external validation using independent pediatric CNS tumor datasets. Second, due to the retrospective setting of the present study, additional prospective studies are necessary to evaluate our conclusions.

CONCLUSION

We presented a systematic comparison of three tumor purity estimation methods across pediatric CNS tumors and found that the RF algorithm is applicable for pediatric CNS tumor purity estimation. MB purity was significantly associated with some classical signaling pathways associated with MB biology and development-related pathways. Furthermore, our analysis showed a minor effect of tumor purity on the survival of pediatric patients with CNS tumors. It is important for future studies of pediatric CNS tumors to take tumor purity into account when analyzing high-throughput data from patient samples.

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 in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

JY, JW, HL, and JM contributed to conceptualization; JY, JW, and ST framed methodology; QW, YZ, BW, LC, ZL, and HZ performed formal analysis; JY, JW, ST, and HL contributed to writing—original draft preparation; JM, HL, and JY helped with writing—review and editing; JM supervised the study; JM assisted with project administration; and JM and HZ acquired funding.

FUNDING

This work was supported by the Shanghai Xin Hua Hospital (JZPI201701 to JM), Shanghai Shenkang Hospital Development Center (16CR2031B to JM), Shanghai Science and Technology Committee (17411951800 to JM), and Chinese National Science Foundation for Young Scholars (81702453 to YZ).

ACKNOWLEDGMENTS

We thank GZ for the technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.707802/full#supplementary-material

REFERENCES

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic Pan-Cancer Analysis of Tumour Purity. *Nat. Commun.* 6, 8971. doi:10.1038/ncomms9971

Basu, S., Campbell, H. M., Dittel, B. N., and Ray, A. (2010). Purification of Specific Cell Population by Fluorescence Activated Cell Sorting (FACS). *J. Vis. Exp.*, 1546. doi:10.3791/1546

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39, 782–795. doi:10.1016/j.immuni.2013.10.003

Bockmayr, M., Mohme, M., Klauschen, F., Winkler, B., Budczies, J., Rutkowski, S., et al. (2018). Subgroup-specific Immune and Stromal Microenvironment in Medulloblastoma. *Oncoimmunology* 7, e1462430. doi:10.1080/2162402x.2018.1462430

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA Methylation-Based Classification of central Nervous System Tumours. *Nature* 555, 469–474. doi:10.1038/nature26000

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., et al. (2012). Absolute Quantification of Somatic DNA Alterations in Human Cancer. *Nat. Biotechnol.* 30, 413–421. doi:10.1038/nbt.2203

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* 164, 550–563. doi:10.1016/j.cell.2015.12.028

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The Prognostic Landscape of Genes and Infiltrating Immune Cells across Human Cancers. *Nat. Med.* 21, 938–945. doi:10.1038/nm.3909

Gröbner, S. N., Worst, B. C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V. A., et al. (2018). The Landscape of Genomic Alterations across Childhood Cancers. *Nature* 555, 321–327. doi:10.1038/nature25480

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Kumar, R., Liu, A. P. Y., Orr, B. A., Northcott, P. A., and Robinson, G. W. (2018). Advances in the Classification of Pediatric Brain Tumors through DNA Methylation Profiling: From Research Tool to Frontline Diagnostic. *Cancer* 124, 4168–4180. doi:10.1002/cncr.31583

Northcott, P. A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C. G., Mack, S., et al. (2011). Medulloblastoma Comprises Four Distinct Molecular Variants. *Jco* 29, 1408–1414. doi:10.1200/jco.2009.27.4324

Ramaswamy, V., and Taylor, M. D. (2017). Medulloblastoma: From Myth to Molecular. *Jco* 35, 2355–2363. doi:10.1200/jco.2017.72.7842

Rhee, J.-K., Jung, Y. C., Kim, K. R., Yoo, J., Kim, J., Lee, Y.-J., et al. (2018). Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types. *Cancer Immunol. Res.* 6, 87–97. doi:10.1158/2326-6066.cir-17-0201

Schmitz, B., Radbruch, A., Kümmel, T., Wickenhauser, C., Korb, H., Hansmann, M. L., et al. (1994). Magnetic Activated Cell Sorting (MACS)--a New Immunomagnetic Method for Megakaryocytic Cell Isolation: Comparison of Different Separation Techniques. *Eur. J. Haematol.* 52, 267–275. doi:10.1111/j.1600-0609.1994.tb00095.x

Schreiber, R. D., Old, L. J., and Smyth, M. J. (2011). Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science* 331, 1565–1570. doi:10.1126/science.1203486

Silver, D. J., Sinyuk, M., Vogelbaum, M. A., Ahluwalia, M. S., and Lathia, J. D. (2016). The Intersection of Cancer, Cancer Stem Cells, and the Immune System: Therapeutic Opportunities. *Neuro Oncol.* 18, 153–159. doi:10.1093/neuonc/nov157

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Wang, J., Garancher, A., Ramaswamy, V., and Wechsler-Reya, R. J. (2018). Medulloblastoma: From Molecular Subgroups to Molecular Targeted Therapies. *Annu. Rev. Neurosci.* 41, 207–232. doi:10.1146/annurev-neuro-070815-013838

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhang, C., Cheng, W., Ren, X., Wang, Z., Liu, X., Li, G., et al. (2017). Tumor Purity as an Underlying Key Factor in Glioma. *Clin. Cancer Res.* 23, 6279–6291. doi:10.1158/1078-0432.ccr-16-2598

Zheng, X., Zhang, N., Wu, H.-J., and Wu, H. (2017). Estimating and Accounting for Tumor Purity in the Analysis of DNA Methylation Data from Cancer Studies. *Genome Biol.* 18, 17. doi:10.1186/s13059-016-1143-5

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 Oncology

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fonc.2021.771173

Imaging-Based Machine Learning Analysis of Patient-Derived Tumor Organoid Drug Response

Erin R. Spiller[1], Nolan Ung[1], Seungil Kim[1], Katherin Patsch[1], Roy Lau[1], Carly Strelez[1], Chirag Doshi[1], Sarah Choung[1], Brandon Choi[1], Edwin Francisco Juarez Rosales[1,2], Heinz-Josef Lenz[3], Naim Matasci[1†] and Shannon M. Mumenthaler[1,3†]*

[1] Lawrence J. Ellison Institute for Transformative Medicine of USC, Los Angeles, CA, United States, [2] Department of Medicine, University of California San Diego, La Jolla, CA, United States, [3] Division of Medical Oncology, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

Three-quarters of compounds that enter clinical trials fail to make it to market due to safety or efficacy concerns. This statistic strongly suggests a need for better screening methods that result in improved translatability of compounds during the preclinical testing period. Patient-derived organoids have been touted as a promising 3D preclinical model system to impact the drug discovery pipeline, particularly in oncology. However, assessing drug efficacy in such models poses its own set of challenges, and traditional cell viability readouts fail to leverage some of the advantages that the organoid systems provide. Consequently, phenotypically evaluating complex 3D cell culture models remains difficult due to intra- and inter-patient organoid size differences, cellular heterogeneities, and temporal response dynamics. Here, we present an image-based high-content assay that provides object level information on 3D patient-derived tumor organoids without the need for vital dyes. Leveraging computer vision, we segment and define organoids as independent regions of interest and obtain morphometric and textural information per organoid. By acquiring brightfield images at different timepoints in a robust, non-destructive manner, we can track the dynamic response of individual organoids to various drugs. Furthermore, to simplify the analysis of the resulting large, complex data files, we developed a web-based data visualization tool, the Organoizer, that is available for public use. Our work demonstrates the feasibility and utility of using imaging, computer vision and machine learning to determine the vital status of individual patient-derived organoids without relying upon vital dyes, thus taking advantage of the characteristics offered by this preclinical model system.

Keywords: patient-derived organoids (PDO), high content imaging, label-free analysis, machine learning, drug response

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Ming-Zhu Jin, Shanghai Jiao Tong University, China
Tiffany Heaster, University of Wisconsin-Madison, United States

****Correspondence:***
Shannon M. Mumenthaler
smumenth@usc.edu

†These authors share last authorship

Specialty section:
This article was submitted to Cancer Molecular Targets and Therapeutics, a section of the journal Frontiers in Oncology

Received: *07 September 2021*
Accepted: *02 December 2021*
Published: *21 December 2021*

Citation:
Spiller ER, Ung N, Kim S, Patsch K, Lau R, Strelez C, Doshi C, Choung S, Choi B, Juarez Rosales EF, Lenz H-J, Matasci N and Mumenthaler SM (2021) Imaging-Based Machine Learning Analysis of Patient-Derived Tumor Organoid Drug Response. Front. Oncol. 11:771173. doi: 10.3389/fonc.2021.771173

INTRODUCTION

High-throughput screening assays have advanced the drug-discovery field by greatly increasing the number of compounds that can be screened and thus the number of positive leads. However, this improvement has yet to produce a corresponding increase in the drugs available for treatment as three-quarters of the drugs that enter clinical trials never make it to market, with a majority failing due to a lack of efficacy (1, 2). Oncology drugs have proven especially challenging, with a predicted success rate of only a 3.4% in clinical trials (3). One important limitation of traditional *in vitro* cancer drug screening methods is the use of oversimplified, immortalized cell lines cultured in 2D, which fails to capture the *in vivo* complexity of human tumors including influences from the surrounding microenvironment and cellular heterogeneity (4–6). To improve the success rate of identifying compounds with promising clinical translation, there is a need for more biomimetic preclinical platforms to carry out these drug testing studies. In this context, patient-derived organoids (PDOs), in which cells obtained from a patient's tumor are grown in a medium that promotes the formation of cellular aggregates that recapitulate important aspects of the original tissue architecture, have gained significant traction in the cancer research field (7–9). Multiple organoid models of human cancers have been developed (10), including gastrointestinal (11), prostate (12), ovarian (13) and pancreatic cancers (14). By more faithfully representing the original physiological environment, these tumor organoid models address some of the limitations of traditional cell line cultures and offer rapid, scalable approaches for patient-specific molecular and phenotypic characterization as well as drug screening (11, 15–17).

Two traditional screening methods typically used to determine compound efficacy are ATP based cell viability assays (18) and vital dyes (VDs) (19). While valuable, both approaches have significant drawbacks in the context of 3D organoid screening. ATP based cell viability assays are disruptive and performed on a pooled population of organoids: as such they do not allow for repeated assaying and mask intra-organoid heterogeneity. Unlike cell viability assays, vital-dye assays are imaging-based and non-disruptive, and therefore, in principle, allow for analysis at multiple timepoints and preserve heterogeneity. However, vital dyes present two significant issues, depending on the specific dye. First, they can have cytotoxic effects and interfere with the outcome readout and, second, they can have transient expression, meaning that the signal indicating a dead cell might peak at a certain timepoint and disappear afterwards. In addition, the per-cell vital dye signal needs to be integrated across multiple cells to obtain a per-organoid viability determination.

Therefore, we propose a label-free high content screening (HCS) method that involves live-cell imaging of colorectal cancer (CRC) patient-derived organoids over time in a robust, non-destructive manner. This approach provides an automated pipeline to visualize cellular dynamics and extract multi-parametric data, which is advantageous for phenotypic screening of PDO models (20). One challenge of HCS platforms is the vast amount of data produced that must be accurately interpreted. To circumvent this bottleneck in analysis pipelines, machine learning (ML) methods can be applied to these large-scale biological data sets. The usefulness of Supervised ML approaches such as linear classifiers and regression models has been demonstrated in analyzing large amounts of data in disparate fields and they are now being used increasingly in the biomedical domain (21–24). Furthermore, computer vision applications can be applied to HCS image data to recognize patterns and changes that are not detectable by the human eye and thus have a huge potential to streamline drug discovery pipelines through screening at a faster pace (25).

Previously, we have shown that imaging-based HCS assays can provide dynamic insight to changes in heterogeneous cellular populations using 2D culture models with a cell-based image analysis method (26). In this study, we trained a linear classifier to discriminate between live and dead PDOs based on a set of morphological and textural features extracted from brightfield images, and then used the trained model to determine drug response of organoids derived from colon cancer patients with heterogeneous clinical histories. Additionally, by collecting the vital status of individual organoids over time we can gain insights into the dynamic aspects of drug response as well as the heterogeneity of response across organoids and across patients. This work showcases the possibilities offered by the application of machine learning approaches to label-free high-content imaging assays.

RESULTS

Generation of a Label-Free Imaging-Based Workflow to Evaluate Patient-Derived Organoids

We established an HCS pipeline that includes label-free temporal imaging of PDOs (**Figure 1A**) coupled with quantitative image analysis using a linear classifier (**Figure 1B**) and data compilation and visualization (**Figure 1C**). To execute this workflow, we utilized PDOs from our biobank generated from CRC patient samples, which includes primary and liver metastatic tumors.

Organoid set up (**Figure 1A**): Briefly, to generate PDOs that recapitulate the morphology of the tissue of origin (**Supplemental Figure 1**), tissue samples were obtained post-surgery, processed, seeded in extracellular matrix, and expanded for future use (see *Methods* section for details). To set up the screening assay, organoids were first digested to a single cell suspension before being seeded into a 96 well plate. Then, using an HCS platform, the samples were imaged at multiple timepoints in brightfield to minimize phototoxicity and photobleaching.

Supervised machine learning algorithm used to classify organoids as live/dead based on phenotypic features (**Figure 1B**): Image analysis was subsequently performed on the maximum intensity projections of multiple z-scan images using a machine learning algorithm that enables users to build a linear classifier by identifying

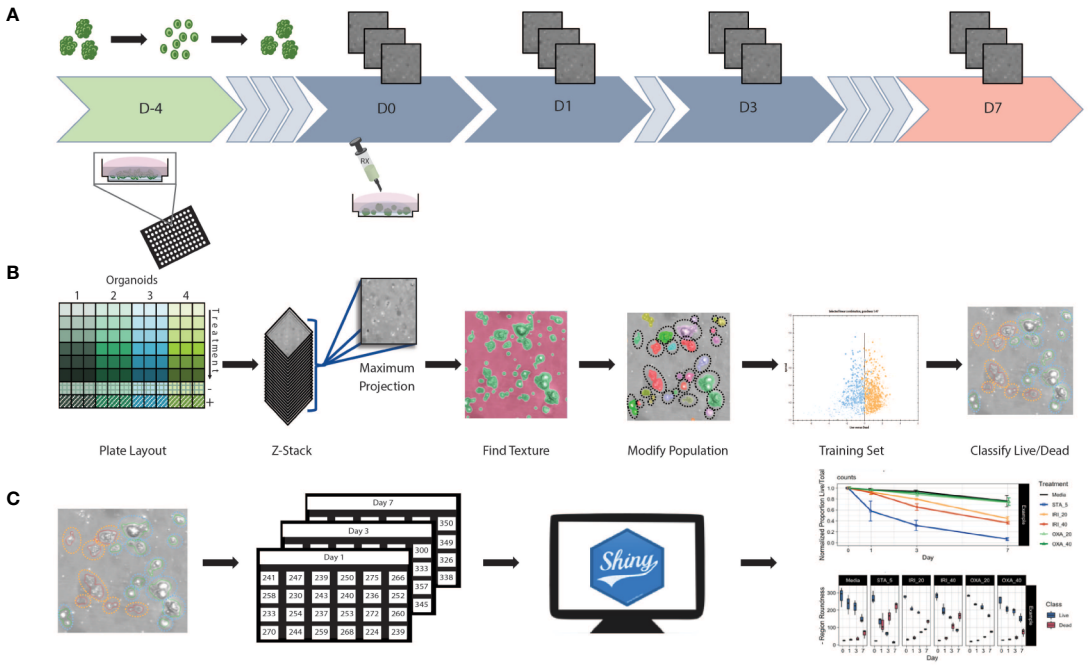


FIGURE 1 | Workflow schematic **(A)** At day -4 organoids were digested to single cells and seeded at 5,000 cells/well in Basement Membrane Extract. Plates were incubated for 4 days allowing organoids to reform. Baseline images were taken at day 0 prior to the initiation of treatment. After initial treatment, plates were re-imaged at days 1, 3, and 7. Media and treatment are refreshed post imaging on day 3, with final measurements taken at day 7. **(B)** Using 96 well plates, multiple patients and/or treatments can be performed with a single assay. Images were obtained in z-stack then combined into a single maximum projection image upon which all further processing and analysis was performed. A textural algorithm was used to identify organoid regions of interest, with a segmenting algorithm applied to split organoids in near proximity to each other. A training set was created by identifying live/dead organoids across untreated and treated samples from each patient. This supervised machine learning algorithm was then applied to experimental data. **(C)** Classification data was compiled into spreadsheets then uploaded to a web-based app for data processing and visualization.

regions of interest (ROIs) that are part of distinct groups. Using a trained feature-based textural machine learning algorithm we divided image regions into two classes: organoid ROIs and background (the ML algorithm was trained on the segmented and annotated images, see the *Methods* section for details). Morphological and textural features were measured for each identified object within the organoid class (**Supplemental Table 1**). STAR morphology features encompass Symmetry properties, Threshold compactness, Axial and Radial properties. Spot-Edge-Ridge (SER) textural features are based on Gaussian derivative images measuring pixel intensity patterns within each ROI. Distributions of all 25 STAR and SER features measured across 6 different PDOs, in media or treated with staurosporine (positive control - apoptosis inducer), are depicted in **Supplemental Figure 2** (see PDO counts and tissue site in **Supplemental Table 2**). At day 3, morphological features are patient-specific and consistent across replicate wells (**Figure 2A**) and unsupervised clustering of the data identified clusters that matched the patient or origin rather than number of days in culture (**Supplemental Figure 3**). Using this ML approach, we can detect morphometric similarities and differences across PDOs.

Statistical analysis and development of data visualization tool (**Figure 1C**): For all 25 textural and morphological features, the measurements for each detected PDO were first summarized (mean value) across all detected objects within a well. Next, the mean and standard deviation were computed from technical replicate wells for each treatment and time-point. A Shiny-based web tool, the Organoizer, was developed to process the data and produce plots for organoid-survival and monitor changes in features over the course of the treatment period (27).

Phenotypic Signatures Correlate to PDO Viability

To create a training data set, we manually classified 179 objects, 80 live (untreated media control) and 99 dead (5 μM staurosporine treated positive control), across 41 images of organoids derived from six different patients at various time points. When applied to new experimental data each detected PDO was assigned to either the "dead" or "live" class by the linear classifier based on 9 significant morphology and texture features chosen by the algorithm to delineate between live and dead PDO categories: SER Valley, SER Edge, SER Ridge, Profile 2/2, Threshold Compactness 60%, Axial Small Length, Area, Ration Width to Length, and Threshold Compactness 50% (**Figure 2B**). Representative images illustrate selected PDO textural and morphological features (i.e., Profile 1/2, and 2/2, SER Edge Ridge and Valley), found to be distinct between live and dead PDOs (**Figure 2C**). The signal to noise ratio of the classifier

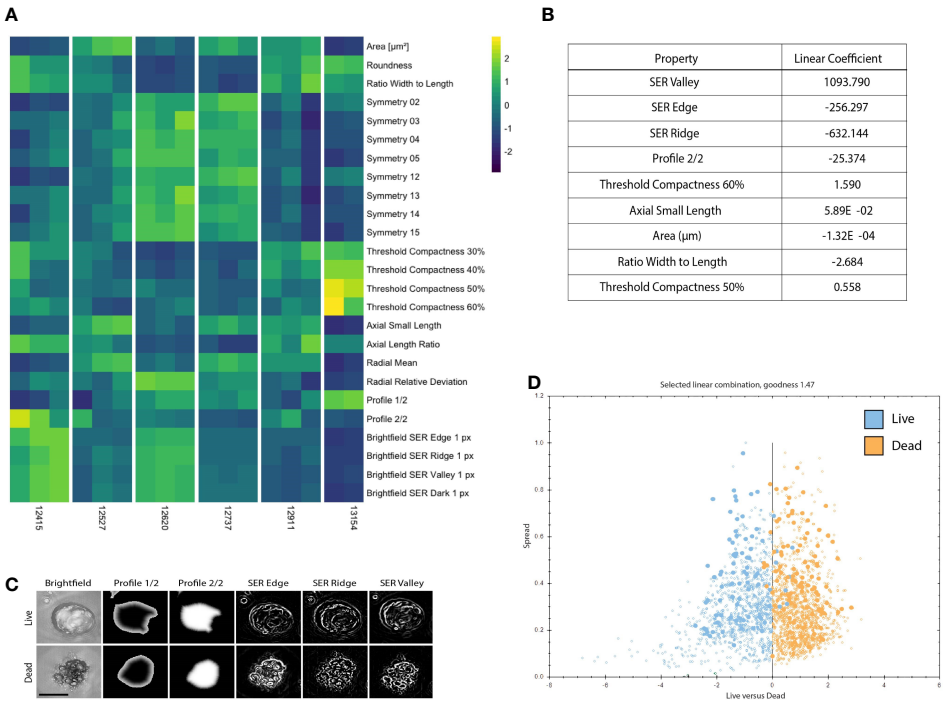


FIGURE 2 | PDOs display distinct texture and morphology features related to viability. **(A)** Heatmap illustrating 25 morphology and texture features (z-score normalized) across 6 different PDOs in the untreated (media only) condition on Day 3. Columns represent replicates for each PDO. **(B)** Features discriminating between live and dead PDOs are listed in order of relevance as indicated by the linear coefficient. **(C)** Representative PDO images of selected morphology and texture features that discriminate between live and dead classification. Scale bar is 50 µm. **(D)** The signal to noise ratio is displayed as goodness of live (blue) versus dead (orange) PDOs manually classified in the training set. Filled circles denote PDOs included in the training set and open circles are classified by the algorithm.

algorithm is expressed as "goodness" based on the distance of the data points from the classifier line, which is visualized using a scatter plot (**Figure 2D**).

As a baseline, we compared our classifier to the visual assessments of trained cell biology experts. Trained cell biologists are adept at visually assessing the health of their cell cultures using brightfield microscopy, however manual classification not only limits throughput but also introduces inter-observer variability (23). To generate a ground truth by visual assessment, we asked 9 scientists to blindly classify images of individual organoids (18 PDOs) as either "live" or "dead" (**Supplemental Figure 4**). Statistical evaluation of inter-rater reliability indicated only moderate agreement between visual classifications (Fleiss' $\kappa = 0.451$, $z = 11.5$, $p = 0.00$; perfect agreement $\kappa=1$), highlighting the inherent variability in subjective manual classification. Using the data set containing the 18 manually classified organoids, we also performed live/dead classifications based on vital dye (VD) intensity from DRAQ7 staining. For our purposes we defined a dead PDO as one that contained at least one DRAQ7$^+$ area ≥ the area of a nucleus. For each organoid we compared the expert consensus classification against DRAQ7 staining results and our linear classifier (**Figure 3A**). We found 78% (14/18) concordance between the linear classifier and expert majority, 61% (11/18) between the linear classifier and DRAQ7, and 61% (11/18) between the expert majority and DRAQ7. In instances where all experts agreed, concordance between the linear classifier and the expert majority increased to 100% (8/8), however expert majority concordance with DRAQ7 only reached 62% (5/8). The strong agreement between the expert classifications and the linear classifier reinforces machine learning as a valuable approach for 3D organoid phenotyping. An important note, the 18 PDO images classified across methods (i.e., experts/ML/VD) were not included in the training set to ensure that there is no "leakage" of information between the training and the testing set.

To further evaluate the performance of the linear classifier using time series data, we compared our algorithm classifications with those made using DRAQ7 for PDO-12620 (**Figure 3B**). We normalized the proportion of live/total PDOs at each timepoint to the proportion of live PDOs at day 0 for untreated and staurosporine treated conditions determined by ML or DRAQ7 staining. For the staurosporine treated group, the ML classifier detected a reduction in the number of live organoids on day 1, whereas DRAQ7 shows a comparable reduction past day 3 (**Figure 3C**). In the untreated group, many organoids are classified as live by both the ML and the VD, with 80% concordance between the two methods. However, the two methods started to diverge upon treatment, with concordance in staurosporine treated organoids dropping to 60% (**Figure 3D**). Additionally, our ML approach allows us to follow individual

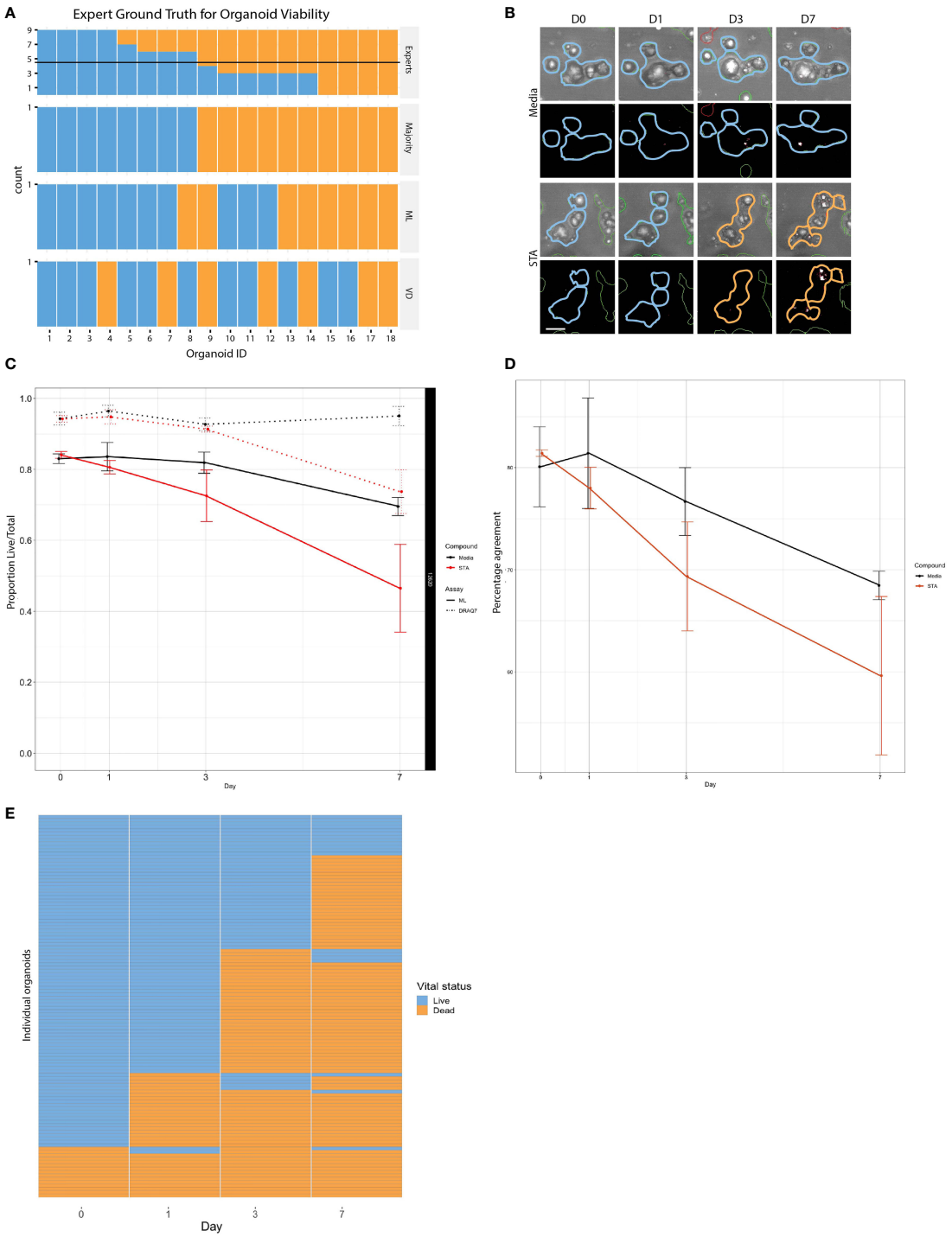


FIGURE 3 | Comparison of live and dead classification by ML and vital dye. **(A)** Classification of 18 different organoid images as either live or dead was determined using three independent methods: tissue culture experts, ML, and DRAQ7, and then compared to determine level of concordance. **(B)** PDOs treated with staurosporine or untreated controls stained with DRAQ7 and classified by ML. Scale bar is 50μm. **(C)** The normalized proportion of live/total organoids (PDO 12620) for the control and staurosporine treated group as determined by both ML and DRAQ7 was plotted over the course of 7 days (error bars: SD of 3 replicate wells per group per timepoint). When classified by ML the difference in response between the treated and untreated groups are seen starting on day 1, whereas VD classification does not start to show separation until after day 3. **(D)** Percentage agreement of ML and DRAQ7 live/dead classification for untreated and staurosporine treated organoids (PDO 12620; error bars are SD of 3 replicate wells each). **(E)** Tracking the vital status of individual organoids (PDO 13154) over 7 days treatment with staurosporine as assessed by our ML classification (N=114 organoids).

organoids over time to determine their vital status (**Figure 3E**). To further clarify the discrepancy between the ML-based and DRAQ7-based classifications we tracked individual PDOs (PDO-13154, N~900) over 7 days (days 0,1,3 and 7) and determined their vital status at each time point by each method (**Supplemental Figure 5**). While the ML classification of dead PDOs increased with time, the population of dead PDOs determined by DRAQ7 decreased. The discordance is most likely due to clearance of cellular debris over time where a PDO previously defined as dead is now DRAQ7 negative. Given staurosporine is an inducer of apoptosis, this result suggests that our ML method may identify dead or dying PDOs more accurately including those that have lost their ability to retain DRAQ7.

Use of Supervised Machine Learning to Track Patient-Specific Drug Response

Tumors evolve over time in response to various stimuli, such as organ-specific microenvironments and drug perturbations. Our approach allows us to characterize the dynamic drug responses of PDOs from both primary and metastatic CRC tumors over time. To accomplish this, we treated PDOs with standard chemotherapy agents: irinotecan, a topoisomerase I inhibitor, and oxaliplatin, an alkylating agent. To interrogate drug specific phenotypic responses, we used heatmaps to examine morphological and textural features within the dead class of PDOs over the course of drug treatment. (**Figure 4A**). Across all PDOs, the feature pattern in the staurosporine-treated group is distinct from the chemotherapy groups. For PDO-12415, the symmetry features in the media control stand out with generally higher feature values compared to the drug treated groups. PDO-12527 and PDO-12911 showed an increase in area and radial mean over time for all treatments, however the symmetry properties of all the PDOs did not show distinct variation across treatment or time. Importantly, the increase in area and radial mean could be attributed to the loss of structure and spreading of dead organoids in response to drug rather than proliferation.

Using our Shiny-based visualization tool we generated box plots of extracted features of interest over time (**Figure 4B** and **Supplemental Figure 6**). We chose to highlight Regional Threshold Compactness 60% due to the unique patterns

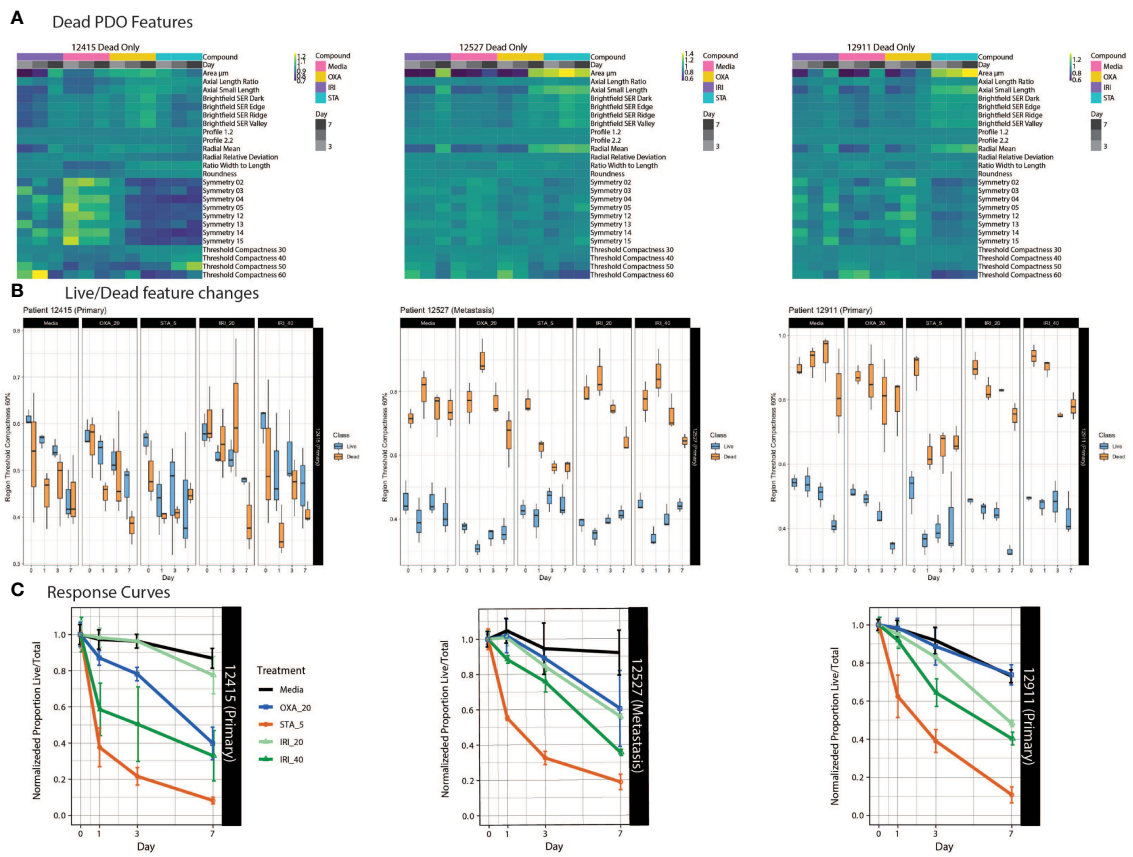


FIGURE 4 | PDO-specific drug responses over time. **(A)** Heat maps of dead PDO features under drug perturbations. Z-score normalized averaged feature values are shown with treatment and time on the x-axis, with features on the y-axis. **(B)** Boxplots of the feature "region threshold compactness 60%" generated using the Organoizer show the variation between the classified live/dead groups. **(C)** Fractions of live/total organoids from untreated media control, irinotecan-treated and oxaliplatin-treated groups are plotted over time to generate dose response curves. Points indicate the mean and bars show the SD.

observed across patients; however, our visualization tool is capable of displaying all features captured. In addition, we plotted dose response curves for each drug treatment (**Figure 4C**). While PDO-12415 (bottom left panel) showed a limited response to irinotecan at 20 μM, a much stronger effect was measured with 40 μM irinotecan starting on day 1. Oxaliplatin at 20 μM also elicited a response with the normalized proportion of live/total dropping 60% by day 7. Analysis of PDO-12527 (bottom middle) revealed a similar response to both oxaliplatin and irinotecan at 20 μM, while irinotecan at 40 μM more effectively killed the PDOs. It appeared that PDO-12911 (bottom right) did not respond to oxaliplatin at 20 μM, as the proportion of live/total PDOs was comparable to the negative control across all 7 days. A slight difference in response timing was seen between the two doses of irinotecan, where the 40 μM dose showed a stronger response at day 3 compared to 20 μM, however by day 7 both showed a response below 40%. Despite temporal variations in response, all three PDOs showed a 60-70% reduction in the proportion of live/total PDOs by day 7. Taken together our ML approach identified PDOs that responded to chemotherapy early in the dosing regimen, highlighting the ability to capture patient-specific drug responses.

DISCUSSION

Given the breadth of biological models used in cancer research, investigations into drug response should span spatial and temporal scales. However, we continue to apply assays such as those measuring ATP-viability that capture a single readout from a sample/well at a fixed timepoint, which results in a limited understanding of the underlying biology. As seen in **Figure 3**, manual classification limits throughput and introduces person-to-person subjectivity. On the other side, VDs proved problematic for determining the viability of 3D organoids - especially once healthy proliferating organoids develop a necrotic core that contains a substantial fraction of dead cells, masking drug effects. This issue is more commonly seen in PDOs that form large structures, which can be the result of rapidly proliferating organoids. Furthermore, dying cells that initially stain positive using VDs, eventually lose their ability to retain the dye and therefore may erroneously be counted as live.

Here we present an object-based image analysis (OBIA) workflow that is designed to fill the gap between cell and population-level analyses, to dynamically interrogate heterogeneous object-based PDOs in response to perturbations including drug treatment. The non-destructive nature of our platform supports temporal monitoring of phenotypic changes, which allows us to capture the appropriate timing of effects. With an OBIA ML approach we can account for variations in inter-PDO samples (size, shape, etc.). With a larger dataset, one could begin to explore possible correlations between organoid features and patient prognosis (28, 29), shedding light on the clinical relevance of these features.

The imaging workflow described herein provides significant advantages; however, it is important to consider the limitations. Imaging consistency plays a large role in the success of a given assay, and deviations in the XY sample placement can influence results when drawing conclusions across timepoints. Patient to patient variability in PDO size influences the parameters needed for proper segmentation and identification of ROIs; therefore, careful consideration needs to be placed on splitting and merging factors during the segmentation algorithm adjustments. Moreover, additional features and/or separate classification algorithms may be needed to accurately separate live and dead PDO categories when treated with diverse classes of drug compounds.

As many biologists are not computer vision or ML experts, analysis platforms that are accessible to non-experts are needed (30, 31). A paper by Falk et al. describes an ImageJ plugin, U-Net, that enables researchers who are not ML experts to benefit from its application to biological data (32). Furthermore, while the computationally intensive parts of the image analysis workflow are done in a reproducible and automated fashion, biologists are still faced with the task of summarizing the data for different timepoints and conditions across thousands of ROIs. To facilitate this step, we have designed an interactive, web-based tool where users can upload the output of the ML analysis and obtain survival curves and feature metrics. Additionally, while the textural and morphological features that best differentiate between live and dead organoids are automatically determined by the linear classifier, it is often useful to be able to visualize differences across all collected features over time. These tools are accessible at http://organoizer.eitm.org and available for download at https://github.com/eitm-org/organoizer.

We focused our attention on the use of a supervised ML linear classifier algorithm to distinguish live versus dead organoids for the purposes of understanding drug response; however, there are many other questions that could be asked using this method. This workflow enables unrestrained exploration of multidimensional features of organoid morphology and texture characteristics to discover new biology within and across patient samples. Here we demonstrate the utility of our ML image analysis method using a smaller sample set; however, this method can be scaled to perform large drug screens on PDOs generated from different cancer types, providing researchers a flexible yet robust platform for posing their own biological questions. Additional artificial intelligence and ML techniques are being applied to image analysis workflows, including unsupervised techniques such as neural networks and deep learning, which recognize outcomes that are not detectable by humans (21–24). Label-free organoid imaging and batch analysis methods using trained neural networks have been developed from several groups (30, 31, 33, 34). Although these approaches provide highly efficient and precise detection, classification, and measurement of organoid objects, these often require programming skills to create a specific code to train the network and process images. Deep learning-based analysis will be very powerful with large datasets, but additional data processing will be needed to extract specific information. Our ML-based method, with linear classifier and data visualization tool, showed great performance with a relatively small patient sample size. Moreover, it generated multiparametric data including patient-specific organoid morphologies and drug responses over time to understand patient heterogeneity.

The rise in patient-derived biobanks, combined with sophisticated image analysis techniques using ML approaches, presents a valuable platform for drug screening and discovery.

ONLINE METHODS

Cell Culture and Reagents

Organoid growth medium consists of base medium (ADMEM/F12 with 10% FBS, 1% penicillin/streptomycin, 1% Glutamax, and 1% HEPES) supplemented with 1X N2 (Sigma Aldrich, 17502048), 1X B-27 (Sigma Aldrich,17504044), 1mM N-Acetylcysteine (Sigma Aldrich, A7250) 50 ng/ml EGF (Life Technologies, PGH 0313), 100 ng/ml Noggin (Tonbo, 21-7075-U500), 10 mM nicotinamide (Sigma, N0636), 500 nM A 83-01 (Calbiochem, 616454-2MG), 10 μM SB202190 (Sigma 47067), and 0.01μM PGE2 (Sigma Aldrich, P5640).

Tissue digestion solution consists of 1.5 mg/ml collagenase (Millipore, 234155), 20 μg/ml hyaluronidase, (MP Biomedicals 100740) and 10 μM Ly27632 (Sigma Y0503).

Generation and Expansion of Human Colorectal Cancer PDOs

Tumor tissue was received from consented patients following Institutional Review Board (IRB) approval at the Norris Comprehensive Cancer Center of USC, Los Angeles CA. Tissue was washed with PBS, minced and digested for 30 minutes at 37°C. Digest suspension was filtered using a 100 μm strainer to remove large residual pieces of tissue, then centrifuged at 189 x g for five minutes. Pellet was washed in DMEM/F12 media (ThermoFisher, 11320033) supplemented with 10% FBS three times and single cells were re-suspended in BME (Culturex® Reduced Growth Factor Basement Membrane Matrix Type 2, Trevigen, 3533-005-02). Cell/BME mixture was plated in 24 well plates with 60 μl per well and incubated upside down at 37°C until solidified (10-20 minutes). Then 500 μl of organoid growth media was layered on top and media was changed as needed. To passage organoids, BME was dissociated with 500 μl/well TrypLE (ThermoFisher Scientific, 12605028) for 5 minutes at 37°C. Organoid suspension was pooled and centrifuged at 450 x g, the pellet was re-suspended in BME and re-plated in a 24 well plate. PDOs used for experiments were ≤ 20 passages in culture.

Drug Treatment Studies

PDOs were harvested from BME using Gentle Cell Dissociation Reagent (Stemcell technologies, 07174), pooling all wells and incubated on ice for 45 minutes then centrifuged for 5 minutes at 189 x g. Supernatant was removed and the pellet was re-suspended in 50% TrypLE with 10 μM Y-27632 (Stemcell Technologies, 72302), incubated at 37°C for 10-15 min with occasional agitation. Alternatively, PDOs were harvested using 500 μl/well TrypLE, incubated at 37°C for 30 minutes. For both methods, PDOs were centrifuged for 5 minutes at 189 x g. Supernatant was removed and the pellet was re-suspended in 1 ml of organoid base medium then filtered through a 40 μm strainer to remove aggregates. Flow through was centrifuged at 189 x g for 5 minutes and the pellet was re-suspended in BME. A 96 well μ-Plate (Ibidi, 89646) was coated with 5 μl of BME/well and incubated at 37°C until BME solidified. The μ-plate was then seeded with 5 μl BME/cell mixture at a concentration of 1000 cells/μl and topped with 70 μl organoid growth medium.

I. Image-Based Organoid Drug Response Assay

Image acquisition. Plates were incubated for four days prior to imaging on the Operetta HCS platform (PerkinElmer). Baseline images were taken on day 0 followed by respective drug treatments: irinotecan (Sigma-Aldrich, I1406), oxaliplatin (Sigma-Aldrich, O9512), staurosporine (Sigma-Aldrich, 569396). Additional images were acquired on days 1, 3, and 7 post drug treatments. Images were acquired in brightfield, with 23 z-stacks ranging from 20-460 μm at increments of 20 μm. On day 3 of the experiment, imaging medium was changed and replaced with fresh medium and drugs.

Image Analysis. Z-stack images were combined into single maximum projection images which were then analyzed using Harmony (PerkinElmer) image analysis software. ROIs were generated using the "Find Texture" supervised ML feature. Training areas of 15 pixels, with texture scaling (2 pixels) were used to define the distance, and region scaling (6 pixels) defines the smoothness of region borders. These ROIs were modified as needed per visual analysis using the "Modify Population" feature to achieve optimal splitting of objects. Specifically, further segmentation was performed to partition the organoid area into multiple, distinct class regions corresponding to individual organoids, by applying a hole-filling algorithm followed by a cluster-by-distance method to detect individual objects within clusters. After objects at the border of the image were removed from the analysis set, morphological and textural features of complete organoids (ROIs) were measured and extracted. A final filtering step based on the object area measurement was applied to exclude small debris as well as large, unsegmented organoid clusters from the data set. In addition, the commercially available PhenoLogic™ ML algorithm (PerkinElmer) was used to classify organoids as live or dead.

II. VD Dead Cell Labeling, Imaging, and Analysis

DRAQ7 (Biolegend, 424001), at a final concentration of 5 μM, was added to plates 30 minutes prior to imaging on day 0. Additional 5 μM DRAQ7 was added on day 3 along with fresh medium and drug. Images were acquired with excitation at 633 nm. Areas positive for DRAQ7 were detected within each organoid ROI. ROIs containing one or more areas of DRAQ7 were classified as dead.

Statistical Analysis

Organoid ROIs were counted, and ROI-level morphological metrics were averaged on a per-well basis at each timepoint and for each class ("dead" vs. "live"). The mean and standard deviation were then computed from replicate wells with the same treatment conditions. Response curves were computed as either the proportion of "live' ROIs over "dead" ROIs (ratio) or as "live" ROIs over all ROIs (proportion). Optionally, the proportion or ratio of "live" organoids can be normalized to the proportion (or

ratio, respectively) on the first day of measurement (usually day 0). Boxplots for each feature across timepoints were also generated. The code is available at https://github.com/eitm-org/organoid_drug_response.

Heatmaps were generated by averaging feature values per well, then taking the average value across wells to get one average value for each unique group. Rows and columns were grouped using hierarchical clustering and rows were scaled using the heatmap package in the R statistical computing language. All analyses were performed using the R statistical language (v. 4.1.0) using the following packages: cowplot (v.1.1.1) (35), eulerr (v. 6.1.1) (36), ggridges (v. 0.5.3) (37), ggthemes (v. 4.2.4), here (v. 1.0.1), irr (v. 0.84.1) (38), knitr (v. 1.33), networkD3 (v. 0.4), pheatmap (v. 1.0.12), plater (v. 1.0.3), readxl (v. 1.3.1), reshape2 (v. 1.4.4), scales (v. 1.1.1), tidyverse (v. 1.3.1) and viridis (v. 0.6.1) (27, 39–49).

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://github.com/eitm-org/organoid_drug_response.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board at the University of Southern California Norris Comprehensive Cancer Center. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ERS performed and analyzed the experiments and wrote the manuscript. NU and EFJR contributed to data analysis and visualization. SK, KP, and BC provided manuscript feedback and assistance with experimental design. CD contributed to image analysis. RL processed patient samples. CS and SC performed preliminary experiments. HJL offered clinical insights and patient samples. NM was responsible for data analysis, visualization, experimental design, and editing of the manuscript. SMM was responsible for study concept and design, interpretation of data, supervision of the study, and editing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by a generous donation made by the Stephenson family, who supported research as part of the Stephenson Family Personalized Medicine Center at the Ellison Institute for Transformative Medicine. It was also funded by a SPRC pilot project awarded through the Ellison Institute for Transformative Medicine.

ACKNOWLEDGMENTS

Thank you to the Stephenson family: Emmet, Toni and Tessa, for their donation of the Operetta HCS, and Operetta CLS platforms and for the establishment of the Stephenson Family Personalized Medicine Center. We would like to express our appreciation to Colin Flinders, PhD for his help with the PDO biobank. We are grateful for the experts who took the time to manually classify PDOs for **Figure 3** and **Supplemental Figure 4**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.771173/full#supplementary-material

REFERENCES

1. Nierode G, Kwon PS, Dordick JS, Kwon SJ. Cell-Based Assay Design for High-Content Screening of Drug Candidates. *J Microbiol Biotechnol* (2016) 26 (2):213–25. doi: 10.4014/jmb.1508.08007
2. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med* (2016) 176(12):1826–33. doi: 10.1001/jamainternmed.2016.6008
3. Wong CH, Siah KW, Lo AW. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* (2019) 20(2):273–86. doi: 10.1093/biostatistics/kxx069
4. Riedl A, Schlederer M, Pudelko K, Stadler M, Walter S, Unterleuthner D, et al. Comparison of Cancer Cells in 2D vs 3D Culture Reveals Differences in AKT-mTOR-S6K Signaling and Drug Responses. *J Cell Sci* (2017) 130(1):203–18. doi: 10.1242/jcs.188102
5. Langhans SA. Three-Dimensional in Vitro Cell Culture Models in Drug Discovery and Drug Repositioning. *Front Pharmacol* (2018) 9. doi: 10.3389/fphar.2018.00006
6. Antoni D, Burckel H, Josset E, Noel G. Three-Dimensional Cell Culture: A Breakthrough in Vivo. *Int J Mol Sci* (2015) 16(3):5517–27. doi: 10.3390/ijms16035517
7. Sato T, Stange DE, Ferrante M, Vries RGJ, van Es JH, van den Brink S, et al. Long-Term Expansion of Epithelial Organoids From Human Colon, Adenoma, Adenocarcinoma, and Barrett's Epithelium. *Gastroenterology* (2011) 141(5):1762–72. doi: 10.1053/j.gastro.2011.07.050
8. Rossi G, Manfrin A, Lutolf MP. Progress and Potential in Organoid Research. *Nat Rev Genet* (2018) 19(11):671–87. doi: 10.1038/s41576-018-0051-9
9. Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernandez-Mateos J, Khan K, et al. Patient-Derived Organoids Model Treatment Response of Metastatic Gastrointestinal Cancers. *Science* (2018) 359(6378):920–6. doi: 10.1126/science.aao2774
10. Fatehullah A, Tan SH, Barker N. Organoids as an *In Vitro* Model of Human Development and Disease. *Nat Cell Biol* (2016) 18(3):246–54. doi: 10.1038/ncb3312
11. van de Wetering M, Francies HE, Francis JM, Bounova G, Iorio F, Pronk A, et al. Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. *Cell* (2015) 161(4):933–45. doi: 10.1016/j.cell.2015.03.053

12. Gao D, Vela I, Sboner A, Iaquinta Phillip J, Karthaus Wouter R, Gopalan A, et al. Organoid Cultures Derived From Patients With Advanced Prostate Cancer. *Cell* (2014) 159(1):176–87. doi: 10.1016/j.cell.2014.08.016
13. Jabs J, Zickgraf FM, Park J, Wagner S, Jiang XQ, Jechow K, et al. Screening Drug Effects in Patient-Derived Cancer Cells Links Organoid Responses to Genome Alterations. *Mol Syst Biol* (2017) 13(11):955. doi: 10.15252/msb.20177697
14. Boj SF, Hwang C-I, Baker LA, Chio IIC, Engle DD, Corbo V, et al. Organoid Models of Human and Mouse Ductal Pancreatic Cancer. *Cell* (2015) 160(1-2):324. doi: 10.1016/j.cell.2014.12.021
15. Huang L, Holtzinger A, Jagan I, BeGora M, Lohse I, Ngai N, et al. Ductal Pancreatic Cancer Modeling and Drug Screening Using Human Pluripotent Stem Cell- and Patient-Derived Tumor Organoids. *Nat Med* (2015) 21 (11):1364–71. doi: 10.1038/nm.3973
16. Phan N, Hong JJ, Tofig B, Mapua M, Elashoff D, Moatamed NA, et al. A Simple High-Throughput Approach Identifies Actionable Drug Sensitivities in Patient-Derived Tumor Organoids. *Commun Biol* (2019) 2:78. doi: 10.1038/s42003-019-0305-x
17. Hill SJ, Decker B, Roberts EA, Horowitz NS, Muto MG, Worley sMJ Jr, et al. Prediction of DNA Repair Inhibitor Response in Short-Term Patient-Derived Ovarian Cancer Organoids. *Cancer Discov* (2018) 8(11):1404–21. doi: 10.1158/2159-8290.CD-18-0474
18. Edmondson R, Adcock AF, Yang L. Influence of Matrices on 3D-Cultured Prostate Cancer Cells' Drug Response and Expression of Drug-Action Associated Proteins.(Report). *PloS One* (2016) 11(6):e0158116. doi: 10.1371/journal.pone.0158116
19. Akagi J, Kordon M, Zhao H, Matuszek A, Dobrucki J, Errington R, et al. Real-Time Cell Viability Assays Using a New Anthracycline Derivative DRAQ7®. *Cytometry A* (2013) 83(2):227–34. doi: 10.1002/cyto.a.22228
20. Berg EL, Hsu YC, Lee JA. Consideration of the Cellular Microenvironment: Physiologically Relevant Co-Culture Systems in Drug Discovery. *Adv Drug Deliv Rev* (2014) 69:190–204. doi: 10.1016/j.addr.2014.01.013
21. Rojas-Moraleda R, Xiong W, Halama N, Breitkopf-Heinlein K, Dooley S, Salinas L, et al. Robust Detection and Segmentation of Cell Nuclei in Biomedical Images Based on a Computational Topology Framework. *Med Image Anal* (2017) 38:90–103. doi: 10.1016/j.media.2017.02.009
22. Rahman MM, Antani SK, Thoma GR. A Learning-Based Similarity Fusion and Filtering Approach for Biomedical Image Retrieval Using SVM Classification and Relevance Feedback. *IEEE Trans Inf Technol BioMed* (2011) 15(4):640–6. doi: 10.1109/TITB.2011.2151258
23. Kraus OZ, Frey BJ. Computer Vision for High Content Screening. *Crit Rev Biochem Mol Biol* (2016) 51:102–9. doi: 10.3109/10409238.2015.1135868
24. Wainberg M, Merico D, Delong A, Frey BJ. Deep Learning in Biomedicine. *Nat Biotechnol* (2018) 36(9):829–38. doi: 10.1038/nbt.4233
25. O'Duibhir E, Paris J, Lawson H, Sepulveda C, Shenton DD, Carragher NO, et al. Machine Learning Enables Live Label-Free Phenotypic Screening in Three Dimensions. *Assay Drug Dev Technol* (2018) 16(1):51–63. doi: 10.1089/adt.2017.819
26. Garvey CM, Spiller E, Lindsay D, Chiang CT, Choi NC, Agus DB, et al. A High-Content Image-Based Method for Quantitatively Studying Context-Dependent Cell Population Dynamics. *Sci Rep* (2016) 6:12. doi: 10.1038/srep29752
27. Chang W, Cheng JJ, Sievert C, Schloerke B, Xie Y, et al. *Shiny: Web Application Framework for R*. (2021). Available at: https://CRAN.R-project.org/package=shiny.
28. Wensink GE, Elias SG, Mullenders J, Koopman M, Boj SF, Kranenburg OW, et al. Patient-Derived Organoids as a Predictive Biomarker for Treatment Response in Cancer Patients. *NPJ Precis Oncol* (2021) 5(1):30. doi: 10.1038/s41698-021-00168-1
29. Verduin M, Hoeben A, De Ruysscher D, Vooijs M. Patient-Derived Cancer Organoids as Predictors of Treatment Response. *Front Oncol* (2021) 11:641980. doi: 10.3389/fonc.2021.641980
30. Borten MA, Bajikar SS, Sasaki N, Clevers H, Janes KA. Automated Brightfield Morphometry of 3D Organoid Populations by OrganoSeg. *Sci Rep* (2018) 8 (1):5319. doi: 10.1038/s41598-017-18815-8
31. Kassis T, Hernandez-Gordillo V, Langer R, Griffith LG. OrgaQuant: Human Intestinal Organoid Localization and Quantification Using Deep Convolutional Neural Networks. *Sci Rep* (2019) 9(1):12479. doi: 10.1038/s41598-019-48874-y
32. Falk T, Mai D, Bensch R, Çiçek ÖN, Abdulkadir A, Marrakchi Y, et al. U-Net: Deep Learning for Cell Counting, Detection and Morphometry. *Nat Methods* (2018) 16:67–70. doi: 10.1038/s41592-018-0261-2
33. Gritti N, Lim JL, Anlas K, Pandya M, Aalderink G, Martinez-Ara G, et al. MOrgAna: Accessible Quantitative Analysis of Organoids With Machine Learning. *Development* (2021) 148(18):dev199611. doi: 10.1242/dev.199611
34. Bian X, Li G, Wang C, Liu W, Lin X, Chen Z, et al. A Deep Learning Model for Detection and Tracking in High-Throughput Images of Organoid. *Comput Biol Med* (2021) 134:104490. doi: 10.1016/j.compbiomed.2021.104490
35. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2021). Available at: https://www.R-project.org/.
36. Wilke CO. *Cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* (2020). Available at: https://CRAN.R-project.org/package=cowplot.
37. Larsson J. *Eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses* (2021). Available at: https://CRAN.R-project.org/package=eulerr.
38. Wilke CO. *ggridges: Ridgeline Plots in "ggplot2."*. (2021). Available at: https://CRAN.R-project.org/package=ggridges.
39. Arnold JB. *ggthemes: Extra Themes, Scales and Geoms for "ggplot2."*. (2021). Available at: https://CRAN.R-project.org/package=ggthemes.
40. Müller K. *Here: A Simpler Way to Find Your Files*. (2020). Available at: https://CRAN.R-project.org/package=here.
41. Gamer M, Lemon J, Singh IFP. *irr: Various Coefficients of Interrater Reliability and Agreement*. (2019). Available at: https://CRAN.R-project.org/package=irr.
42. Xie Y. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. (2021). Available at: https://yihui.org/knitr/.
43. Allaire JJ, Gandrud C, Russell K, Yetman CJ. *NetworkD3: D3 JavaScript Network Graphs From R*. (2017). Available at: https://CRAN.R-project.org/package=networkD3.
44. Kolde R. *Pheatmap: Pretty Heatmaps*. (2019). Available at: https://CRAN.R-project.org/package=pheatmap.
45. Hughes SM. Plater: Read, Tidy, and Display Data From Microtiter Plates. *J Open Source Softw* (2016) 1(7):106. doi: 10.21105/joss.00106
46. Wickham H, Bryan J. *readxl: Read Excel Files*. (2019). Available at: https://CRAN.R-project.org/package=readxl.
47. Wickham H. Reshaping Data With the Reshape Package. *J Stat Softw* (2007). 21(12):1–20. doi: 10.18637/jss.v021.i12
48. Wickham H, Seidel D. *Scales: Scale Functions for Visualization*. (2020). Available at: https://CRAN.R-project.org/package=scales.
49. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. (2019) 4:1686. doi: 10.21105/joss.01686
50. Garnier S, Ross N, Rudis R, Camargo PA, Sciaini M, Scherer C. *Viridis - Colorblind-Friendly Color Maps for R*. (2021). doi: 10.5281/zenodo.4679424

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 Oncology

MINI REVIEW
published: 21 January 2022
doi: 10.3389/fonc.2021.805592

Capturing Biomarkers and Molecular Targets in Cellular Landscapes From Dynamic Reaction Network Models and Machine Learning

Susan D. Mertins [1,2,3*†]

[1] Department of Science, Mount St. Mary's University, Emmitsburg, MD, United States, [2] Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Limited Liability Company (LLC), Frederick, MD, United States, [3] BioSystems Strategies, Limited Liability Company (LLC), Frederick, MD, United States

Computational dynamic ODE models of cell function describing biochemical reactions have been created for decades, but on a small scale. Still, they have been highly effective in describing and predicting behaviors. For example, oscillatory phospho-ERK levels were predicted and confirmed in MAPK signaling encompassing both positive and negative feedback loops. These models typically were limited and not adapted to large datasets so commonly found today. But importantly, ODE models describe reaction networks in well-mixed systems representing the cell and can be simulated with ordinary differential equations that are solved deterministically. Stochastic solutions, which can account for noisy reaction networks, in some cases, also improve predictions. Today, dynamic ODE models rarely encompass an entire cell even though it might be expected that an upload of the large genomic, transcriptomic, and proteomic datasets may allow whole cell models. It is proposed here to combine output from simulated dynamic ODE models, completed with omics data, to discover both biomarkers in cancer *a priori* and molecular targets in the Machine Learning setting.

Keywords: biomarkers, molecular targets, drug discovery, drug development, pharmacodynamic modeling, ODE modeling, machine learning

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Mohit Kumar Jolly, Indian Institute of Science (IISc), India

***Correspondence:**
Susan D. Mertins
smertins@biosystemsstrategies.com

***Present address:**
Susan D. Mertins, BioSystems Strategies, LLC, Frederick, MD, United States

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *30 October 2021*
Accepted: *31 December 2021*
Published: *21 January 2022*

Citation:
Mertins SD (2022) Capturing Biomarkers and Molecular Targets in Cellular Landscapes From Dynamic Reaction Network Models and Machine Learning. Front. Oncol. 11:805592. doi: 10.3389/fonc.2021.805592

INTRODUCTION

Understanding biological systems is challenging as the detail and complexity of such dynamic entities cannot be grasped using human ken and intuition. And disease states such as cancer further the difficulties in addressing living entities. Thus, investigators have created networks of cellular systems that encompass many components, connect those parts in some fashion, and then interrogate their usefulness for addressing predictability and/or to study the networks themselves (1). This has been well studied in the case of transcription factor networks that characterize the phenomenon of epithelial to mesenchymal transition (EMT) in cancer (2–6). Recent work demonstrated that the topology of transcription factor regulatory networks that were parameter free (using a Boolean approach) or were parameter agnostic (using a random parameter generator) was important in limiting changes in cell state that may promote disease progression (3).

While important findings can be gleaned from understanding regulation of gene expression, the effectors are not included in such modeling nor are site-specific details such as binding affinity or catalytic activities such as phosphorylation that are likely to influence a particular cell behavior. Therefore, the remainder of the Mini Review will focus on the modeling approaches that include the depth and likely parameters that may improve useful predictions.

Modeling cell behaviors using the mathematics underpinning biochemical reactions has been a research topic for decades when first approached by Tyson and others in understanding the cell cycle in the 1980s (7). It was and still is clear that cellular networks and pathways are dynamic and the overall contribution to cell behavior mattered. Since then, a vast array of biological models described by ordinary and partial differential equations (ODE, PDE) have been published, new software has been created to entice bench scientists to advance their findings computationally [for example, RuleBender (8) and Virtual Cell (9)], and now, an unprecedented amount of data to supply those models is available. For the oncology field, melding both mathematical models and omics data holds the promise to select the most effective molecular targets and any concurrent biomarkers. The future promise consists of whole cell models that lead to decisions of personalized therapies needed to predict tumor regression.

At the cellular level, mathematical models describe reaction networks (e.g., signal transduction pathways and metabolic cycles) effectively, typically with output that is difficult to predict. For example, binding of a growth factor to its receptor is a reaction that is dependent on characteristics such as binding affinity and concentration and leads to downstream events of interest occurring with time. Further, if positive and/or negative feedback loops are considered, cellular function becomes less predictable by inspection, if at all. But, operationally, in this example, the change in concentration of bound and unbound receptors and growth factor with respect to time can be tracked through species in the ODEs and downstream effects delineated. Furthermore, properties such as molecular diffusion can be described with PDEs, thus including spatial considerations as well. Many other biological properties have been described and include cytoskeleton formation (10), vesicular transport (11), and gene expression (12). Hundreds of such mathematical models have been deposited in the BioModels database (13) (RIDD: SCR_001993).

It is the premise of this Mini Review to describe how computational dynamic ODE models describing biochemical reaction networks can be analyzed by Machine Learning (ML) algorithms capable of predicting desirable outcomes for the two of the present challenges in oncology: discovering biomarkers associated with positive patient outcomes and novel molecular targets not generally considered druggable.

SELECTED ODE MODELS

As experimentalists have supplied an understanding of the essential knowledge of biochemical pathways underscoring cell behavior, ODE models offered predictions on some of the more interesting aspects. For example, stimulus response, in general, was thought to occur in a linear fashion such that the greater concentration of a growth factor, the greater the response. However, some components in pathways appeared not to follow this linear rule. Rather, an all-or-none response occurred. ODE modeling supplied a mechanism for such ultrasensitivity, demonstrating a phenomenon that occurs when proteins with enzymatic function are acting at saturation (14, 15). Other important discoveries ensued and included uncovering bistability, a behavior dependent on initial conditions. Bistability is defined as having two stable states at one stimulus level (16, 17). Parenthetically, bistability has also been defined to describe whole cell states such as those found in EMT and is distinct from that which is noted here (3–5). Functionally, bistability at the biochemical reaction level is of interest because, dependent on conditions such as saturation of an enzyme and high initial stimulus, cellular response can resemble a toggle switch in the on position. Thus, immediate downregulation may not be needed in these instances (16, 18). Finally, oscillatory behavior of important transcription factors such p53 and their regulators have been demonstrated through ODE modeling and in cells (19, 20). In addition, oscillatory levels of phosphorylated kinases have been characterized by an amplitude and frequency and were shown to be regulated and define outcomes such as the decision to proliferate (21). Below are two relevant examples of ODE models that can be exploited to discover biomarkers and molecular targets in oncology.

MAPK Signaling

One of the most well-studied signal transduction pathways in oncology is that which triggers cell proliferation *via* growth factor stimulation. In one such reaction network, EGF, a growth factor, binds to its cognate receptor, which in turn, dimerizes and activates its kinase domain through conformational adjustments. Trans-autophosphorylation occurs next which forms the initial sites for adapter binding. GRB2 binding through its SH2 domain to phosphorylated tyrosine then binds SOS, a GTPase exchange factor. Critically, SOS replaces GDP with GTP on RAS, thus activating it for downstream binding of RAF, a kinase. A kinase cascade ensues ultimately leading to a phosphorylated kinase (phospho-ERK1/2) capable of triggering gene transcription necessary for cell growth. Importantly, several positive and negative feedback loops regulate the pathway and when considered in an ODE model, oscillatory behavior of phosphorylated ERK1/2 protein level occurs. While this description adequately corresponds to non-oncogenic signaling, disruption in the reaction rates by mutations leads to unpredictable outcomes dependent on pathway protein levels (22, 23). **Figure 1** (upper portion) depicts a limited Contact Map of MAPK signaling.

Cell Cycle Arrest or Apoptosis Decision

Cell decisions regulated by p53 have been readily modeled as it is of deep interest to determine the conditions in which the outcome of cell cycle arrest or apoptosis occurs following genomic insult. This is of particular importance since p53 is

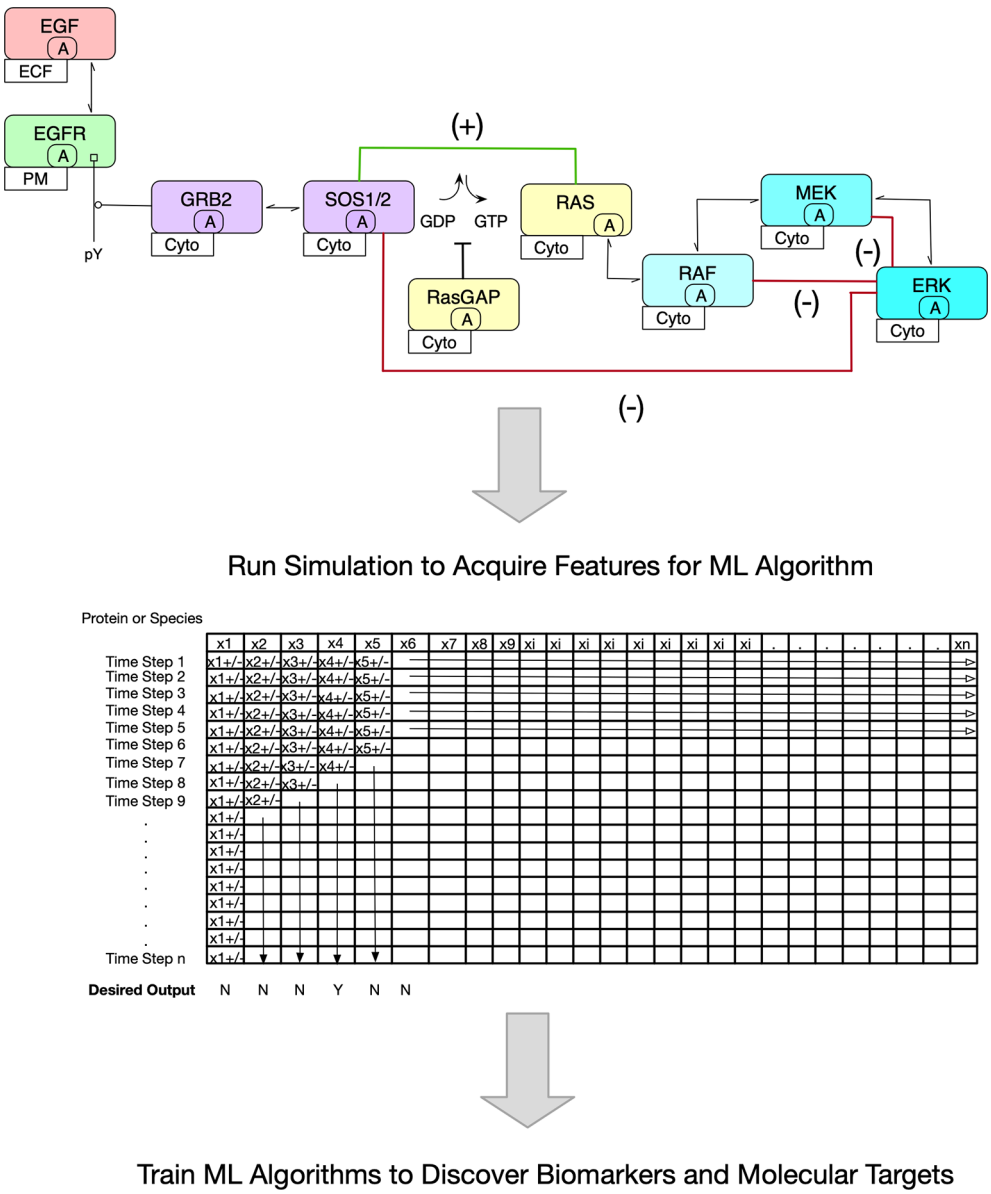


FIGURE 1 | Overview of Proposed Methodology. The premise of this Mini Review is pictured. In the upper portion, a stylized version of the MAPK signaling pathway is shown. The contact map shown is the basis of an ODE Model. Note the inclusion of positive (green line) and negative (red lines) feedback loops. Because of these regulatory networks, it is inherently difficult to predict outcomes such as the decision to proliferate. Further, it is more challenging to ascertain pharmacologic interventions. In the lower portion, an abstract matrix is shown that depicts the changing protein concentrations with time once an ODE model is simulated. It is these data that are useful to train ML algorithms to discover biomarkers and novel molecular targets.

commonly mutated in cancer and underpins tumorigenesis. Lipniacki's group developed a sophisticated model wherein irradiation induced DNA damage, triggering p53 transcription and subsequent downstream reactions (24, 25). The p53 network was created using the rule-based modeling technique that defines biochemical interactions such as protein complex formation and enzymatic activities. In the three-stage model, a steady state was achieved first through simulation and solutions through deterministic algorithms. Next, irradiation was modeled over a range (1-10 Gy). In the last relaxation phase, cell behavior was characterized. Notably, the predicted outcomes were contingent on the degree of irradiation. Oscillatory behavior of p53 levels was observed as has been demonstrated *in vitro* and was found to define the conditions that lead to programmed cell death. Finally, their findings provided support for the heterogeneity in response to DNA damage observed in tumor cell lines. [N.B. For an overview of the methodology of ODE modeling in apoptosis, see (26)].

PROTEOMIC DATABASES

One of the challenges of ODE modeling is the need to insert quantitative parameters such as protein concentration in the reaction network. Parenthetically, precise measurements of reaction rates, enzymatic activity, and binding affinities are also critically important in computational modeling. The advent of critical technologies that can determine such include labeled and label-free assays such and SILAC and LC-MS/MS, respectively (27, 28). Both tumor cell lines and tumor tissue have been studied in this fashion and databases exist to exploit (29–32). And these quantitative proteomics efforts can evaluate post-translational modifications as well (33). The direct analysis of datasets from these studies has identified biomarkers, therapeutic targets, and drug resistance mechanisms. Further, and importantly, direct interaction networks can be constructed, but offer a static interpretation of a dynamic living cell.

The National Cancer Institute's Office of Cancer Clinical Proteomics developed one such database that contains proteomic evaluations from 13 different tumor types collected since 2006 (31, 32, 34) (RIDD: SCR_017135). A series of studies merged both proteomic and genomic data to better classify actionable mutations that are expressed as proteins with certainty rather than through sequence alone. This is even more critical since transcriptomics does not necessarily confirm translation in as high as 50% of all transcripts measured. But importantly, studies utilizing this database have provided important new classifications in cancer histologies. One study evaluated the proteogenomics of pediatric brain tumors and found new subgroups with wild type BRAF and novel networks that overlap with the mutant gene (35). Thus, new therapeutic trials could be proposed for these challenging tumor types.

SILAC methods have been melded with mathematical modeling in the study of dynamic systems. For example, Yilmaz et al. investigated proteosomal processing of NFκB subunits in mouse embryonic fibroblasts *via* labeling studies and mathematical modeling to discover the dynamics of the system under activation (36). In another study, CHO cell extracts were processed in a glycoproteomic approach to understand N-glycan processing and found a kinetic description of the pathway (37). Finally, global protein synthesis rates were studied by pulse labeling and compared to mRNA synthesis rates in mathematical models, thus, providing critical quantitative parameters useful for future studies as well (38). In summary, SILAC and LC-MS/MS methods have supported mathematical modeling and hold further promise.

MACHINE LEARNING

Applying ML algorithms to uncover cancer diagnoses from histologies, to determine therapeutic decisions, and predict outcomes is becoming pervasive in light of omics studies (39, 40). ML, in an overview, can sort through vast inputs and discover connections not likely to be found by human inspection. An alternative application for ML in oncology could include the development of hypotheses subject to future study. Thus, ML is suited to intake vast inputs and to realize relationships not previously expected. It is the thesis of this Mini Review to offer ways to meld ODE modeling and ML to discover biomarkers and molecular targets, ultimately aiding drug discovery and predictive oncology.

Biomarker and Molecular Target Discovery

It is instructive to describe a published ODE model in more detail for EGFR signaling encompassing the MAPK pathway (**Figure 1**) to demonstrate how it might be utilized in ML with existing proteomic databases. In two models published by Creamer and colleagues (41) and Kochanczyk and colleagues (22), the basic signaling pathway is described for EGF binding to the family of cognate receptor tyrosine kinases with subsequent dimerization triggering autophosphorylation on their intracellular tails at multiple sites. Next, adaptors bind with some affinity *via* SH2 or PTB domains. In canonical MAPK signaling, bound GRB2 anchors SOS1/2, a GTPase exchange protein. During activation, SOS1/2 binds RAS and exchanges GDP for GTP, and so activates RAS. A kinase cascade follows RAF1 dimerization and binding to RAS. RAF1 phosphorylates MEK1/2 and it, in turn, phosphorylates ERK1/2 which translocates to the nucleus to modulate transcription regulating essential genes for cell proliferation. The ODE model of EGFR signaling as described here includes reaction rates that underlie binding affinities, turnover rates, phosphorylation and dephosphorylation rates, and both positive and negative feedback loops (22). Including these important regulatory features (the loops) is critical since unexpected behaviors emerge such oscillatory ERK1/2 phosphorylation levels and multiple states that include a steady state, a monostable one, or a bistable one dependent on the mathematical nature of the feedback loops. As an aside, even at this basic understanding, choosing a molecular target would be challenging.

Establishing the reaction network such as the one above by ODE equations can be readily accomplished with rule-based modeling using BioNetGen, a programming language and software that can further simulate the model over time (8). The simulations can be deterministically or stochastically solved with well accepted algorithms. Deterministic solutions are reproducible, resulting in the same output with every simulation. In contrast, stochastic simulations apply randomness to the solution, hence output is variable within a certain distribution and can be averaged. However, it is important to note that biological systems are noisy and the latter solutions may be more relevant. Simple output includes changes in species (molecule) abundance at each time step modeled. In the end, a vast amount of data is generated and fairly complex models with thousands of reactions can be coded (42). **Figure 1** (lower portion) pictorially describes a matrix of protein concentrations (in the columns) that change at each time step during a deterministic simulation (in the rows).

Parameterization of ODE models is challenging and typically, the scientific literature can provide many (43). It is anticipated that the proteomic databases described herein will readily provide protein abundance for ODE models and thus, may reflect both normal and disease states in separate models. Databases such as BRENDA (44, 45) (RIDD: SCR_002997) and Binding Database (46) contain curated reaction rates and affinities culled from the published literature. Thus, with a completed ODE model of interest, experimental protein abundance and measured parameters underlie their usefulness.

For ML algorithms, each time step can be represented by copy number (abundance) for an individual protein and thus, become features. What would be of interest for both biomarker and molecular target discovery is which state (i.e., time point and copy numbers) would be predictive of a desired outcome such as inhibition of cell proliferation and/or induction of cell death through one or more of the many known mechanisms. In order to achieve this, a training set would be needed that describes a signal transduction pathway or pathways anticipated to be central to cancer pathogenesis and clinically relevant. For example, the p53 ODE model such as the one described above intersects a critical cellular decision in light of DNA damage, determining cell cycle arrest or programmed cell death. The proteins in this training set model would be derived from tumor cell lines or tissues. Next a simulation would be completed resulting in a matrix of protein abundance at each time step (**Figure 1**). Now, the investigator has hundreds, if not thousands of models with and without the desired outcome that act as the training dataset for ML. The first analysis of such a ML model would be to find biomarkers (i.e., abundances of particular proteins) that correlate with the predicted outcome.

An alternative approach utilizing the same ODE model can aid in the discovery of novel molecular targets. In this case, proteins can be knocked out virtually (individually) through simple programming, a simulation run for each knock-out, and outcome collected. The ML algorithm would identify the connection between the presumptive molecular target and programmed cell death in this instance. Thus, a virtual high throughput screen has been completed using only computational effort.

CONCLUSION

It has been proposed in this Mini Review to apply ML algorithms to discover biomarkers and molecular targets through the creation of ODE models of signaling pathways in cancer (47). While ODE modeling is more labor intensive than the ML analysis, the complex systems of cancer cell biology can be studied in this novel way leading to knowledge that will be readily apply to the pharmaceutical challenges ahead. In addition, ongoing advances in ODE modeling combined with tissue level simulations also show promise. For example, mathematical models of metabolism and cell proliferation indicated new molecular targets (48) and such multicellular models can further predict outcomes such as necrosis and growth arrest (49), cancer cell migration (50), and immune cell invasion (51). It can be envisioned that the computational efforts described herein can contribute to proposed Digital Twins for personalized medicine in cancer (52).

AUTHOR CONTRIBUTIONS

SM wrote the manuscript, completed the revisions, and approved of the submitted and revised version.

FUNDING

SM was funded by Mount St. Mary's University as Assistant Professor and BioSystems Strategies, LLC as Founder and CEO.

ACKNOWLEDGMENTS

The author wishes to acknowledge the expert advice from Dr. M. Kathy Jung in editing and revising the manuscript. In addition, the author thanks the generous support from Dr. Eric Stahlberg whose insights underpin the ideas set forth herein.

REFERENCES

1. Kauffman S. Gene Regulation Networks: A Theory For Their Global Structure and Behaviors. *Curr Top Dev Biol* (1971) 6:145–82. doi: 10.1016/S0070-2153(08)60640-7
2. Steinway SN, Zañudo JGT, Michel PJ, Feith DJ, Loughran TP, Albert R. Combinatorial Interventions Inhibit Tgfβ-Driven Epithelial-to-Mesenchymal Transition and Support Hybrid Cellular Phenotypes. *NPJ Syst Biol Appl* (2015) 1:15014. doi: 10.1038/npjsba.2015.14
3. Hari K, Sabuwala B, Subramani BV, La Porta CAM, Zapperi S, Font-Clos F, et al. Identifying Inhibitors of Epithelial-Mesenchymal Plasticity Using a Network Topology-Based Approach. *NPJ Syst Biol Appl* (2020) 6(1):15. doi: 10.1038/s41540-020-0132-1
4. Pillai M, Jolly MK. Systems-Level Network Modeling Deciphers the Master Regulators of Phenotypic Plasticity and Heterogeneity in Melanoma. *iScience* (2021) 24(10):103111. doi: 10.1016/j.isci.2021.103111
5. Udyavar AR, Wooten DJ, Hoeksema M, Bansal M, Califano A, Estrada L, et al. Novel Hybrid Phenotype Revealed in Small Cell Lung Cancer by a Transcription Factor Network Model That Can Explain Tumor Heterogeneity. *Cancer Res* (2017) 77(5):1063–74. doi: 10.1158/0008-5472.CAN-16-1467
6. Lang J, Nie Q, Li C. Landscape and Kinetic Path Quantify Critical Transitions in Epithelial-Mesenchymal Transition. *Biophys J* (2021) 120(20):4484–500. doi: 10.1016/j.bpj.2021.08.043
7. Tyson JJ, Hannsgen KB. Cell Growth and Division: A Deterministic/Probabilistic Model of the Cell Cycle. *J Math Biol* (1986) 23(2):231–46. doi: 10.1007/BF00276959
8. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS. BioNetGen: Software for Rule-Based Modeling of Signal Transduction Based on the Interactions of Molecular Domains. *Bioinformatics* (2004) 20(17):3289–91. doi: 10.1093/bioinformatics/bth378
9. Moraru II, Schaff JC, Slepchenko BM, Blinov ML, Morgan F, Lakshminarayana A, et al. Virtual Cell Modelling and Simulation Software Environment. *IET Syst Biol* (2008) 2(5):352–62. doi: 10.1049/iet-syb:20080102
10. Novak IL, Slepchenko BM, Mogilner A, Loew LM. Cooperativity Between Cell Contractility and Adhesion. *Phys Rev Lett* (2004) 93(26 Pt 1):268109. doi: 10.1103/PhysRevLett.93.268109
11. Shea SM, Karnovsky MJ, Bossert WH. Vesicular Transport Across Endothelium: Simulation of a Diffusion Model. *J Theor Biol* (1969) 24(1):30–42. doi: 10.1016/S0022-5193(69)80004-4
12. Larson DR, Singer RH, Zenklusen D. A Single Molecule View of Gene Expression. *Trends Cell Biol* (2009) 19(11):630–7. doi: 10.1016/j.tcb.2009.08.008
13. Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, et al. BioModels-15 Years of Sharing Computational Models in Life Science. *Nucleic Acids Res* (2020) 48(D1):D407–15. doi: 10.1093/nar/gkz1055
14. Goldbeter A, Koshland DE. An Amplified Sensitivity Arising From Covalent Modification in Biological Systems. *Proc Natl Acad Sci USA* (1981) 78(11):6840–4. doi: 10.1073/pnas.78.11.6840

15. Kim SY, Ferrell JE. Substrate Competition as a Source of Ultrasensitivity in the Inactivation of Wee1. *Cell* (2007) 128(6):1133–45. doi: 10.1016/j.cell.2007.01.039
16. Pomerening JR, Sontag ED, Ferrell JE. Building a Cell Cycle Oscillator: Hysteresis and Bistability in the Activation of Cdc2. *Nat Cell Biol* (2003) 5 (4):346–51. doi: 10.1038/ncb954
17. Ferrell JE, Ha SH. Ultrasensitivity Part III: Cascades, Bistable Switches, and Oscillators. *Trends Biochem Sci* (2014) 39(12):612–8. doi: 10.1016/j.tibs.2014.10.002
18. Trunnell NB, Poon AC, Kim SY, Ferrell JE. Ultrasensitivity in the Regulation of Cdc25C by Cdk1. *Mol Cell* (2011) 41(3):263–74. doi: 10.1016/j.molcel.2011.01.012
19. Ma L, Wagner J, Rice JJ, Hu W, Levine AJ, Stolovitzky GA. A Plausible Model for the Digital Response of P53 to DNA Damage. *Proc Natl Acad Sci USA* (2005) 102(40):14266–71. doi: 10.1073/pnas.0501352102
20. Hat B, Kochańczyk M, Bogdał MN, Lipniacki T. Feedbacks, Bifurcations, and Cell Fate Decision-Making in the P53 System. *PLoS Comput Biol* (2016) 12(2): e1004787. doi: 10.1371/journal.pcbi.1004787
21. Albeck JG, Mills GB, Brugge JS. Frequency-Modulated Pulses of ERK Activity Transmit Quantitative Proliferation Signals. *Mol Cell* (2013) 49(2):249–61. doi: 10.1016/j.molcel.2012.11.002
22. Kochańczyk M, Kocieniewski P, Kozłowska E, Jaruszewicz-Błońska J, Sparta B, Pargett M, et al. Relaxation Oscillations and Hierarchy of Feedbacks in MAPK Signaling. *Sci Rep* (2017) 7:38244–59. doi: 10.1038/srep38244
23. Stites EC, Shaw AS. Quantitative Systems Pharmacology Analysis of KRAS G12C Covalent Inhibitors. *CPT Pharmacometrics Syst Pharmacol* (2018) 7 (5):342–51. doi: 10.1002/psp4.12291
24. Bogdał MN, Hat B, Kochańczyk M, Lipniacki T. Levels of Pro-Apoptotic Regulator Bad and Anti-Apoptotic Regulator Bcl-xL Determine the Type of the Apoptotic Logic Gate. *BMC Syst Biol* (2013) 7:67. doi: 10.1186/1752-0509-7-67
25. Hat B, Jaruszewicz-Błońska J, Lipniacki T. Model-Based Optimization of Combination Protocols for Irradiation-Insensitive Cancers. *Sci Rep* (2020) 10 (1):12652. doi: 10.1038/s41598-020-69380-6
26. Rehm M, Prehn JH. Systems Modelling Methodology for the Analysis of Apoptosis Signal Transduction and Cell Death Decisions. *Methods* (2013) 61 (2):165–73. doi: 10.1016/j.ymeth.2013.04.007
27. Li N, Li J, Desiderio DM, Zhan X. SILAC Quantitative Proteomics Analysis of Ivermectin-Related Proteomic Profiling and Molecular Network Alterations in Human Ovarian Cancer Cells. *J Mass Spectrom* (2021) 56(1):e4659. doi: 10.1002/jms.4659
28. Su F, Zhou F-F, Zhang T, Wang D-W, Zhao D, Hou X-M, et al. Quantitative Proteomics Identified 3 Oxidative Phosphorylation Genes With Clinical Prognostic Significance in Gastric Cancer. *J Cell Mol Med* (2020) 24 (18):10842–54. doi: 10.1111/jcmm.15712
29. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, et al. Connecting Genomic Alterations to Cancer Biology With Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* (2013) 3(10):1108–12. doi: 10.1158/2159-8290.CD-13-0219
30. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res* (2015) 14(6):2707–13. doi: 10.1021/pr501254j
31. Guo T, Luna A, Rajapakse VN, Koh CC, Wu Z, Liu W, et al. Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. *iScience* (2019) 21:664–80. doi: 10.1016/j.isci.2019.10.059
32. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-Generation Characterization of the Cancer Cell Line Encyclopedia. *Nature* (2019) 569(7757):503–8. doi: 10.1038/s41586-019-1186-3
33. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics Identifies New Therapeutic Targets of Early-Stage Hepatocellular Carcinoma. *Nature* (2019) 567(7747):257–61. doi: 10.1038/s41586-019-0987-8
34. Rodriguez H, Zenklusen JC, Staudt LM, Doroshow JH, Lowy DR. The Next Horizon in Precision Oncology: Proteogenomics to Inform Cancer Diagnosis and Treatment. *Cell* (2021) 184(7):1661–70. doi: 10.1016/j.cell.2021.02.055
35. Petralia F, Tignor N, Reva B, Koptyra M, Chowdhury S, Rykunov D, et al. Integrated Proteogenomic Characterization Across Major Histological Types of Pediatric Brain Cancer. *Cell* (2020) 183(7):1962–85.e31. doi: 10.1016/j.cell.2020.10.044
36. Yılmaz ZB, Kofahl B, Beaudette P, Baum K, Ipenberg I, Weih F, et al. Quantitative Dissection and Modeling of the NF-κb P100-P105 Module Reveals Interdependent Precursor Proteolysis. *Cell Rep* (2014) 9(5):1756–69. doi: 10.1016/j.celrep.2014.11.014
37. Arigoni-Affolter I, Scibona E, Lin C-W, Brühlmann D, Souquet J, Broly H, et al. Mechanistic Reconstruction of Glycoprotein Secretion Through Monitoring of Intracellular N-Glycan Processing. *Sci Adv* (2019) 5(11): eaax8930. doi: 10.1126/sciadv.aax8930
38. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global Quantification of Mammalian Gene Expression Control. *Nature* (2011) 473(7347):337–42. doi: 10.1038/nature10098
39. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, et al. Rethinking Drug Design in the Artificial Intelligence Era. *Nat Rev Drug Discov* (2020) 19(5):353–64. doi: 10.1038/s41573-019-0050-3
40. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine Learning in Medicine: A Practical Introduction. *BMC Med Res Methodol* (2019) 19(1):64. doi: 10.1186/s12874-019-0681-4
41. Creamer MS, Stites EC, Aziz M, Cahill JA, Tan CW, Berens ME, et al. Specification, Annotation, Visualization and Simulation of a Large Rule-Based Model for ERBB Receptor Signaling. *BMC Syst Biol* (2012) 6:107. doi: 10.1186/1752-0509-6-107
42. Dolan DWP, Zupanic A, Nelson G, Hall P, Miwa S, Kirkwood TBL, et al. Integrated Stochastic Model of DNA Damage Repair by Non-Homologous End Joining and P53/P21-Mediated Early Senescence Signalling. *PLoS Comput Biol* (2015) 11(5):e1004246. doi: 10.1371/journal.pcbi.1004246
43. Mitra ED, Hlavacek WS. Parameter Estimation and Uncertainty Quantification for Systems Biology Models. *Curr Opin Syst Biol* (2019) 18:9–18. doi: 10.1016/j.coisb.2019.10.006
44. Schomburg I, Chang A, Schomburg D. BRENDA, Enzyme Data and Metabolic Information. *Nucleic Acids Res* (2002) 30(1):47–9. doi: 10.1093/nar/30.1.47
45. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Res* (2021) 49(D1):D498–508. doi: 10.1093/nar/gkaa1025
46. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res* (2016) 44(D1):D1045–53. doi: 10.1093/nar/gkv1072
47. Patterson EA, Whelan MP. A Framework to Establish Credibility of Computational Models in Biology. *Prog Biophys Mol Biol* (2017) 129:13–9. doi: 10.1016/j.pbiomolbio.2016.08.007
48. Roy M, Finley SD. Computational Model Predicts the Effects of Targeting Cellular Metabolism in Pancreatic Cancer. *Front Physiol* (2017) 8:217. doi: 10.3389/fphys.2017.00217
49. Ghaffarizadeh A, Heiland R, Friedman SH, Mumenthaler SM, Macklin P. PhysiCell: An Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems. *PLoS Comput Biol* (2018) 14(2):e1005991. doi: 10.1371/journal.pcbi.1005991
50. Schumacher LJ, Maini PK, Baker RE. Semblance of Heterogeneity in Collective Cell Migration. *Cell Syst* (2017) 5(2):119–27. doi: 10.1016/j.cels.2017.06.006
51. Vipond O, Bull JA, Macklin PS, Tillmann U, Pugh CW, Byrne HM, et al. Multiparameter Persistent Homology Landscapes Identify Immune Cell Spatial Patterns in Tumors. *Proc Natl Acad Sci USA* (2021) 118(41): e2102166118. doi: 10.1073/pnas.2102166118
52. Hernandez-Boussard T, Macklin P, Greenspan EJ, Gryshuk AL, Stahlberg E, Syeda-Mahmood T, et al. Digital Twins for Predictive Oncology Will be a Paradigm Shift for Precision Cancer Care. *Nat Med* (2021) 27(12):2065–6. doi: 10.1038/s41591-021-01558-5

Conflict of Interest: The author had a position as Guest Researcher at Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, LLC, and is Founder and CEO of BioSystems Strategies, LLC.

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 Oncology

TECHNOLOGY AND CODE
published: 16 February 2022
doi: 10.3389/fonc.2021.806603

Deep Learning-Based Mapping of Tumor Infiltrating Lymphocytes in Whole Slide Images of 23 Types of Cancer

Shahira Abousamra[1], Rajarsi Gupta[2], Le Hou[1], Rebecca Batiste[3], Tianhao Zhao[3], Anand Shankar[4], Arvind Rao[4], Chao Chen[2], Dimitris Samaras[1], Tahsin Kurc[2] and Joel Saltz[2]*

[1] Department of Computer Science, Stony Brook University, Stony Brook, NY, United States, [2] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States, [3] Department of Pathology, Stony Brook University, Stony Brook, NY, United States, [4] Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, United States

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Ole Winther, University of Copenhagen, Denmark
Konstantinos Zormpas-Petridis, Institute of Cancer Research (ICR), United Kingdom

***Correspondence:**
Shahira Abousamra
sabousamra@cs.stonybrook.edu

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *01 November 2021*
Accepted: *31 December 2021*
Published: *16 February 2022*

Citation:
Abousamra S, Gupta R, Hou L, Batiste R, Zhao T, Shankar A, Rao A, Chen C, Samaras D, Kurc T and Saltz J (2022) Deep Learning-Based Mapping of Tumor Infiltrating Lymphocytes in Whole Slide Images of 23 Types of Cancer. Front. Oncol. 11:806603. doi: 10.3389/fonc.2021.806603

The role of tumor infiltrating lymphocytes (TILs) as a biomarker to predict disease progression and clinical outcomes has generated tremendous interest in translational cancer research. We present an updated and enhanced deep learning workflow to classify 50x50 um tiled image patches (100x100 pixels at 20x magnification) as TIL positive or negative based on the presence of 2 or more TILs in gigapixel whole slide images (WSIs) from the Cancer Genome Atlas (TCGA). This workflow generates TIL maps to study the abundance and spatial distribution of TILs in 23 different types of cancer. We trained three state-of-the-art, popular convolutional neural network (CNN) architectures (namely VGG16, Inception-V4, and ResNet-34) with a large volume of training data, which combined manual annotations from pathologists (strong annotations) and computer-generated labels from our previously reported first-generation TIL model for 13 cancer types (model-generated annotations). Specifically, this training dataset contains TIL positive and negative patches from cancers in additional organ sites and curated data to help improve algorithmic performance by decreasing known false positives and false negatives. Our new TIL workflow also incorporates automated thresholding to convert model predictions into binary classifications to generate TIL maps. The new TIL models all achieve better performance with improvements of up to 13% in accuracy and 15% in F-score. We report these new TIL models and a curated dataset of TIL maps, referred to as *TIL-Maps-23*, for *7983* WSIs spanning *23* types of cancer with complex and diverse visual appearances, which will be publicly available along with the code to evaluate performance.

Code Available at: https://github.com/ShahiraAbousamra/til_classification.

Keywords: TIL maps, digital histopathology, whole slide images, tumor infiltrating lymphocytes, deep learning, large scale analysis

1 INTRODUCTION

Tumor infiltrating lymphocytes (TILs) have gained importance as a biomarker in translational cancer research for predicting clinical outcomes and guiding treatment. As our collective understanding of tumor immune responses in cancer expands, clinical research studies have shown that high densities of TILs correlate with favorable clinical outcomes (1), such as longer disease-free survival (2) and/or improved overall survival in multiple types of cancer (3). Studies also suggest that the spatial distribution of TILs within complex tumor microenvironments may play an important role in cancer prognosis (4–6). These findings have led to efforts to characterize the abundance and spatial distribution of TILs in cancer tissue samples to further our understanding of tumor immune interactions and help develop precision medicine applications in oncology (7–11).

Computational image analysis of whole slide images (WSIs) of cancer tissue samples has become a very active area of translational biomedical research. The goals are to gain novel insights into cancer and the tumor microenvironment, including tumor immune responses, through the search for biomarkers to predict outcomes and treatment response. Modern digital microscopes scan whole slide tissue samples at very high image resolutions, ranging from 50,000x50,000 pixels to over 100,000x100,000 pixels. The increasing availability of such gigapixel WSIs has stimulated the development of image analysis methods for detection, segmentation, and classification of microanatomic regions, structures, cells, and other objects in tissue images. Therefore, we utilized advances in computer vision and machine learning to quantitatively characterize TILs to complement qualitative microscopic evaluation of cancer tissue samples by pathologists. Deep learning has become the preferred approach for a variety of image analysis tasks in recent years (12–17) since these methods can analyze raw image data and do not require specified instructions to identify and quantify engineered image features. Furthermore, deep learning-based image analysis workflows have been shown to consistently produce more accurate results and generalize to new datasets better than previous image analysis methods in computational pathology.

Several projects have implemented methods to detect and classify lymphocytes in tissue images. Eriksen et al. (18) employed a commercial system to count CD3+ and CD8+ cells in tissue images that were obtained from stage II colon cancer patients and stained with an immunohistochemistry (IHC) protocol. Swiderska-Chadaj (19) also trained a deep learning model with a dataset of 171,166 annotated CD3+ and CD8+ cells in images of IHC stained tissue specimens from breast, prostate and colon cancer cases. Garcia et al. (20) proposed a deep learning model to count TILs in IHC images of gastric cancer tissue samples by using a model trained with 70x70 square pixel patches extracted from biopsy micrographs scanned at 40x magnification and labeled by pathologists. PathoNet, developed by Negahbani et al. (21), implements a deep learning model based on the U-Net architecture (22) for detection and classification of Ki-67 and TILs in breast cancer cases.

Methods were also developed to study TILs in Hematoxylin and Eosin (H&E) stained tissue images. Budginaite et al. (23) developed a deep learning workflow based on the Micro-Net architecture (24) and multi-layer perceptrons to identify lymphocytes in tissue images from breast and colorectal cancer cases. Corredor et al. (25) investigated the spatial patterns of TILs in early stage non-small cell lung cancer cases with the goal of predicting cancer recurrence. Jaber et al. (26) investigated TILs in non-small cell lung cancer cases by employing deep learning architectures and support vector machines to classify 100x100 square micron patches in WSIs. Acs et al. (27) developed a computerized TIL scoring method using QuPath software (28) to cluster melanoma cancer patients into those with favorable prognosis and those with poor prognosis. Linder et al. (29) evaluated the use of deep learning for TIL analysis in tissue images of testicular germ cell tumors by using commercial image analysis software and implementing a two stage workflow in which the first stage processed WSIs to detect regions that contained TILs and the second stage counted the TILs in those regions, demonstrating how deep learning-based methods can be used successfully for TIL detection in germ cell cancer. Amgad et al. (30) proposed a deep learning workflow based on a fully convolutional network architecture developed by Long et al. (31) to identify tumor, fibroblast, and lymphocyte nuclei and tumor and stroma regions. Le et al. (32) developed deep learning models for segmentation of tumor regions and detection of TIL distributions in whole slide images of breast cancer tissues by training models based on VGG16, Inception-V4, and Resnet-34 architectures that used WSIs from The Surveillance, Epidemiology, and End Results (SEER) Program at the National Cancer Institute (NCI) and the Cancer Genome Atlas (TCGA) repository.

Despite an increasing number of projects, there are few large scale datasets of WSIs that are publicly available to study TILs. Moreover, most of the previous projects targeted specific types of cancer from particular organ sites. The classification of TILs can be challenging in large datasets of WSIs across multiple types of cancer from different organ sites for many reasons. Deep learning models need to distinguish TILs from cancer cells that are intrinsically complex across a wide spectrum of growth patterns, cellular and nuclear morphologies, and other histopathologic features associated with specific types of cancer, which vary by organ site, state of cellular differentiation, and stage of cancer (e.g. primary organ site versus a metastatic tumor deposit). Computational image analysis of pathology WSIs is also complicated by variations in image properties from differences in scanning with different types of digital slide scanners and varying tissue staining laboratory protocols. **Figure 1** shows an example of identifying TILs in a WSI and the heterogeneity of the appearance and distribution of TILs in different tissue samples. Before our work, the largest TIL dataset was generated by Saltz et al. (33), where 5202 WSIs from 13 cancer types were analyzed.

In this paper we describe a deep learning workflow that was utilized to generate a large dataset of TIL maps, referred to here as the TIL-Maps-23 dataset. Unlike the previous work that studied TILs in mostly common types of cancer, we trained a deep learning model with the goal of analyzing WSIs from a much wider range of different types of cancer. We adopted the same approach of patch-wise classification as in (33), where each WSI is partitioned into non-overlapping patches of size 50 x 50 square microns. A trained deep

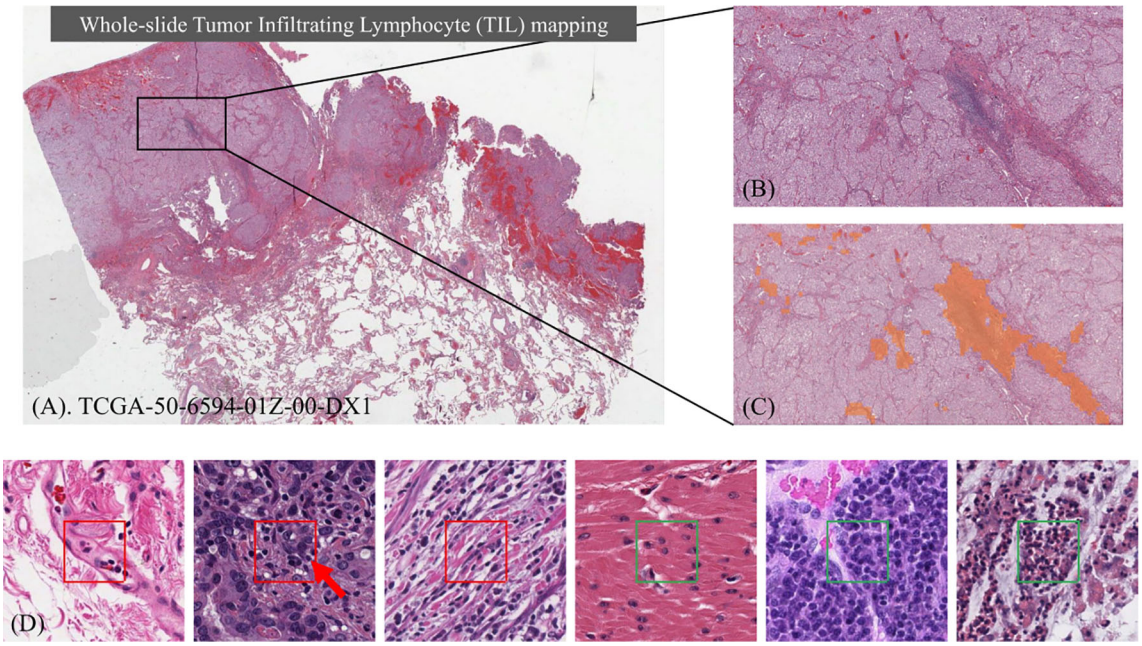


FIGURE 1 | Identifying Tumor Infiltrating Lymphocyte (TIL) regions in gigapixel pathology WSIs. **(A)** H&E stained WSI of lung adenocarcinoma. **(B)** Example of a region of tissue. **(C)** Example of a TIL map overlaid on the region of tissue. **(D)** Examples of TIL positive (framed in red) and negative (framed in green) patches. A lymphocyte is typically dark, round to ovoid, and relatively small compared to tumor and normal nuclei. Sample patches show the heterogeneity in TIL regions and how it can be challenging to differentiate TIL positive and TIL negative regions.

learning CNN model classifies each patch as TIL-positive or TIL-negative and then compiled to generate a TIL map of the WSI. While a classification at the cellular level allows finer grain analysis, patch-level classification offers several advantages. First, it requires much less annotation time and effort. The pathologist can just mark regions as TIL positive or TIL negative and then we can sample patches from these regions. On the other hand, cell-level annotations require marking each individual lymphocyte cell in a patch. Second, optimizing nuclear classification is more challenging over multiple cancer types and needs much more data. Our approach allows us to scale the dataset to develop a model to span more cancer types with much less effort. Third, the identifying lymphocytes at a 50 microns resolution provides valuable and interpretable information about the spatial distributions of TILs across large sets of WSIs to study many samples from a particular type of cancer and/or compare the role of TILs in different types of cancer, which can be further studied in downstream correlative analyses. In an earlier work (33), we applied spatial statistics to patch-level TIL predictions in WSIs and demonstrated that spatial clustering patterns of TILs correlate with molecular features and clinical outcomes. In another work (32), we computed TIL infiltration amounts by combining patch-level TIL predictions with tumor segmentation results in breast cancer and showed correlations between TIL infiltration and survival that was stratified by molecular subtype.

The work presented in this manuscript focuses on an improved deep learning workflow for patch-level TIL prediction and generation of a large dataset of TIL predictions across multiple cancer types. We plan to carry out additional studies to ascertain the clinical relevance of TIL predictions in future works. Our work improves on the earlier work done by Saltz et al. (33) in several ways. The previous work trained two CNN deep learning models, one for detecting lymphocytes and the other for segmenting necrosis regions by using convolutional neural networks (CNNs) developed in-house. The necrosis segmentation model was used to eliminate false TIL-positive predictions in necrotic regions of tissues, which required two separate training datasets. This new and improved deep learning workflow employs a single CNN by adapting popular, engineered classification networks and using a combination of manual annotations and machine-generated annotations as training data. Moreover, the previous work included a manual thresholding step in order to generate the final binary TIL maps. This step consisted of a patch sampling process and a manual review of the sampled patches to set TIL-positive/TIL-negative thresholds for different WSIs. The new workflow implements an automated mechanism for computing thresholds to map model predictions to binary classifications. This eliminates the manual thresholding step of the previous work. After all of these improvements, we present the TIL-Maps-23 dataset for 23 types of cancer, which is the largest collection of curated TIL maps across both common and rare types of cancer to date.

2 MATERIALS AND METHODS

The overall analysis workflow is illustrated in **Figure 2**. The workflow consists of training data generation, model training,

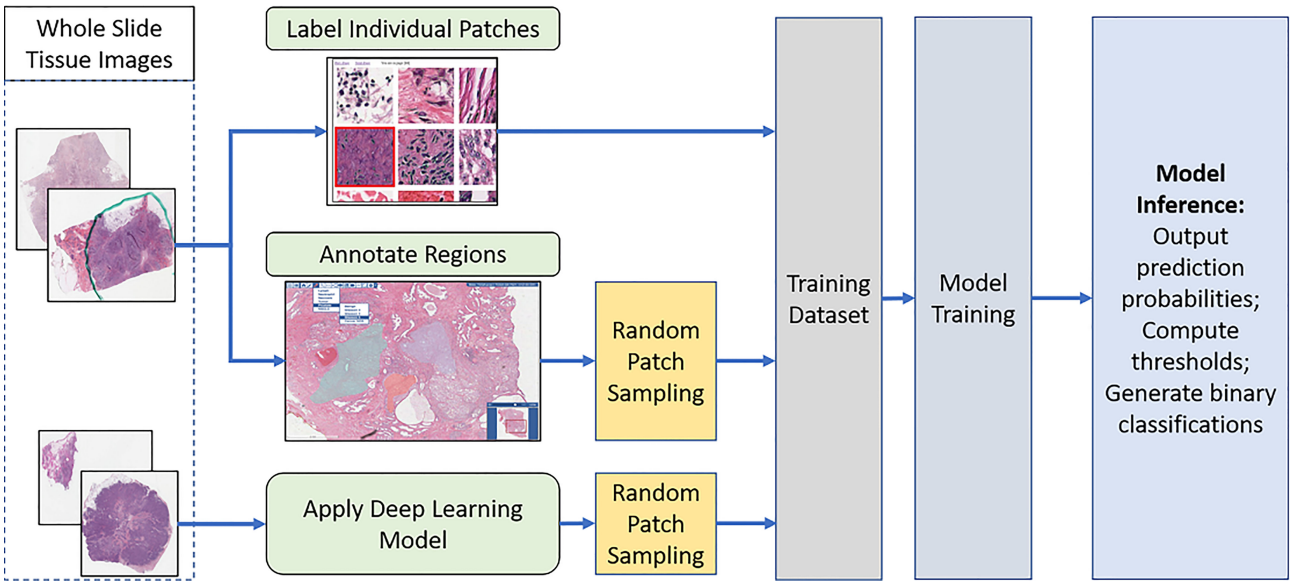


FIGURE 2 | The overall analysis workflow: Training data generation, model training, inference (prediction), and computing threshold values. Training data is generated *via* a combination of manual annotations and model-generated predictions. A trained model generates predictions in the form of the probability that a patch is TIL-positive. The probability values are mapped to binary classifications by applying an automatically computed threshold.

and inference steps. The training dataset is generated by combining labels from manual patch-level and region-level annotations, as well as classification predictions generated by the deep learning model developed in (33). The inference step (**Figure 3**) partitions WSIs into patches, outputs patch-level probability values, and executes an automated method to compute thresholds for mapping the probability values to binary classifications.

2.1 Generating Training Dataset

We created a training dataset by combining manually annotated patches (strong annotations) from 18 TCGA cancer types (ACC, BRCA, COAD, ESCA, HNSC, KIRC, LIHC, LUAD, MESO, OV, PAAD, PRAD, SARC, SKCM, TGCT, THYM, UCEC, and UVM) and model-generated annotations from 4 TCGA cancer types (CESC, LUSC, READ, and STAD). For the model-generated annotations, we sampled a set of patches classified

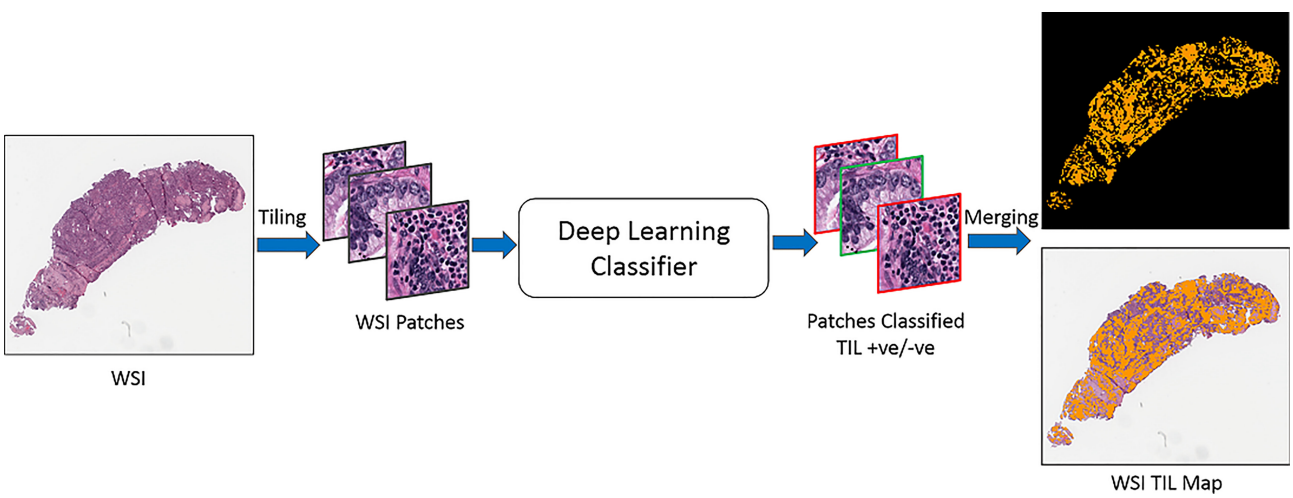


FIGURE 3 | TIL inference step. An input WSI is partitioned into disjoint patches of 50x50 square microns. Each patch is processed by the trained classification model and assigned a probability value. The probability values are mapped to binary classifications. A TIL map covering the entire WSI is generated.

by the model in (33). The model-generated annotations are employed not only as a cost-saving mechanism to reduce manual annotation workload but also to increase diversity in texture and appearance of tissue data. Variations in texture and appearance are often the case with H&E images, especially with a dataset like TCGA which comes from multiple sites, each using their own slide scanners and staining protocols. We have shown previously in (34) that combining manual annotations with model-generated annotations for cancer types with scarce or no manual annotations gives better results compared to using manual annotations alone.

The manual annotations are generated in 2 ways. First, patches of 150 x 150 square microns are randomly sampled from the WSIs. Pathologists annotate the center 50 x 50 square micron sub-patch in each patch. The annotation indicates whether the center sub-patch is TIL-positive or TIL-negative. Using a 150 x 150 square micron patch allows pathologists to see the surrounding tissue for a more informed decision on the label of the center sub-patch. Only the center sub-patch is used in training. A patch is labeled TIL-positive if it has at least 2 lymphocytes or plasma cells in the center sub-patch. Second, pathologists mark TIL-positive and TIL-negative regions on WSIs, where TIL-positive regions are regions with a significant amount of lymphocytes and/or plasma cells. Patches of 50 x 50 square microns are randomly sampled from these regions, where each patch is assigned the same label as the source region.

The model-generated annotations are collected from classifications produced by the previous model in (33). This model employed a human-in-the-loop TIL classification procedure, where a manual threshold step was applied to the predicted TIL probability maps in order to produce binary classifications. In our work, we randomly sampled TIL-positive and TIL-negative patches from the binary classifications.

2.2 Deep Neural Network Models and Training

We trained 3 models with different networks: VGG-16 (35), ResNet-34 (36), and Inception-V4 (37). These networks are engineered for image classification. They have been shown to be powerful classifiers on the ImageNet dataset (38) and have been adopted in various computer vision applications. The main differences between the 3 networks can be summarized as follows: VGG-16 has a basic convolutional neural network architecture; ResNet-34 is much deeper and features skip connections that allow a more stable training of the deeper network; and Inception-v4 is an even a deeper network, where each block in the network utilizes residual connections and convolutional layers of various sizes to capture features at different resolutions and reception fields.

Each network is initialized with weights from the respective pre-trained model on ImageNet. The batch normalization layers are dropped. Each input image (patch) is scaled with bilinear interpolation to match the network's pre-training input size (i.e., 224 x 224 pixels for VGG-16, 299 x 299 pixels for Inception-V4, and 100 x 100 for ResNet-34). The input image is normalized to the range [−1, 1] for VGG-16 and Inception-V4 by $img = (\frac{img}{255} - 0.5) \times 2$. For ResNet-34, the input image is normalized with the same mean and standard deviation vectors as the pre-trained model. The training phase implements data augmentation, including random rotation and flipping, shifting of input patches left/right and up/down by a random number of pixels in the range of [−20, +20], and color augmentation *via* small variations to brightness and color in the hue, saturation, and lightness (HSL) space. All of the networks were trained end-to-end using the cross entropy loss.

2.3 Determining Binary Classification Thresholds

The trained models output a probability value for each patch in an input WSI. This creates a probability map for the entire WSI. The final binary prediction (TIL positive or TIL negative) is obtained by thresholding the probability map. If the probability of a patch is greater than or equal to the threshold value, the patch is classified as TIL-positive. Otherwise, it is classified as TIL-negative.

A default threshold value of 0.5 was used during training to evaluate a model's performance in each training epoch. At the end of the training phase, the threshold value was fine-tuned for the inference phase. A threshold value in the range [0.4, 0.6] was selected for each model based on the performance of the model on a small *hold-out* dataset. We evaluated two methods for selecting the threshold value for each model. The first method relies on the true positive rate (TPR) and the false positive rate (FPR) (39). The optimal (FPR, TPR) pair is (0,1). The threshold selection method minimizes the FPR and maximizes the TPR. **Figure 4** shows an example receiver operating characteristic (ROC) curve ($x = FPR$, $y = TPR$). The length of the line from the (0,1) point and intersecting the curve at (fpr, tpr) is $\sqrt{\text{fpr}^2 + (1 - \text{tpr})^2}$. By selecting the threshold value that minimizes the distance from (0,1) to the curve, FPR and (1-TPR) are minimized. The second method is based on the Youdin Index, which is commonly used to select a threshold that maximizes TPR - FPR (40). In our experiments, both methods resulted in almost identical binary classification maps. The threshold values selected for the VGG-16, ResNet-34, and Inception-V4 models were 0.4, 0.56, and 0.41, respectively.

2.4 Software Support for Training Data Generation and Review of Analysis Results

The WSIs in the image dataset are loaded to a software platform, called Quantitative Imaging in Pathology (QuIP), for training data generation and review of the model predictions. QuIP consists of multiple services, implemented as micro-services with software containers, and a set of Web-based applications that support viewing of WSIs, annotation of image regions and patches, and interactive viewing of model predictions as heatmaps overlaid on WSIs (41).

One of the web applications is a markup and annotation tool with multiple class label selections (**Figure S2** in supplementary material). This tool enables annotations of full-resolution whole slide tissue images. The user can draw a polygon to mark up a

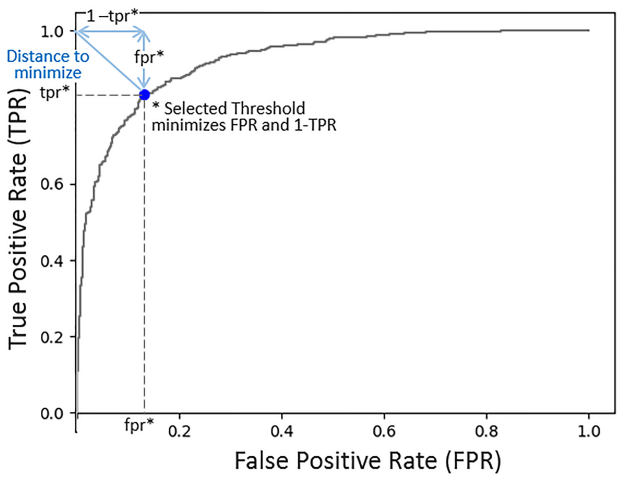


FIGURE 4 | Model probability threshold selection. The objective is to minimize the false positive rate (FPR) and maximize the true positive rate (TPR). This ROC plot (FPR, TPR), illustrates that by minimizing the distance from the point (0,1) to the curve, we are minimizing $FPR^2 + (1 - TPR)^2$, thus achieving our objective. The selected threshold corresponds to the FPR and TPR at which the line intersects with the curve minimizing the distance to the point (0,1).

region and select a label from a pull-down menu to label the region. Multiple regions and classes can be annotated in an image. In addition to marking regions, pathologists can annotate individual patches. Another web application is used for this purpose. A set of patches are displayed to the user who can assign a label to each patch by clicking on the patch. To minimize the number of mouse clicks (or taps on touch screens) for the binary classification case, we assume a default class for all patches. The user clicks on patches that belong to the alternative class only.

Manual examination of model predictions requires interactive interrogation and visual analytic tools that link these results with the underlying images. QuIP implements two tools for this purpose; the FeatureMap tool and the heatmap viewer/editor. The FeatureMap tool converts probability maps into low resolution heatmaps, called featuremaps, which can be visualized at a lower image resolution than at the resolution of whole slide images (**Figure 5A**). Each pixel in a featuremap image corresponds to a patch in the WSI. The goal is to let a user rapidly go through a set of images without having to load heatmaps on full-resolution images and pan and zoom in the images. After reviewing a featuremap, the user can click anywhere on the featuremap image and visualize the region at full image resolution using the heatmap viewer/editor. The heatmap viewer/editor allows a user to access full-resolution heatmap representation of a probability map overlaid on the input WSI and re-label algorithm predictions (**Figure 5B**). The user can click on an area in a heatmap, zoom and pan, and interactively examine the areas of interest. If the user determines that predictions in some areas should be corrected, the user switches to the heatmap editor and annotates a set of patches to be positive or negative on the WSI. The FeatureMap and heatmap viewer/editor tools rely on the backend data management and indexing services of QuIP, namely PathDB for managing images and FeatureMap data and FeatureDB for managing probability maps and user annotations.

2.5 Evaluating Model Performance

We evaluated the performances of the trained models *via* two methods: patch-level classification accuracy and region categorical classification performance.

For patch-level classification accuracy, we collected manually labeled test patches and measured the performance of each model with these patches using the accuracy and F-score metrics. The accuracy metric represents the percent of correctly classified patches and is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\,\% \tag{1}$$

Here TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. The F-score measures the balance of model precision and how many of the positive patches are correctly classified (i.e. recalled). It is computed as:

$$Precision = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

For the region categorical classification performance, we adopted the evaluation method implemented in (33). We evaluated the correlation between predictions from the models and annotations (labels) from the pathologists, both quantitatively and qualitatively using *super-patches*. Super-patches make it easier to collect a large number of annotations from multiple pathologists. This evaluation method provides a higher level of evaluation that is beyond individual patches and offers a quantification of the correlation between a model's predictions and a pathologist's perception of TIL distribution.

A super-patch is defined as a large 800 x 800 square pixel patch at 20x magnification (i.e., a super-patch covers a 400 x 400 square micron area in tissue). The deep learning models classify 100 x 100 square pixel patches at 20x magnification. Hence, each super-patch is divided into an 8 x 8 grid, and each patch (of 100 x 100 square pixels) is classified as TIL-Positive or TIL-Negative. **Figure 6** shows an example of a super-patch and the labeling of its patches.

In our work, each super-patch was annotated by one to three pathologists as Low TIL, Medium TIL, or High TIL, based on the perceived fraction of the area of the TIL-positive patches. The *score* of a deep learning model for a given super-patch is the number of patches classified as TIL-positive by the model. Hence, each super-patch gets assigned a score between 0 to 64.

We use the polyserial correlation method (42, 43) to quantify the correlation between the model scores and the pathologist annotations. Polyserial correlation measures the inferred latent correlation between a continuous variable and an ordered

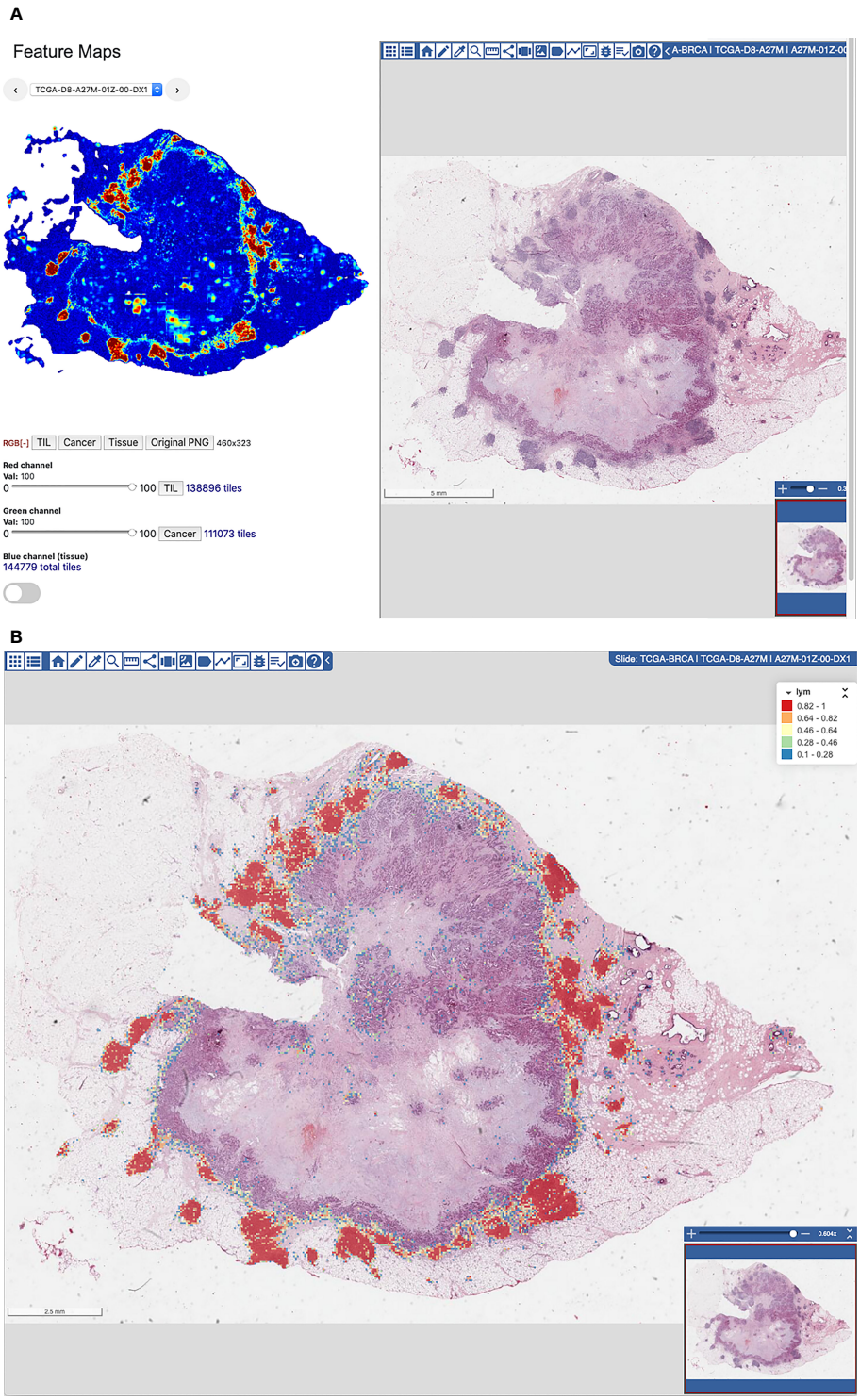


FIGURE 5 | **(A)** FeatureMap along with a view of the tissue image. **(B)** Heatmap viewer and editor for viewing of heatmaps on full-resolution WSIs and for fine-grain re-labeling of patches to generate additional training data.

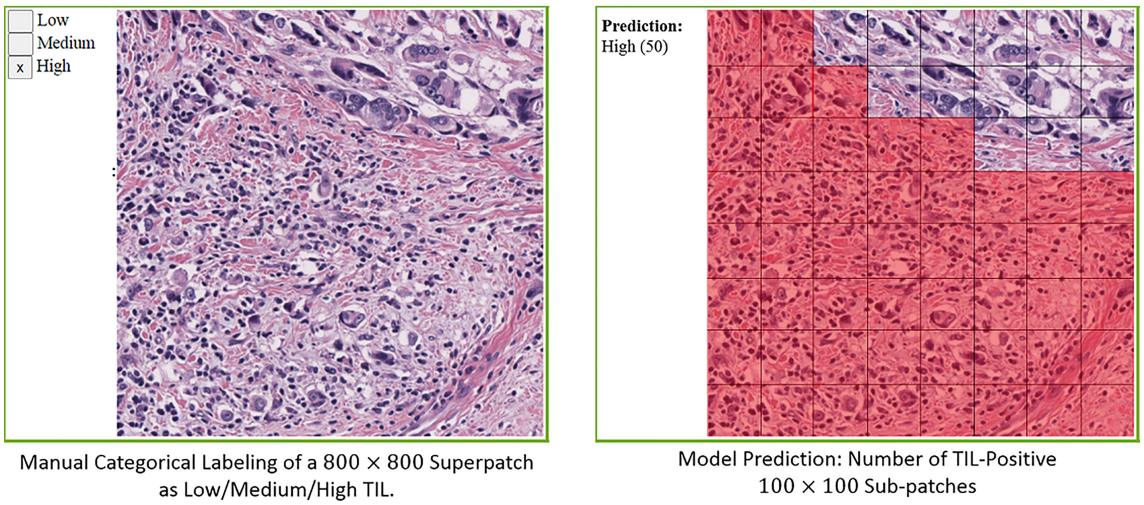


FIGURE 6 | Illustration of a superpatch labeling and prediction.

categorical variable, which, in our case, represent scoring by the model and the rounded average TIL-positive annotations from the pathologists, respectively. We also used violin plots for the qualitative evaluation of the correlation between the model scoring and the pathologists' categorical labels. Violin plots can be viewed as box plots that show the smoothed probability density distribution rotated on each side.

3 RESULTS

3.1 Dataset and Implementation Details

The number of patches in training and test sets are given in **Tables S1** and **S2** in the supplementary material. On average, 19 WSIs per cancer type were used in manually annotated training data and 117 WSIs per cancer type were used in model-generated training annotations. There were 351,272 patches in total in the training dataset. Out of these patches, 282,065 were manually annotated and 69,207 were patches from the model-generated annotations dataset. The model-generated annotations allowed us to reduce the manual annotation effort by 19% and increase training data diversity by covering 22 cancer types (the training dataset did not include patches from BLCA), while maintaining a good ratio of strong annotations to model-generated annotations.

We trained three models with popular networks, namely Inception V4 (37), VGG-16 (35) and ResNet-34 (36), as described in Section 2. The models were trained with the Adam optimizer using a learning rate of 0.00005 and a batch size of 128.

3.2 Patch-Level Classification Accuracy

We collected 327, 299, 326, and 299 of manually labeled test patches from BRCA, LUAD, SARC, and OV, respectively, and 888 patches in total from the other cancer types with 47 patches per cancer type on average. **Tables 1** and **2** show the accuracy and F-score, respectively, for the three models, as well as the model trained in (33), referred to as the *Baseline* model in the tables. The columns LUAD, BRCA, SARC, and OV show the performance numbers in each metric for the patches collected from these four cancer types. The columns *Other* and *All* show the performance values with the 888 patches from the other cancer types and with all of the patches, respectively. The column *13 Cancer Types* shows the performance comparison between the Baseline model and the newer models with patches from the 13 cancer types (BLCA, BRCA, CESC, COAD, LUAD, LUSC, PAAD, PRAD, READ, SKCM, STAD, UCEC, and UVM) analyzed in the previous work (33). The results show that the new models outperformed the Baseline model by up to 13% in accuracy and 15% in F-score.

TABLE 1 | Evaluation of patch classification accuracy.

Model Name	LUAD	BRCA	SARC	OV	Other*	13 cancer types**	All
Baseline	73.60%	74.90%	–	–	–	79.56%	–
VGG-16	83.28%	**88.38%**	94.17%	88.29%	82.52%	83.32%	86.02%
ResNet-34	84.28%	86.24%	91.41%	87.29%	82.10%	82.45%	85.14%
Incep-V4	**86.29%**	87.16%	**96.93%**	**94.31%**	**82.53%**	**83.68%**	**87.43%**

*Compare result for each of LUAD, BRCA, SARC, OV, *Other: patches from other cancer types in the set of 23 types used in training, **13 cancer types: subset of test patches belonging to the 13 cancer types the baseline model with human in the loop (Baseline) (33) was trained on, All: all test patches from all the 23 cancer types. Best accuracy in each dataset is indicated in bold.*

TABLE 2 | Patch classification F-score results.

Model Name	LUAD	BRCA	SARC	OV	Other*	13 cancer types**	All
Baseline	0.78	0.77	–	–	–	0.85	–
VGG-16	0.85	0.88	0.92	0.84	0.85	0.86	0.86
ResNet-34	0.87	0.87	0.88	0.82	0.86	0.86	0.86
Incep-V4	**0.89**	**0.89**	**0.96**	**0.93**	**0.87**	**0.88**	**0.89**

*Compare result for each of LUAD, BRCA, SARC, OV, *Other: patches from other cancer types in the set of 23 types used in training, **13 cancer types: subset of test patches belonging to the 13 cancer types the baseline model with human in the loop (Baseline) (33) was trained on, All: all test patches from all the 23 cancer types. Best F-score in each dataset is indicated in bold.*

All of the new models performed well, attaining high accuracy and F-score values. In most of the cases, the Inception V4 model achieved better performance, in the range of 1–5% higher values, than the other models.

3.3 Region Categorical Classification

We collected manual annotations on 4,198 randomly selected super-patches from the 23 cancer types. **Table 3** shows the polyserial correlation coefficient for each model for super-patches from individual cancer types. The last column in the bottom set of the table is the polyserial correlation coefficient with respect to the collective set of super-patches and the mean and standard deviation over the correlation coefficients of the individual cancer types. The results show that no single model is consistently better than the other models. The Inception V4 model achieves a higher mean score as shown in the *ALL* column of the table. The correlation coefficients are the lowest for KIRC. The nuclei of cells in KIRC are generally small, dark, and rounded, which gives the tumor cells a similar appearance to lymphocytes. Thus, the deep learning models classify them incorrectly and overestimate TIL regions. **Figure 7** shows some of the super-patches that were incorrectly scored by the Inception V4 model. The left panel in the figure shows the categorical label (Low, Medium and High) of the super-patch assigned by the pathologists as well as the model prediction and the number of patches classified as TIL-positive by the model in parentheses. For the sake of presentation in the figure, the model prediction is described as Low, if the model score is $0 \leq$score ≤ 21, Medium if the score is $22 \leq$score ≤ 42, and High >42. Similar low correlations were obtained with super-patches from OV. The Inception V4 model resulted in under-estimation in 14 cases versus over-estimation in 9 cases of the OV super-patches. **Figure 8** shows various sample results from the model with the OV super-patches, illustrating the discrepancy between the model scoring and the pathologists' classifications. The polyserial correlation coefficient is greater than or equal to 0.8 for 13 cancer types (ACC, BRCA, ESCA, HNSC, LIHC, MESO, PAAD, PRAD, READ, SARC, SKCM, TGCT, and UVM), between 0.7 and 0.8 for 5 cancer types (LUSC, THYM, STAD, BLCA, and UCEC) and below 0.7 for 5 cancer types (COAD, CESC, OV, LUAD, and KIRC).

Figure 9 shows the violin plots for scores from each deep learning model against the rounded average of pathologists' annotations. The visual representations of the density distributions and the median values indicate that the VGG-16 model tends to under-estimate TILs. The ResNet-34 and Inception-V4 models are more consistent with the pathologist categorical labeling, where the Inception-V4 model performs better.

3.4 TIL Area Estimation

After we evaluated the performance of these TIL models and visually confirmed how well TILs were being classified in WSIs across 23 types of cancer, the next step was to utilize the best TIL model to analyze all of the available diagnostic DX1 TCGA WSIs in these types of cancer to characterize the abundance and spatial distribution of TILs as a potential biomarker. Based on our

TABLE 3 | Superpatches evaluation using polyserial correlation coefficient.

Model Name	ACC (147)	BLCA (64)	BRCA (348)	CESC (61)	COAD (65)	ESCA (312)	HNSC (324)	KIRC (319)
Baseline	–	0.720	0.552	**0.679**	0.329	–	–	–
VGG-16	0.879	**0.787**	0.745	0.592	0.688	0.777	**0.904**	0.515
ResNet-34	0.925	0.740	0.797	0.654	0.658	**0.810**	0.883	**0.599**
Incep-V4	**0.963**	0.744	**0.797**	0.667	**0.695**	0.805	0.897	0.598
Model Name	**LIHC (248)**	**LUAD (63)**	**LUSC (65)**	**MESO (271)**	**OV (158)**	**PAAD (440)**	**PRAD (66)**	**READ (62)**
Baseline	–	0.615	0.658	–	–	0.695	0.819	0.706
VGG-16	**0.891**	0.670	**0.830**	**0.840**	0.565	**0.886**	**0.885**	0.702
ResNet-34	0.872	**0.733**	0.775	0.805	0.527	0.874	0.862	0.715
Incep-V4	0.854	0.617	0.789	0.818	**0.635**	0.870	0.818	**0.811**
Model Name	**SARC (299)**	**SKCM (67)**	**STAD (63)**	**TGCT (303)**	**THYM (324)**	**UCEC (64)**	**UVM (64)**	**ALL (4198)**
Baseline	–	0.666	0.728	–	–	0.692	0.681	–
VGG-16	0.912	0.816	0.713	**0.859**	0.774	0.667	0.896	0.807 (0.77 ± 0.12)
ResNet-34	**0.932**	0.794	**0.821**	0.799	0.765	**0.766**	0.899	0.808 (0.78 ± 0.10)
Incep-34	0.921	**0.822**	0.752	0.823	**0.790**	0.742	**0.913**	**0.820 (0.79 ± 0.10)**

The number in brackets indicated the number of superpatches in the respective cancer type. Baseline is the model developed in (33).
Highest polyserial correlation in each dataset (cancer type) is indicated in bold.

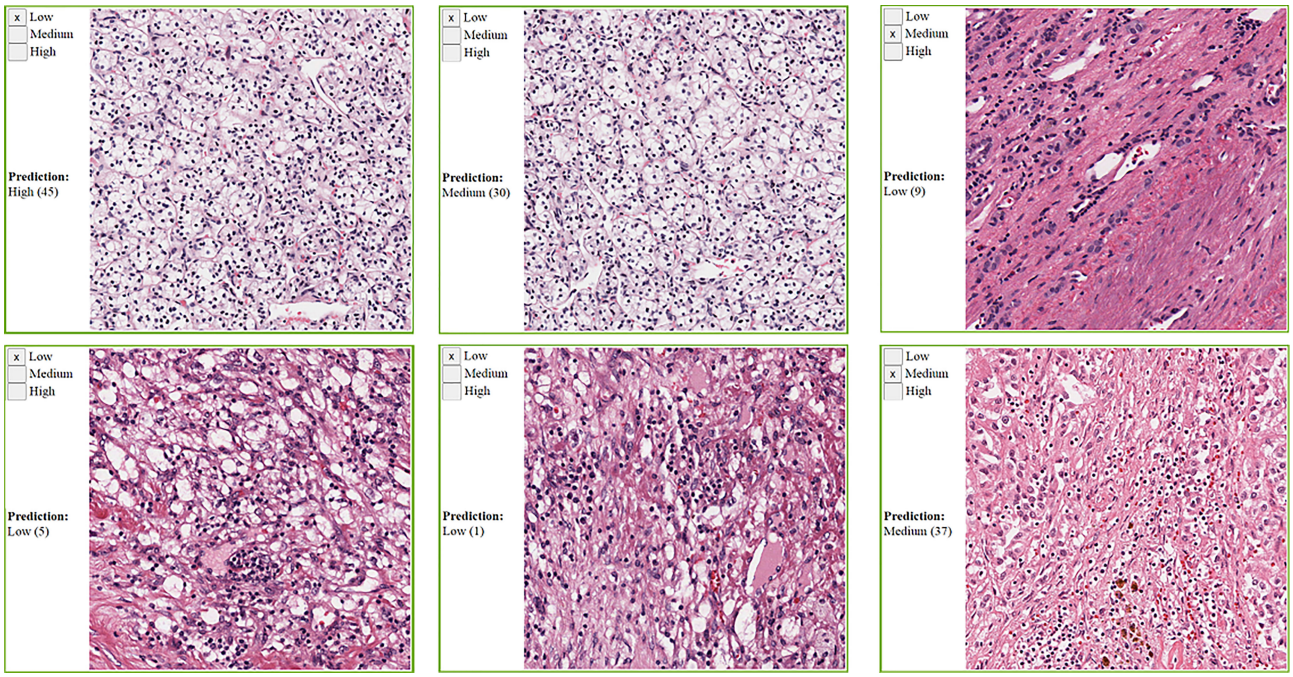


FIGURE 7 | Sample KIRC super-patches, showing the categorical label and the Inception model prediction. KIRC is challenging because other cell types nuclei can look like lymphocytes. The model prediction is displayed as a category and a score between brackets. The models' scoring is a value in the range 0 to 64. We roughly interpret it as: Low if 0 ≤score≤ 21, Medium if 22 ≤score ≤ 42, and High otherwise. Top row: cases where the category approximated from the model scoring does not match the pathologists' label. Bottom row: cases where the category approximated from the model scoring matches the pathologists' label.

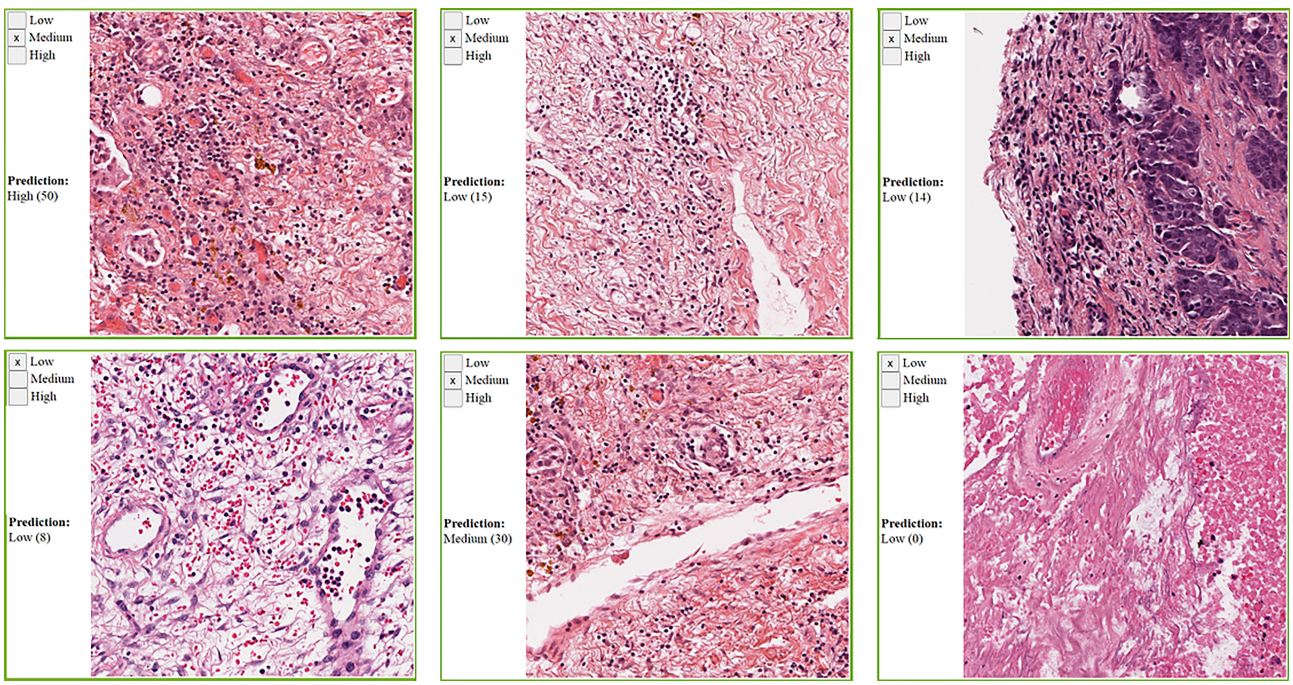


FIGURE 8 | Sample OV super-patches, showing the categorical label and the Inception model prediction. The model prediction is displayed as a category and a score between brackets. The models' scoring is a value in the range 0 to 64. We roughly interpret it as: Low if 0 ≤core ≤ 21, Medium if 22 ≤core ≤ 42, and High otherwise. Top row: cases where the category approximated from the model scoring does not match the pathologists' label. Bottom row: cases where the category approximated from the model scoring matches the pathologists' label.

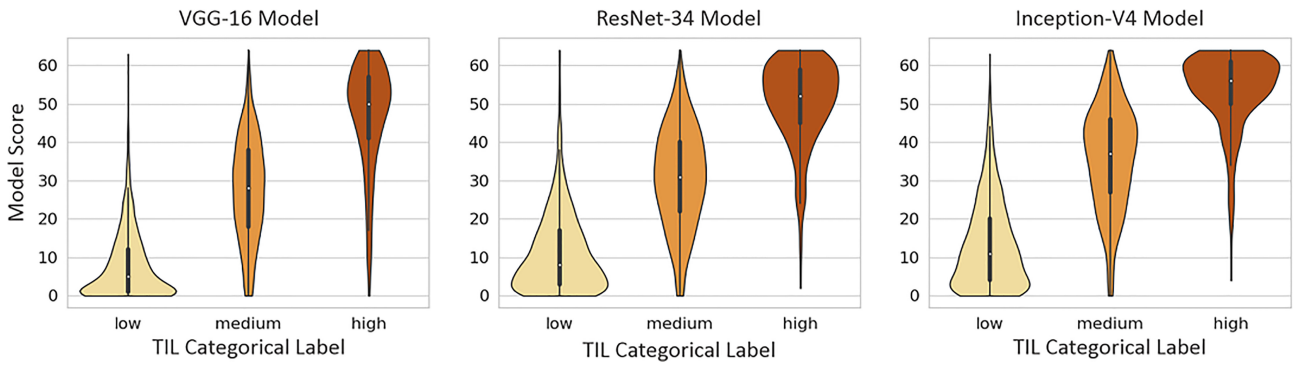


FIGURE 9 | Violin plots of each model's scores against super-patch categorical labels (Low, Medium, and High TIL).

evaluations, we utilized the Inception model to analyze all diagnostic DX1 TCGA WSIs since it had the highest patch classification accuracy and F-score and best overall performance on the super-patches. We used the Inception-V4 TIL model to generate all of the TIL maps in this dataset and compute the estimated average area that is infiltrated by TILs per WSI in the dataset across 23 types of cancer. The results are summarized in **Table 4** and demonstrate how computational pathology is very useful in characterizing TILs as a biomarker, which can be very helpful in guiding future clinical research in precision oncology and immunotherapy by supporting cohort discovery by identifying potential types of cancer with high abundance of intra- and peri-tumoral TILs.

4 DISCUSSION

We described and evaluated a deep learning workflow that creates TIL maps to facilitate the quantitative characterization of TILs and map their spatial distributions in H&E WSIs of cancer tissue specimens. Since H&E staining is routinely performed for diagnostic histopathologic evaluation of tissue samples, we developed this workflow to analyze TILs in H&E WSIs that are becoming more commonly available as digital pathology is being more commonly adopted in clinical laboratories. Studies have shown that the host immune system is capable of controlling tumor growth through the activation of adaptive and innate immune surveillance mechanisms (44) and that the spatial context and nature of cellular heterogeneity of the tumor microenvironment are important in cancer prognosis (1, 4, 45, 46). This has led to TILs becoming important in the clinical arena with increasing importance in precision medicine (47–49). Thus, having the ability to quantify TILs in diagnostic H&E WSIs of tissue images is becoming incredibly important as we collectively expand our understanding about tumor immune interactions and their role in disease progression, recurrence, treatment response, and survival.

Therefore, our goal was to develop a robust computational pathology workflow for H&E WSIs to reliably characterize TILs in the tumor microenvironment in a uniform manner. We generated TIL maps to complement traditional microscopic examination so that pathologists and research scientists could interpret the abundance and distribution of TILs alongside the assessment of invasive growth patterns and other histopathologic features across 23 types of cancer. The interest in harnessing the power of TILs to fight cancer continues to grow with advances in immunotherapy, chemoradiation regimens, and other treatment modalities, which has led to important translational cancer research initiatives by the International Immuno-Oncology Biomarker Working Group in creating standardized visual reporting guidelines for pathologists to evaluate TILs in breast cancer and other solid tumors (49–54). Even though pathologists can follow the guidelines and perform qualitative and semi-quantitative assessments of TILs in cancer, the task is highly challenging, subjective, and prone to intra- and interobserver variability. Our results show that the new TIL models are quite

TABLE 4 | Estimated percent TIL area (mean±standard deviation) across WSIs in the dataset TIL-Maps-23.

Cancer Type	TIL Area	Cancer Type	TIL Area	Cancer Type	TIL Area
ACC	1.96 ± 5.15	BLCA	8.60 ± 8.23	BRCA	6.37 ± 7.38
CESC	15.69 ± 11.57	COAD	9.60 ± 6.62	ESCA	11.34 ± 8.45
HNSC	13.54 ± 10.36	KIRC	6.74 ± 8.43	LIHC	7.80 ± 8.27
LUAD	14.29 ± 11.31	LUSC	15.59 ± 10.29	MESO	7.64 ± 8.03
OV	3.94 ± 4.96	PAAD	10.42 ± 7.78	PRAD	5.73 ± 6.52
READ	9.04 ± 6.23	SARC	6.44 ± 9.28	SKCM	13.42 ± 14.46
STAD	15.29 ± 13.24	TGCT	14.51 ± 14.19	THYM	52.89 ± 26.88
UCEC	7.87 ± 8.40	UVM	2.20 ± 2.34	–	–

useful for both qualitative and quantitative evaluation of TILs in WSIs. The TIL maps are also very useful for discerning how much of the tissue samples contain mononuclear lymphoplasmacytic infiltrates and their spatial distribution in individual cancer tissue samples and across several different kinds of cancer from various organ sites. And most importantly, these new models perform better than the model developed in the earlier work, which was limited to 13 different types of cancer (33).

We attribute the better results to the use of state-of-the-art engineered networks and our larger and more diverse training dataset that includes both computer-generated annotations and manual annotations. Having the capability to computationally analyze WSIs to study fascinating patterns of tumor immune interactions with reliable and reproducible methods represents a highly significant opportunity for cancer research to help improve cancer treatment and clinical management. This novel data about the quantity and distribution TILs from H&E WSIs is also important as a biomarker for downstream correlative prognostic studies with clinical, radiologic, laboratory, molecular, and pharmacologic data. Moreover, these kinds of analyses facilitate large-scale research to elucidate deeper mechanistic understanding of the role of tumoral immunity in disease progression and treatment response across both common and rarer types of cancer. Furthermore, the identification and quantification of other image features would allow for the formulation of higher-order relationships to explore the role of TIL infiltrates in cancer immunology with respect to histologic patterns of tumor growth, tumor grade, tumor heterogeneity, cancer recurrence, and metastasis.

In this work, we used three popular network architectures, VGG16, Inception V4, and ResNet-34, to train models for the detection and classification of TILs in tissue images. There are other state-of-the-art networks, such as Xception (55) and EfficientNet (56), which have shown excellent performance in image classification tasks. Our choice of the networks is primarily based on the fact that we have used these selected networks for other projects. Since deep learning is a rapidly evolving field, future work will explore incorporating other deep learning architectures into our workflow to further improve performance and expand the capabilities and applicability of our workflow. We utilized our models to generate TIL maps, referred to here as the *TIL-Maps-23* dataset, in 7983 H&E WSIs in 23 tumor types in the TCGA data repository from among approximately 12,000 diagnostic WSIs from 33 cancer types.

The *TIL-Maps-23* dataset covers 70% of the TCGA cancer types and 67% of the diagnostic TCGA WSIs. Beyond the information embedded in pathology WSIs, the TCGA dataset also includes demographic, clinical, and molecular data derived from multiple molecular platforms, which presents a readily available opportunity to integrate image-derived features, such as TIL-tumor distance distributions or TIL spatial cluster distributions, with rich molecular and clinical data to gain a more comprehensive understanding about tumor immune interactions and the role of TILs as a biomarker. To the best of our knowledge, this is the largest set of TIL maps to date. The list of cancer types included in the dataset is in **Table 5**. In addition to making our models and Tensorflow CNN codes publicly available, we are also releasing the dataset of TIL maps with the intention of motivating translational cancer research and algorithmic development for image analysis in computational pathology.

5 CONCLUSION

The growth of cancer immunotherapy has created tremendous interest in characterizing the abundance and spatial distribution of TILs in cancer tissue samples in order to explore their clinical significance to help guide treatment. As the footprint of Digital Pathology rapidly expands in translational cancer research and clinical laboratories with the recent FDA approval of whole slide imaging for primary diagnostic use, it is widely expected that a large majority of pathology slides will be routinely digitized within the next 5-10 years. In parallel, advances in machine learning, computer vision, and computational hardware resources have led to an increased focus on deep learning-based techniques for segmentation and classification of various features of tissue microanatomy in WSIs, including regions, microanatomic structures, cells, nuclei, and other features. The characterization of TIL infiltrated tissue in WSIs at a resolution of 50 microns by using our methods goes far beyond what can be reproducibly and scalably observed by human beings across hundreds and thousands of tissue samples. Tools and methodologies that augment or enable such

TABLE 5 | The list of cancer types in TIL-Maps-23, the number of WSIs for each cancer type, and the polyserial correlation coefficients for the Inception-V4 model, sorted in descending order.

Cancer Type	# WSIs	Polyserial Correlation Coefficient
Adrenocortical carcinoma (ACC)	323	0.96
Sarcoma (SARC)	255	0.92
Uveal melanoma (UVM)	80	0.91
Head and Neck squamous cell carcinoma (HNSC)	450	0.90
Pancreatic adenocarcinoma (PAAD)	189	0.87
Liver hepatocellular carcinoma (LIHC)	365	0.85
Mesothelioma (MESO)	175	0.82
Prostate adenocarcinoma (PRAD)	403	0.82
Skin cutaneous melanoma (SKCM)	448	0.82
Testicular germ cell tumors (TGCT)	154	0.82
Esophageal carcinoma (ESCA)	156	0.81
Rectum adenocarcinoma (READ)	165	0.81
Breast invasive carcinoma (BRCA)	1068	0.80
Lung squamous cell carcinoma (LUSC)	484	0.79
Thymoma (THYM)	121	0.79
Stomach adenocarcinoma (STAD)	434	0.75
Bladder urothelial carcinoma (BLCA)	386	0.74
Uterine corpus endometrial carcinoma (UCEC)	506	0.74
Colon adenocarcinoma (COAD)	453	0.69
Cervical squamous cell carcinoma (CESC)	268	0.67
Ovarian serous cystadenocarcinoma (OV)	106	0.64
Lung adenocarcinoma (LUAD)	480	0.62
Kidney renal clear cell carcinoma (KIRC)	514	0.60

characterizations can improve the practice of pathology while we march towards realizing the goal of precision oncology.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories can be found below: https://stonybrookmedicine.box.com/v/til-results-new-model.

AUTHOR CONTRIBUTIONS

SA, RG, LH, CC, DS, TK, and JS contributed to the design of the deep learning workflow. SA implemented the workflow and carried out the experiments for evaluation. SA, RG, LH, AS, AR, and JS designed the experimental evaluation. RG, RB, and TZ contributed to the data annotation. SA, RG, JS, and TK led the generation of the TIL-Maps-23 dataset. TK, JS, and RG led the development of the software for training data generation and management and visualization of images and TIL maps. SA, RG, CC, DS, TK, and JS edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Institutes of Health (NIH) and National Cancer Institute (NCI) grants UH3-CA22502103, U24-CA21510904, 1U24CA180924-01A1, 3U24CA215109-02, and 1UG3CA225021-01 as well as generous private support from Bob Beals and Betsy Barton. AR and AS were partially supported by NCI grant R37-CA214955 (to AR), the University of Michigan (U-M) institutional research funds and also supported by ACS grant RSG-16-005-01 (to AR). AS was supported by the Biomedical Informatics & Data Science Training Grant (T32GM141746). This work was enabled by computational resources supported by National Science Foundation grant number ACI-1548562, providing access to the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center, and also a DOE INCITE award joint with the MENNDL team at the Oak Ridge National Laboratory, providing access to Summit high performance computing system. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

This work used the high-performance computing systems provided by the Extreme Science and Engineering Discovery Environment, the Summit high performance computing system at Oak Ridge National Laboratory, and the GPU cluster at the Institute for AI-Driven Discovery and Innovation at Stony Brook University. We acknowledge Dr. John Van Arnam, MD for participation in the annotation effort, and thank Dr. Beatrice Knudsen, MD/PhD and Dr. Kenneth R. Shroyer, MD/PhD for thoughtful input and discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.806603/full#supplementary-material

REFERENCES

1. Mlecnik B, Bindea G, Pagès F, Galon J. Tumor Immunosurveillance in Human Cancers. *Cancer Metastasis Rev* (2011) 30:5–12. doi: 10.1007/s10555-011-9270-7
2. Loi S, Drubay D, Adams S, Pruneri G, Francis P, Lacroix-Triki M, et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. *J Clin Oncol* (2019) 37:JCO.18.01010. doi: 10.1200/JCO.18.01010
3. Angell H, Galon J. From the Immune Contexture to the Immunoscore: The Role of Prognostic and Predictive Immune Markers in Cancer. *Curr Opin Immunol* (2013) 25:261–7. doi: 10.1016/j.coi.2013.03.004
4. Mlecnik B, Tosolini M, Kirilovsky A, Berger A, Bindea G, Meatchi T, et al. Histopathologic-Based Prognostic Factors of Colorectal Cancers are Associated With the State of the Local Immune Reaction. *J Clin Oncol* (2011) 29:610–8. doi: 10.1200/JCO.2010.30.5425
5. Badalamenti G, Fanale D, Incorvaia L, Barraco N, Listì A, Maragliano R, et al. Role of Tumor-Infiltrating Lymphocytes in Patients With Solid Tumors: Can a Drop Dig a Stone? *Cell Immunol* (2019) 343:103753. doi: 10.1016/j.cellimm.2018.01.013
6. Idos G, Kwok J, Bonthala N, Kysh L, Gruber S, Qu C. The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Sci Rep* (2020) 10. doi: 10.1038/s41598-020-60255-4
7. Thorsson V, Gibbs DL, Brown SD, Wolf D, S Bortone D, Ou Yang T, et al. The Immune Landscape of Cancer. *Immunity* (2018) 48:812–30. doi: 10.1016/j.immuni.2018.03.023
8. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The Evaluation of Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer: Recommendations by an International Tils Working Group 2014. *Ann Oncol* (2014) 26:259–71. doi: 10.1093/annonc/mdu450
9. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immuno-Oncology Biomarkers Working Group: Part 2. *Adv Anat Pathol* (2017) 24(6):311–35. doi: 10.1097/PAP.0000000000000161
10. John M, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors. *Adv Anat Pathol* (2016) 24 (6):311–35. doi: 10.1097/PAP.0000000000000161
11. Plesca I, Tunger A, Müller L, Wehner R, Lai X, Grimm MO, et al. Characteristics of Tumor-Infiltrating Lymphocytes Prior to and During Immune Checkpoint Inhibitor Therapy. *Front Immunol* (2020) 11:364. doi: 10.3389/fimmu.2020.00364
12. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q. Deep Learning for Image-Based Cancer Detection and Diagnosis- A Survey. *Pattern Recogn* (2018) 83:134–49. doi: 10.1016/j.patcog.2018.05.014
13. Xing F, Xie Y, Su H, Liu F, Yang L. Deep Learning in Microscopy Image Analysis: A Survey. *IEEE Trans Neural Networks Learn Syst* (2017) 29:4550–68. doi: 10.1109/TNNLS.2017.2766168
14. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

15. Deng S, Zhang X, Yan W, Eric I, Chang C, Fan Y, et al. Deep Learning in Digital Pathology Image Analysis: A Survey. *Front Med* (2020) 14(4):470–87. doi: 10.1007/s11684-020-0782-9
16. Srinidhi CL, Ciga O, Martel AL. Deep Neural Network Models for Computational Histopathology: A Survey. *Med Image Anal* (2020) 67:101813. doi: 10.1007/s11684-020-0782-9
17. Dimitriou N, Arandjelović O, Caie PD. Deep Learning for Whole Slide Image Analysis: An Overview. *Front Med* (2019) 6:264. doi: 10.3389/fmed.2019.00264
18. Eriksen AC, Andersen JB, Kristensson M, Christensen RD, Hansen TF, Kjær-Frifeldt S, et al. Computer-Assisted Stereology and Automated Image Analysis for Quantification of Tumor Infiltrating Lymphocytes in Colon Cancer. *Diagn Pathol* (2017) 12:1–14. doi: 10.1186/s13000-017-0653-0
19. Swiderska-Chadaj Z, Pinckaers H, van Rijthoven M, Balkenhol M, Melnikova M, Geessink O, et al. Learning to Detect Lymphocytes in Immunohistochemistry With Deep Learning. *Med Image Anal* (2019) 58:101547. doi: 10.1016/j.media.2019.101547
20. Garcia E, Hermoza R, Castanon CB, Cano L, Castillo M, Castanneda C. Automatic Lymphocyte Detection on Gastric Cancer Ihc Images Using Deep Learning. In: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. Thessaloniki, Greece: IEEE (2017). p. 200–4.
21. Negahbani F, Sabzi R, Jahromi BP, Firouzabadi D, Movahedi F, Shirazi MK, et al. Pathonet Introduced as a Deep Neural Network Backend for Evaluation of Ki-67 and Tumor-Infiltrating Lymphocytes in Breast Cancer. *Sci Rep* (2021) 11:1–13. doi: 10.1038/s41598-021-86912-w
22. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich, Germany: Springer (2015). p. 234–41.
23. Budginaitė E, Morkūnas M, Laurinavičius A, Treigys P. Deep Learning Model for Cell Nuclei Segmentation and Lymphocyte Identification in Whole Slide Histology Images. *Informatica* (2021) 32:23–40. doi: 10.15388/20-INFOR442
24. Raza SEA, Cheung L, Shaban M, Graham S, Epstein D, Pelengaris S, et al. Micro-Net: A Unified Model for Segmentation of Various Objects in Microscopy Images. *Med Image Anal* (2019) 52:160–73. doi: 10.1016/j.media.2018.12.003
25. Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, et al. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non–Small Cell Lung Cancer. *Clin Cancer Res* (2019) 25:1526–34. doi: 10.1158/1078-0432.CCR-18-2013
26. Jaber M, Beziaeva L, Benz S, Reddy S, Rabizadeh S, Szeto C. A30 Tumor-Infiltrating Lymphocytes (Tils) Found Elevated in Lung Adenocarcinomas (Luad) Using Automated Digital Pathology Masks Derived From Deep-Learning Models. *J Thorac Oncol* (2020) 15:S22. doi: 10.1016/j.jtho.2019.12.059
27. Acs B, Ahmed FS, Gupta S, Wong PF, Gartrell RD, Pradhan JS, et al. An Open Source Automated Tumor Infiltrating Lymphocyte Algorithm for Prognosis in Melanoma. *Nat Commun* (2019) 10:1–7. doi: 10.1038/s41467-019-13043-2
28. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. Qupath: Open Source Software for Digital Pathology Image Analysis. *Sci Rep* (2017) 7:1–7. doi: 10.1038/s41598-017-17204-5
29. Linder N, Taylor JC, Colling R, Pell R, Alveyn E, Joseph J, et al. Deep Learning for Detecting Tumour-Infiltrating Lymphocytes in Testicular Germ Cell Tumours. *J Clin Pathol* (2019) 72:157–64. doi: 10.1136/jclinpath-2018-205328
30. Amgad M, Sarkar A, Srinivas C, Redman R, Ratra S, Bechert CJ, et al. Joint Region and Nucleus Segmentation for Characterization of Tumor Infiltrating Lymphocytes in Breast Cancer. *Med Imaging 2019: Digital Pathol (Int Soc Opt Photonics)* (2019) 10956:109560M. doi: 10.1117/12.2512892
31. Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA: IEEE (2015). p. 3431–40.
32. Le H, Gupta R, Hou L, Abousamra S, Fassler D, Torre-Healy L, et al. Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer. *Am J Pathol* (2020) 190:1491–504. doi: 10.1016/j.ajpath.2020.03.012
33. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep* (2018) 23(1):181–93.e7. doi: 10.1016/j.celrep.2018.03.086
34. Abousamra S, Hou L, Gupta R, Chen C, Samaras D, Kurc T, et al. Learning From Thresholds Fully Automated Classification of Tumor Infiltrating Lymphocytes for Multiple Cancer Types. *CoRR* (2019).
35. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Int Conf Learn Representations (ICLR)* (2015) 1–14.
36. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *IEEE Conf Comput Vision Pattern Recogn (CVPR)* (2016) 770–8. doi: 10.1109/CVPR.2016.90
37. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-V4, Inception-Resnet and the Impact of Residual Connections on Learning. *Thirty-First AAAI Conf Artif Intell* (2017) 4278–84.
38. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. *Proc IEEE Conf Comput Vision Pattern Recogn* (2009) 248–55. doi: 10.1109/CVPR.2009.5206848
39. Unal I. Defining an Optimal Cut-Point Value in Roc Analysis: An Alternative Approach. *Comput Math Methods Med* (2017) 2017:1–14. doi: 10.1155/2017/3762651
40. Youden WJ. Index for Rating Diagnostic Tests. *Cancer* (1950) 3:32–5. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
41. Saltz J, Sharma A, Iyer G, Bremer E, Wang F, Jasniewski A, et al. A Containerized Software System for Generation, Management, and Exploration of Features From Whole Slide Tissue Images. *Cancer Res* (2017) 77:e79–82. doi: 10.1158/0008-5472.CAN-17-0316
42. Drasgow F. Polychoric and Polyserial Correlations. *Wiley StatsRef: Stat Reference Online* (2006). doi: 10.1002/0471667196.ess2014.pub2
43. Olsson U, Drasgow F, Dorans N. The Polyserial Correlation Coefficient. *Psychometrika* (1982) 47:337–47. doi: 10.1007/BF02294164
44. Galon J, Angell HK, Bedognetti D, Marincola FM. The Continuum of Cancer Immunosurveillance: Prognostic, Predictive, and Mechanistic Signatures. *Immunity* (2013) 39:11–26. doi: 10.1016/j.immuni.2013.07.008
45. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat Rev Cancer* (2012) 12:298–306. doi: 10.1038/nrc3245
46. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science* (2006) 313:1960–4. doi: 10.1126/science.1129139
47. Barnes M, Sarkar A, Redman R, Bechert C, Srinivas C. Abstract P5-03-08: Development of a Histology-Based Digital Pathology Image Analysis Algorithm for Assessment of Tumor Infiltrating Lymphocytes in Her2+ Breast Cancer. *Cancer Res* (2018) 78:P5–03–08–P5–03–08. doi: 10.1158/1538-7445.SABCS17-P5-03-08
48. Steele KE, Tan TH, Korn R, Dacosta K, Brown C, Kuziora M, et al. Measuring Multiple Parameters of CD8+ Tumor-Infiltrating Lymphocytes in Human Cancers by Image Analysis. *J Immunother Cancer* (2018) 6(1):20. doi: 10.1186/s40425-018-0326-x
49. Amgad M, Stovgaard ES, Balslev E, Thagaard J, Chen W, Dudgeon S, et al. Report on Computational Assessment of Tumor Infiltrating Lymphocytes From the International Immuno-Oncology Biomarker Working Group. *NPJ Breast Cancer* (2020) 6:1–13. doi: 10.1038/s41523-020-0154-2
50. Kos Z, Roblin E, Kim R, Michiels S, Gallas B, Chen W, et al. Pitfalls in Assessing Stromal Tumor Infiltrating Lymphocytes (Stils) in Breast Cancer. *NPJ Breast Cancer* (2020) 6. doi: 10.1038/s41523-020-0156-0
51. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma *In Situ*, Metastatic Tumor Deposits and Areas for Further Research. *Adv Anat Pathol* (2017) 24:235–51. doi: 10.1097/PAP.0000000000000162
52. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in Melanoma,

Gastrointestinal Tract Carcinomas, Non-Small Cell Lung Carcinoma and Mesothelioma, Endometrial and Ovarian Carcinomas, Squamous Cell Carcinoma of the Head and Neck, Genitourinary Carcinomas, and Primary Brain Tumors. *Adv Anat Pathol* (2017) 24:311–35. doi: 10.1097/PAP.0000000000000161

53. Gupta R, Le H, Arnam J, Belinsky D, Hasan M, Samaras D, et al. Characterizing Immune Responses in Whole Slide Images of Cancer With Digital Pathology and Pathomics. *Curr Pathobiology Rep* (2020) 8:1–16. doi: 10.1007/s40139-020-00217-7
54. Dudgeon S, Wen S, Hanna M, Gupta R, Amgad M, Sheth M, et al. A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study. *J Pathol Inf* (2021) 12:45. doi: 10.4103/jpi.jpi_83_20
55. Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii: IEEE (2017). p. 1251–8.
56. Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning* (2019) 97:6105–14.

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 Artificial Intelligence

ORIGINAL RESEARCH
published: 02 March 2022
doi: 10.3389/frai.2022.826402

Developing a Cancer Digital Twin: Supervised Metastases Detection From Consecutive Structured Radiology Reports

Karen E. Batch[1], Jianwei Yue[1], Alex Darcovich[1], Kaelan Lupton[1], Corinne C. Liu[2], David P. Woodlock[2], Mohammad Ali K. El Amine[3], Pamela I. Causa-Andrieu[2], Lior Gazit[4], Gary H. Nguyen[4], Farhana Zulkernine[1], Richard K. G. Do[2] and Amber L. Simpson[1,5]*

[1] School of Computing, Queen's University, Kingston, ON, Canada, [2] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [3] Department of Graduate Medical Education, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [4] Department of Strategy and Innovation, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [5] Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON, Canada

OPEN ACCESS

Edited by:
George Zaki,
Frederick National Laboratory for Cancer Research (NIH), United States

Reviewed by:
Braja Gopal Patra,
NewYork-Presbyterian, Weill Cornell Medical Center, United States
Hong-Jun Yoon,
Oak Ridge National Laboratory (DOE), United States

***Correspondence:**
Karen E. Batch
karen.batch@queensu.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: 30 November 2021
Accepted: 27 January 2022
Published: 02 March 2022

Citation:
Batch KE, Yue J, Darcovich A, Lupton K, Liu CC, Woodlock DP, El Amine MAK, Causa-Andrieu PI, Gazit L, Nguyen GH, Zulkernine F, Do RKG and Simpson AL (2022) Developing a Cancer Digital Twin: Supervised Metastases Detection From Consecutive Structured Radiology Reports. Front. Artif. Intell. 5:826402. doi: 10.3389/frai.2022.826402

The development of digital cancer twins relies on the capture of high-resolution representations of individual cancer patients throughout the course of their treatment. Our research aims to improve the detection of metastatic disease over time from structured radiology reports by exposing prediction models to historical information. We demonstrate that Natural language processing (NLP) can generate better weak labels for semi-supervised classification of computed tomography (CT) reports when it is exposed to consecutive reports through a patient's treatment history. Around 714,454 structured radiology reports from Memorial Sloan Kettering Cancer Center adhering to a standardized departmental structured template were used for model development with a subset of the reports included for validation. To develop the models, a subset of the reports was curated for ground-truth: 7,732 total reports in the lung metastases dataset from 867 individual patients; 2,777 reports in the liver metastases dataset from 315 patients; and 4,107 reports in the adrenal metastases dataset from 404 patients. We use NLP to extract and encode important features from the structured text reports, which are then used to develop, train, and validate models. Three models—a simple convolutional neural network (CNN), a CNN augmented with an attention layer, and a recurrent neural network (RNN)—were developed to classify the type of metastatic disease and validated against the ground truth labels. The models use features from consecutive structured text radiology reports of a patient to predict the presence of metastatic disease in the reports. A single-report model, previously developed to analyze one report instead of multiple past reports, is included and the results from all four models are compared based on accuracy, precision, recall, and F1-score. The best model is used to label all 714,454 reports to generate metastases maps. Our results suggest that NLP models can extract cancer progression patterns from multiple consecutive reports and predict the presence of metastatic disease in multiple organs with higher performance when compared with a single-report-based prediction. It demonstrates a promising automated approach to

label large numbers of radiology reports without involving human experts in a time- and cost-effective manner and enables tracking of cancer progression over time.

Keywords: digital twins, cancer, metastases, machine learning, radiology, natural language processing (NLP), convolutional neural network (CNN), recurrent neural network (RNN)

INTRODUCTION

Healthcare is increasingly tailoring treatments to the needs of individual patients, an approach known as personalized medicine. To achieve this, the engineering concept of Digital Twins is proposed to develop virtual patients that can be computationally treated to find optimal treatment strategies (Björnsson et al., 2020). These models are *in silico* high-resolution representations of an individual based on available molecular, physiological, and other data, which has the potential for vast improvements in patient care (Björnsson et al., 2020; Croatti et al., 2020). Personalized medicine stems from the assumption that refined mathematical models of patients will result in more precise and effective medical interventions (Bruynseels et al., 2018). This approach uses fine-grained information on individuals to identify deviations from the individual's normal to develop or select treatment focusing on a patient's individual clinical characterization such as diversity of symptoms, severity, and genetic traits, as well as environmental and lifestyle factors over time (Bruynseels et al., 2018; National Institutes of Health, 2020). Previously believed to be impossible, digital models of patients are becoming a reality with the wide-spread availability of molecular as well as other clinical data and substantial increase in computational power.

Much of the information contained in a medical record is in the form of free-text or semi-structured text data from clinical notes. Radiology reports in particular capture information critical to the treatment and management of cancer patients. Therefore, the development of a Cancer Digital Twin from routinely acquired radiology reports offers a unique opportunity to study cancer response and progression throughout a patient's journey. Manual extraction of data from CT reports requires substantial domain expertise and is prohibitively time-consuming to perform across all cancer patients. As a result, little is known about metastatic progression outside of cancer clinical trials, where response rates are most typically calculated. Data extraction from radiology reports by natural language processing (NLP) is now increasingly performed (Pons et al., 2016), including in large populations of patients with cancer, so the potential application to Digital Twins is attractive. To date, the application of NLP to radiology reports for the classification of metastatic disease has been limited to bone and brain metastases (Senders et al., 2019; Groot et al., 2020) or generalized cancer outcomes (Kehl et al., 2019). We previously presented an ensemble voting model to detect metastases from individual radiology reports for different organs using NLP (Do et al., 2021). This model considered only single reports for a given patient using standard term frequency-inverse document frequency (TF-IDF) techniques. The application of NLP to large-scale labeling of CT reports would facilitate the development of a Digital Twin and offer new insights into patterns of metastatic progression across cancer sites. Identification of such patterns will allow for the development of the high-resolution representation required for virtual patients to be effective when modeling a cancer patient's disease progression over time. This coupled with the generation of a large database of patterns of spread, early detection, and prediction of how an individual patient will progress will be possible.

Time is a critical aspect of medical data. *When* an event occurs, or the order of events that occurred, is as important as the events themselves. Studies have been conducted to incorporate the information contained in free-text clinical notes with temporal data points for ICU-related tasks (Caballero Barajas and Akella, 2015; Khadanga et al., 2019; Huang et al., 2020). Clinical notes have high-dimensionality and are sparsely recorded, creating a computational challenge compared with traditional structured time-series data (Huang et al., 2020) such as ICU data. There has been little investigation into using radiology reports throughout a patient's cancer treatment to improve the detection of metastatic spread in radiology reporting. Our research aims to fill this gap and develop a map of disease spread in individual patients over time.

In this paper, we extend our model presented in Do et al. (2021) to incorporate consecutive, multi-report prediction using several convolutional and recurrent neural network (RNN) approaches to improve detection accuracy. We present three NLP models that generate weak labels for semi-supervised classification of CT reports when exposed to multiple consecutive reports throughout a patient's treatment history.

MATERIALS AND METHODS

Dataset Description

The data for this study consists of consecutive radiology reports for CT examinations of the chest, abdomen, and pelvis, performed between July 1, 2009, and March 26, 2021, at Memorial Sloan Kettering Cancer Center (MSKCC). Only reports following the departmental standardized structured template introduced in July 2009 were included; any reports which deviated from the template were omitted from analysis. The complete dataset includes 714,454 reports. Each report consists of "findings" for 13 individual organs (lungs, pleura, thoracic nodes, liver, spleen, adrenal glands, renal, abdominopelvic nodes, pelvic organs, bowel, peritoneum, bones, and soft tissues) and an overall "impression" field. The reports in this dataset are semi-structured as shown in **Figure 1**. In the findings section, the radiologists report observations using free text under individual headings for each organ (e.g., lung, liver). Important findings are summarized using free text in

FINDINGS:

LUNGS: New small anterior pneumothoraces bilaterally. Patient is status post left upper lobe, left lower lobe, right lower lobe, right upper lobe wedge resections. There is a small postoperative cavity within the right upper lobe that measures 2.4 x 2.4 cm and is surrounded by adjacent consolidation. Right lower lobe postoperative changes are identified. Within the left lower lobe, there is rounded consolidation, likely postoperative in etiology. Minimal left upper lobe groundglass opacity is identified. There is a persistent left upper lobe pulmonary nodule that measures 0.8 cm. Additional previously seen pulmonary nodules appear to have been resected.

PLEURA/PERICARDIUM: No pericardial fluid. Trace bilateral pleural effusions.

THORACIC NODES: Subcentimeter mediastinal lymph nodes. Surgical clips are present within the left axilla.

ADRENAL GLANDS: unremarkable

BONES/SOFT TISSUES: There are no suspicious osseous lesions.

OTHER: There are foci of air within the left chest wall and posterior to the right scapula. The patient is status post cholecystectomy.

IMPRESSION: Since outside CT dated February DATE,
1. Left upper lobe pulmonary nodule that measures 0.8 cm.
2. Status post bilateral pulmonary wedge resections with postoperative changes and small anterior pneumothoraces bilaterally.
3. Trace bilateral pleural effusions.
4. Subcentimeter mediastinal lymph nodes

FIGURE 1 | Example report of a chest CT following the template implemented in July 2009. The "Findings" section contains observations specific to each organ sites, while the "Impression" section can contain observations pertaining to any organ.

the impression section at the end of the report. Of note, non-observations are often as important as observations. This is to say that if there are "no changes" reported for a certain organ, it could mean that in a previous report metastases were identified, and they remain as they were. It could also mean that there are no lesions of interest. Standardized reporting improves clarity and consistency of clinical reports and is increasingly preferred compared to free-text reports (Renshaw et al., 2018).

Three of the 13 available organs were selected for the study: lungs, liver, and adrenal glands. The lungs and liver were selected as the most common sites of metastases while adrenal glands are one of the least common sites. Subsets of the complete dataset were annotated for ground truth by a radiologist. Each report in the ground truth set was labeled for the presence or absence of metastases. For each patient in the ground truth set, five radiologists were instructed to read all reports available before deciding the presence or absence of metastases at each time point. If after reviewing all the available reports, the radiologists were unsure about the presence or absence of metastases in a particular patient, they were instructed to skip those reports. This resulted in the following number of annotated reports: 7,732 in the lung metastases dataset from 867 individual patients; 2,777 in the liver metastases dataset from 315 patients; and 4,107 in the adrenal metastases dataset from 404 patients. Annotated reports were used to train a single-report ensemble prediction model for each organ. Once the model accuracy plateaued, the dataset was deemed to be of adequate size for that organ, resulting in differing quantities of annotated reports for each organ. Each of the three datasets (lung metastases, liver metastases, adrenal metastases) were randomly split into training (70%), testing (15%), and validation (15%) sets (see **Table 1**). All models were trained, tested, and validated using the same data splits to ensure accurate performance comparison at each stage.

Data Preprocessing

The raw text data consisted of organ observations from the report, each associated with a patient. To transform the data into a format for multi-report analysis, individual reports were grouped by patient and ordered chronologically from oldest to newest. For each report r_t of a patient, all previous reports ($t = 0, 1, \ldots, n$, where n is the target report) were concatenated into a single document. For example, if the target report is the first report associated with the patient, the resulting document would consist only of this report. If the target report is the third report

TABLE 1 | Model performance results for the baseline single-report metastases prediction model and the three novel multi-report metastases prediction models.

Model	Metric	Training			Testing			Validation		
		Lung (*n* = 5,413)	**Liver (*n* = 1,943)**	**Adrenal (*n* = 2,874)**	**Lung (*n* = 1,160)**	**Liver (*n* = 417)**	**Adrenal (*n* = 617)**	**Lung (*n* = 1,160)**	**Liver (*n* = 417)**	**Adrenal (*n* = 616)**
TF-IDF ensemble model (Baseline)	Accuracy	99.69% (±0.15%)	**99.95% (±0.10%)**	99.23% (±0.32%)	92.33% (±1.53%)	90.12% (±2.86%)	96.60% (±1.43%)	93.80% (±1.39%)	92.50% (±2.53%)	96.10% (±1.53%)
	Precision	0.9977 (±0.00)	**1.0000 (±0.00)**	**1.0000 (±0.00)**	0.8553 (±0.02)	0.9060 (±0.03)	**0.9444 (±0.02)**	0.9080 (±0.02)	0.8990 (±0.03)	**1.0000 (±0.00)**
	Recall	0.9833 (±0.00)	0.9983 (±0.00)	0.8932 (±0.01)	0.6733 (±0.03)	0.7794 (±0.04)	0.4595 (±0.04)	0.6860 (±0.03)	0.8310 (±0.04)	0.5000 (±0.04)
	F1-score	0.9904 (±0.00)	**0.9991 (±0.00)**	0.9436 (±0.01)	0.7535 (±0.02)	0.8379 (±0.04)	0.6182 (±0.04)	0.7815 (±0.02)	0.8637 (±0.03)	0.6667 (±0.04)
Simple CNN	Accuracy	99.93% (±5.21%)	99.85% (±7.59%)	**100% (±0.00%)**	**97.41% (±0.91%)**	**98.56% (±1.14%)**	**99.03% (±0.77%)**	96.64% (±1.04%)	98.56% (±1.14%)	99.51% (±0.55%)
	Precision	0.9956 (±0.00)	0.9950 (±0.00)	**1.0000 (±0.00)**	**0.9526 (±0.01)**	0.9851 (±0.01)	0.9429 (±0.02)	**0.9526 (±0.01)**	0.9746 (±0.02)	0.9592 (±0.02)
	Recall	**1.0000 (±0.00)**	**1.0000 (±0.00)**	**1.0000 (±0.00)**	0.8960 (±0.02)	0.9706 (±0.02)	**0.8919 (±0.02)**	0.8564 (±0.02)	0.9746 (±0.02)	**0.9792 (±0.01)**
	F1-score	0.9978 (±0.00)	0.9975 (±0.00)	**1.0000 (±0.00)**	0.9234 (±0.02)	0.9778 (±0.01)	**0.9167 (±0.02)**	0.8920 (±0.02)	0.9746 (±0.02)	0.9691 (±0.02)
Augmented CNN	Accuracy	**99.98% (±0.04%)**	99.90% (±0.14%)	99.97% (±0.06%)	**97.41% (±0.91%)**	**98.56% (±1.14%)**	98.87% (±0.83%)	96.81% (±1.01%)	**99.04% (±0.94%)**	**99.68% (±0.45%)**
	Precision	**0.9989 (±0.00)**	0.9966 (±0.00)	0.9952 (±0.00)	0.9388 (±0.01)	0.9710 (±0.02)	0.9167 (±0.02)	0.9467 (±0.01)	**0.9831 (±0.01)**	**0.9792 (±0.01)**
	Recall	**1.0000 (±0.00)**	**1.0000 (±0.00)**	**1.0000 (±0.00)**	**0.9109 (±0.02)**	**0.9853 (±0.01)**	**0.8919 (±0.02)**	0.8511 (±0.02)	**0.9831 (±0.01)**	**0.9792 (±0.01)**
	F1-score	**0.9994 (±0.00)**	0.9983 (±0.00)	0.9976 (±0.00)	**0.9246 (±0.02)**	0.9781 (±0.01)	0.9041 (±0.02)	**0.8964 (±0.02)**	**0.9831 (±0.01)**	**0.9792 (±0.01)**
Bidirectional LSTM	Accuracy	97.97% (±0.38%)	99.23% (±0.39%)	99.72% (±0.19%)	96.66% (±1.03%)	**98.56% (±1.14%)**	98.70% (±0.89%)	**97.16% (±0.96%)**	98.32% (±1.23%)	99.03% (±0.77%)
	Precision	0.9052 (±0.01)	0.9798 (±0.01)	0.9660 (±0.01)	0.8465 (±0.02)	**0.9853 (±0.01)**	0.8919 (±0.02)	0.8404 (±0.02)	0.9661 (±0.02)	0.9375 (±0.02)
	Recall	0.9366 (±0.01)	0.9873 (±0.00)	0.9803 (±0.01)	0.8976 (±0.02)	0.9781 (±0.01)	**0.8919 (±0.02)**	**0.9054 (±0.02)**	0.9702 (±0.02)	0.9375 (±0.02)
	F1-score	0.9206 (±0.01)	0.9835 (±0.01)	0.9731 (±0.01)	0.8713 (±0.02)	**0.9817 (±0.01)**	0.8919 (±0.02)	0.8717 (±0.02)	0.9682 (±0.02)	0.9375 (±0.02)

Organ datasets are split into three subsets for training (70%), testing (15%), and validation (15%). The n values correspond to the size of the sets. The highest values for each organ in each performance metric are bolded. Values in parentheses are within the 95% confidence interval rounded to two decimal places.

associated with the patient, the resulting document consists of the patient's first, second, and third reports concatenated together in chronological order. The radiology reports often included dates and lesion measurements. These text patterns were identified using regular expressions and replaced with the text "date" and "measurement", respectively. This is done to shrink the size of the vocabulary as well as to capture the higher-level concept of a date or measurement being present in the text. Since the measurements themselves were not included in any analysis, it was beneficial to remove them from the vocabulary space. Target values (i.e., labels) were encoded from "Yes" and "No" values to binary values 1 and 0, respectively.

Model Development

Three models were developed to predict the presence of metastases over time in each of the three target organs namely, a simple convolutional neural network (CNN), a CNN augmented with an attention layer (referred to as the Augmented CNN), and a bidirectional Long Short-Term Memory (Bi-LSTM) model. Convolutional neural network extract spatial features which allows for the maintenance of context when analyzing text in NLP applications. Adding the attention layer to the CNN allows for increased explainability and allows the model to learn and give higher importance to features later in the sequence. The Bi-LSTM was selected because it learns context of information and the sequence of patterns by traversing the text in two directions to create superior text embedding. For the purposes of this study, we combine multiple consecutive reports of a patient consisting of observations made by a radiologist into a single document which is used as the input to the model. To evaluate the benefit of looking at multiple consecutive reports compared to only one report, the single-report model described previously by our group (Do et al., 2021) was used as a baseline. The models are compared based on the following metrics: accuracy, precision, recall, and F_1 measure. F_1 measure is considered the most important metric because F_1 is the harmonic mean of precision and recall and provides a better measure of incorrectly classified cases than accuracy. In the cases of identifying potential metastases, the cost of missing positive cases (false negatives) is much greater than the cost of false positives, which is reflected in the F_1 score. F_1 also mitigates the effect of imbalanced class distribution, which can be masked behind accuracy scores.

Baseline Model

We previously presented an ensemble voting model to detect metastases from individual radiology reports for different organs using NLP (Do et al., 2021). This model is used as the baseline for performance evaluation of the multi-report prediction models presented in the current paper. Briefly, this baseline model processes the raw text data using a TF-IDF method. The processed data are passed through an ensemble voting model built with a logistic regression (LR) model, a support vector machine (SVM), a random forest (RF) model, and an extreme gradient boosting (XGBoost) model. The specifications for each model are given in the following paragraph. Ensemble models use a "voting" strategy to select the best prediction based on predictions made by multiple underlying statistical models. Voting can be done using either a hard vote counter or a soft vote counter. In hard voting, the final classification is made based on a strict count of the predictions made by the underlying models, while soft voting gives higher importance to certain models. In soft voting, the models in the ensemble are ranked using a simple weighting algorithm to determine the relative importance given to each model's predictions. The algorithm compares the accuracy, precision, and recall metrics of all models on the training set to assign the weights. These values are used such that the best-performing classification model's prediction is given the highest importance when tallying the votes. In addition to this ranked weighting, the confidence values for each model's prediction are leveraged in making the final prediction. Our model uses soft voting. Importance calculations were done for each organ to better optimize model performance by location. This means that the weights assigned to the individual models for predicting liver metastases may be different from those for predicting lung metastases.

The LR model is configured with a regularization strength of 15.0, it uses balanced class weighting, which automatically adjusts the weights inversely proportional to class frequencies in the input data. It uses the Newton-CG optimization algorithm solver to handle multinomial loss in the multiclass prediction problem. The SVM uses a linear kernel and all other default parameters. The RF model is built to have 2,000 trees with bootstrapping and the maximum number of features used when building a tree is set to the square root of the number of features seen during *fit*. The XGBoost model uses default configuration.

Simple Convolutional Neural Network

Text data from the radiology reports must be converted into a numeric vector representation to be used as inputs to machine learning models. Recent studies (Zuccon et al., 2015; Zhao and Mao, 2018; Verma et al., 2021) have shown different text encoding approaches having different complexities and ability to represent contextual information. One of the popular approaches is called word embedding, which includes word context and transforms each word to a numeric vector capturing semantic information (Ghannay et al., 2016). The transformation allows different words having similar meanings to have vector representations that are close together in the embedding space. For the convolutional neural network models, the text data is transformed using the Tokenizer from TensorFlow (Abada et al., 2015). Tokenizer creates a vocabulary of all the unique terms in the training corpus and allows for vectorization of the text corpus by turning each document into a sequence of integers, where each integer is the index of a token in a dictionary. All punctuation is removed from the text when it is processed through Tokenizer. When any text is processed by the Tokenizer, only the known words are processed while the unknown words are ignored. This processed data is then fed as input to the convolutional layers of the model.

The idea behind convolutions in computer vision is to learn filters that transform adjacent pixels into single values. A CNN for NLP learns which combinations of adjacent words are associated with a given concept, meaning they can augment the existing techniques by leveraging the representation of language

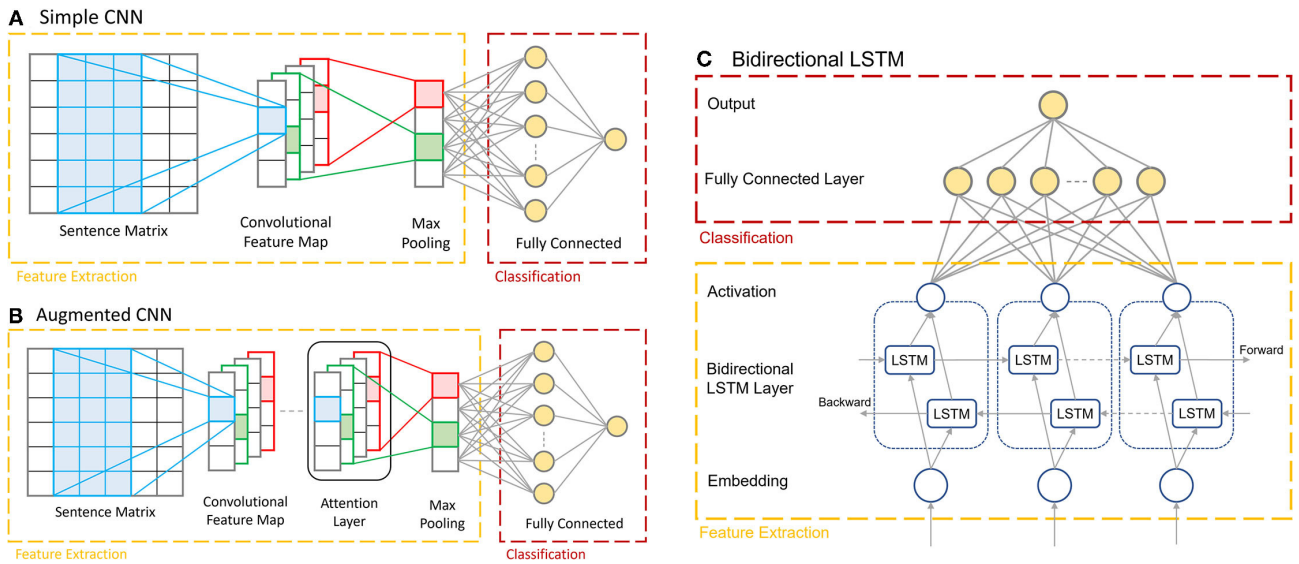


FIGURE 2 | The architectures of the three multi-report prediction models. **(A)** The Simple CNN architecture consisting of the embedding layer, 1D convolutional layer, max pooling layer, and dense layers. **(B)** The Augmented CNN, consisting of the same architecture as the Simple CNN with an added Attention Layer before the max pooling layer. **(C)** The Bi-LSTM, with the two LSTM layers processing inputs in opposite directions.

to learn which phrases in clinical text are relevant for a given medical condition. In a CNN, a text is first represented as a sequence of word embeddings in which each word is projected into a distributed representation. Words that occur in similar concepts are trained to have similar embeddings, meaning misspellings, synonyms, and abbreviations of an original word learn similar embeddings, leading to similar results. Therefore, a database of synonyms and common misspellings is therefore not required.

Embedded text is the input to the convolutional layer. Convolutions detect a signal from a combination of adjacent inputs, and each convolution operation applies a filter of trained parameters to an input-window of specific width. A filter is applied to every possible word window in the input to produce a feature map. The feature map is then reduced using a pooling operation. It is possible to combine multiple convolutions per length and of different lengths to evaluate phrases from 1 to 5 words long, for example. A final fully connected feed forward layer helps compute the probability of whether the text refers to a patient with a certain disease condition.

The CNN model (**Figure 2A**) is built using Keras (Chollet, 2015), which consists of an embedding layer with an embedding dimension of 50, a 1D convolutional layer with a filter size of 64, a kernel size of 3 and using ReLU activation, followed by a global max pooling layer, and finally two fully connected dense layers containing 10 nodes and 1 node, respectively. The final single output node generates a binary decision of whether the input corresponds to the presence of a certain disease condition or not. The penultimate layer uses ReLU activation, and the ultimate layer uses sigmoid activation to make the final prediction. The model is optimized using the ADAM optimizer and the binary cross-entropy loss function.

Augmented CNN

The Augmented CNN (**Figure 2B**) consists of the same architecture as the Simple CNN model as describe above with one added layer: the Keras Sequential Self Attention layer (SeqSelfAttention). This layer implements an attention mechanism when processing sequential data to learn important text embedding and attend to that information (increase weight values) when extracting data features. It is added after the convolutional layer in the model, and its output is fed as the input to the global max pooling layer. The attention layer is configured to use multiplicative attention, an attention width of 1, and uses sigmoid attention activation. The remaining layers of the model are the same as in the Simple CNN.

Bi-Directional LSTM

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks are a powerful type of RNN. One of the main limitations of the basic RNNs is that they lose critical information when dealing with long sequences. Long Short-Term Memories are explicitly designed to avoid such problems and retain information from long sequences of data to learn dependencies in data that are far apart. Thus, the model can remember and also forget certain information it has previously seen. These models consist of a cell state c_{t-1} (i.e., the memory of the network) and a hidden state h_{t-1} (used to make predictions) with three gates that allow the gradient to flow unchanged. The three gates are a forget gate, an input gate, and an output gate. The *forget gate* determines what information are going to be thrown away from the cell state. This gate is essentially a sigmoid function, taking hidden state h_{t-1} and data x_t as input, and outputs a number between 0 and 1 for each element in the cell state. A 0 means to completely throw away the information while

a 1 means keep all the detail of that element. The *input gate* determines what new information is going to be retained in the cell state (update the cell state from c_{t-1} to $\tilde{c}_t$). This layer has two parts: a sigmoid layer and a tanh layer. The sigmoid layer takes hidden state h_{t-1} and data x_t as input and determines which values to update by assigning a number between 0 and 1 to each element computed in the tanh layer. The tanh layer transforms the data x_t and hidden state h_{t-1} to a number between -1 and 1. Next, the product of both layers yields the update to the cell state. The cell state is updated by multiplying the output from the forget gate elementwise, ensuring only critical information can flow down the sequence. Next, the results from the input gate are added elementwise to the cell state. This completes the cell state update, yielding c_t. We use the freshly updated cell state c_t to update the hidden state. In the *output gate*, it first passes the hidden state h_{t-1} and data x_t through a sigmoid layer. Then, c_t is passed through a tanh layer and these results are multiplied together, yielding the new hidden state h_t.

Bidirectional RNN models are two combined RNN models, one model processing data sequentially from beginning to end, while the other received input data in the opposite direction, from end to start. These models perform data analysis simultaneously and their results (predictions) are combined and passed to the dense layers.

For the Bi-LSTM (**Figure 2C**), a self-created dictionary is used for the word embedding. Each unique word in the reports is extracted and sorted in the order of alphabet. Each word is then assigned an index, reserving the first two indices for padding (0) and unknown (1) values, respectively. This model encodes the input documents as vectors consisting of values corresponding to the word's index in the vocabulary dictionary. This means that each input vector for this model depends on the length of the original report, which is variable. The documents are not padded initially but will be padded to the same length for each batch while they are passing through the data generator function. This data is then passed through into the two LSTM layers, processing the data in opposite directions.

The Bi-LSTM was also developed using Keras. The first layer is an embedding layer with input dimension equal to the size of the vocabulary and an output size of 64. This is followed by the Bidirectional LSTM layer provided in the Keras library, which uses the tanh activation function and sigmoid recurrent activation. The output dimension of this layer is 64. The final two layers of the model are similar to those found in both the Simple and Augmented CNN models; the penultimate layer is a fully connected dense layer with 64 nodes and ReLU activation, while the ultimate layer is a dense layer with one output node with no activation.

RESULTS

Metastatic disease was present in 16.6% (1,287/7,733) of the reports in the lung dataset, 30.5% (848/2,777) of the reports in the liver dataset, and in 7.1% (291/4,107) of the adrenal gland dataset. These distributions were consistent in the training, testing, and validation sets. Prediction accuracies exceeded 96% across all organs and all models during validation, with the lowest accuracy being the Simple CNN predicting the presence of lung metastases. The F_1 scores are especially promising, showing balanced precision and recall scores in all models. The F_1 scores demonstrate that the Augmented CNN is the most balanced model, though all models' F_1 scores for the lung dataset were below 0.90. The F_1 scores for lung metastases detection are consistently the lowest, though always scoring above 0.87. The performance metrics of the three models on the validation dataset are presented in **Table 1**. We compare these results with the performance of the baseline model outlined in section Baseline Model, which predicts the presence of metastases from single reports, in contrast to the three deep learning models, which include information from previous reports concatenated in chronological order.

The training and testing results for the baseline TF-IDF Ensemble Voting model are as follows: in training, the model scored 99.69 $\pm$ 0.001%, 0.9977, 0.9833, and 0.9904 (accuracy, precision, recall, F_1 score) on the lung dataset, 99.95%, 1.00, 0.9983, and 0.9991 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.23%, 1.00, 0.8932, and 0.9436 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. In testing, the model scored 92.33%, 0.8553, 0.6733, and 0.7535 (accuracy, precision, recall, F_1 score) on the lung dataset, 90.12%, 0.9060, 0.7794, and 0.8379 (accuracy, precision, recall, F_1 score) on the liver dataset, and 96.60%, 0.9444, 0.4595, and 0.6182 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. During validation, the model scored 93.80%, 0.9080, 0.6860, and 0.7815 (accuracy, precision, recall, F_1 score) on the lung dataset, 92.50%, 0.8990, 0.8310, and 0.8637 (accuracy, precision, recall, F_1 score) on the liver dataset, and 96.10%, 1.00, 0.5000, and 0.6667 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. The results from the baseline model are also presented in **Table 1**.

The complete results for the Simple CNN on each dataset are as follows: in training, the model scored 99.93%, 0.9956, 1.00, and 0.9978 (accuracy, precision, recall, F_1 score) on the lung dataset, 99.85%, 0.9950, 1.00, and 0.9975 (accuracy, precision, recall, F_1 score) on the liver dataset, and 100%, 1.00, 1.00, and 1.00 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. In testing, the model scored 97.41%, 0.9526, 0.8960, and 0.9234 (accuracy, precision, recall, F_1 score) on the lung dataset, 98.56%, 0.9851, 0.9706, and 0.9778 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.03%, 0.9429, 0.8919, and 0.9167 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. During validation, the model scored 96.64%, 0.9526, 0.8564, and 0.8920 (accuracy, precision, recall, F_1 score) on the lung dataset, 98.56%, 0.9746, 0.9746, and 0.9746 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.51%, 0.9592, 0.9792, and 0.9691 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. The results from the Simple CNN model are also presented in **Table 1**.

The complete results for the Augmented CNN on each dataset are as follows: in training, the model scored 99.98%, 0.9989, 1.00, and 0.9994 (accuracy, precision, recall, F_1 score) on the lung dataset, 99.90%, 0.9966, 1.00, and 0.9983 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.97%, 0.9952, 1.00, and

0.9976 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. In testing, the model scored 97.41%, 0.9388, 0.9109, and 0.9246 (accuracy, precision, recall, F_1 score) on the lung dataset, 98.56%, 0.9710, 0.9853, and 0.9781 (accuracy, precision, recall, F_1 score) on the liver dataset, and 98.87%, 0.9167, 0.8919, and 0.9041 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. During validation, the model scored 96.81%, 0.9467, 0.8511, and 0.8964 (accuracy, precision, recall, F_1 score) on the lung dataset, 99.04%, 0.9831, 0.9831, and 0.9831 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.68%, 0.9792, 0.9792, and 0.9792 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. The results from the Augmented CNN model are also presented in **Table 1**.

The complete results for the Bi-LSTM on each dataset are as follows: in training, the model scored 97.97%, 0.9052, 0.9366, and 0.9206 (accuracy, precision, recall, F_1 score) on the lung dataset, 99.23%, 0.9798, 0.9873, and 0.9835 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.72%, 0.9660, 0.9803, and 0.9731 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. In testing, the model scored 96.66%, 0.8465, 0.8976, and 0.8713 (accuracy, precision, recall, F_1 score) on the lung dataset, 98.56%, 0.9853, 0.9781, and 0.9817 (accuracy, precision, recall, F_1 score) on the liver dataset, and 98.70%, 0.8919, 0.8919, and 0.8919 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. During validation, the model scored 97.16%, 0.8404, 0.9054, and 0.8717 (accuracy, precision, recall, F_1 score) on the lung dataset, 98.32%, 0.9661, 0.9702, and 0.9682 (accuracy, precision, recall, F_1 score) on the liver dataset, and 99.03%, 0.9375, 0.9375, and 0.9375 (accuracy, precision, recall, F_1 score) on the adrenal gland dataset. The results from the Bi-LSTM model are also presented in **Table 1**.

DISCUSSION

We developed three novel models for detecting metastatic disease in three separate organs using NLP over multiple consecutive radiology reports. Both the CNN models and the Bi-LSTM model demonstrated high performance in accomplishing this task. Our results demonstrate the added predictive power of exposing an NLP model to historical patient information. Indeed, F_1 score increased from 0.7815, 0.8637, and 0.6667 to 0.8964, 0.9831, and 0.9792 in the lung, liver, and adrenal gland datasets, respectively, when multiple reports were considered. Accuracy, precision, and recall all improved with the multi-report model. The best-performing model—the Augmented CNN—achieved the highest F_1 scores at all three organ sites during validation. Through the model development process, the model performance remained consistent through training, testing, and validation. During training, the models were exposed to records with varying number of concatenated reports, meaning the models have been trained to detect metastases with varying amounts of available information so the models can be used at any point within a patient's course of treatment.

Performance on the lung dataset was lowest for all three models, with F_1 scores of 0.8920, 0.8964, and 0.8717 achieved by the Simple CNN, the Augmented CNN, and the Bi-LSTM, respectively. This is likely because the lung can be subject to a large variety of ailments, such as infections, some of which overlap in appearance with metastatic disease because they appear to radiologists as pulmonary nodules. In analyzing the model's decision-making by extracting the most predictive terms from the vocabulary, it was identified that the presence of measurements in a report were highly indicative of the presence of a metastatic disease. This is not surprising since radiologists commonly rely on measurements to document response to treatment in their report. Use of this feature is excellent in detecting metastases in the liver and adrenal glands, however there are more types of lesions that are measured for the lung, including benign lung nodules. Predicting based on the presence of measurements in the case of lung metastases therefore, results in higher frequency of false positives. The F_1 scores for liver and adrenal metastases predictions on the validation sets exceeded 0.9691, and the Bi-LSTM was only slightly lower at 0.9375 when predicting the presence of adrenal metastases.

There are many papers that describe the usage of NLP for text mining clinical notes, linking events described in notes to time series data (typically for prediction of mortality or length of stay) (Caballero Barajas and Akella, 2015; Khadanga et al., 2019; Huang et al., 2020). A recent study used NLP and deep learning for case-level context for classifying pathology reports has demonstrated the success of CNNs, RNNs, and attention models (Gao et al., 2020), such as those presented in our study. While presenting similar models, this previous study focused on several multi-class classification problems, while our study focuses on the binary classification of the presence or absence of metastatic disease. Both studies demonstrate the benefit of capturing case-level context from consecutive reports compared with single-report prediction, however our models demonstrate higher F1 scores overall. To our knowledge, ours is the first demonstration of multi-report detection in consecutive radiology reports. Specifically, we consider the order that metastases appear for each patient by concatenating reports but do not consider the length of time between metastatic events. Given the overall high performance of our models, factoring in the actual time may not be warranted for simple detection of labels. As we advance our methods to metastatic phenotype identification, the goal of our cancer digital twin, time will likely be an important factor. When included in a digital twin, the series of metastatic cancer labels will show how the individual's state is changing over time and construct the high-resolution representations required. We were the first to demonstrate the benefit of semi-structured narrative reports in the largest study using NLP for identifying metastatic disease (Do et al., 2021), combined with the new models proposed in the current study, we are unlocking the potential of using cancer digital twins for anticipating cancer response and progression.

Our study is not without limitations. It is important to note that the human annotators had access to slightly different information compared to what the models had access to at the time of prediction. The human annotators had access to both historical and future reports, while the models only had the text from previous reports concatenated to the target report to

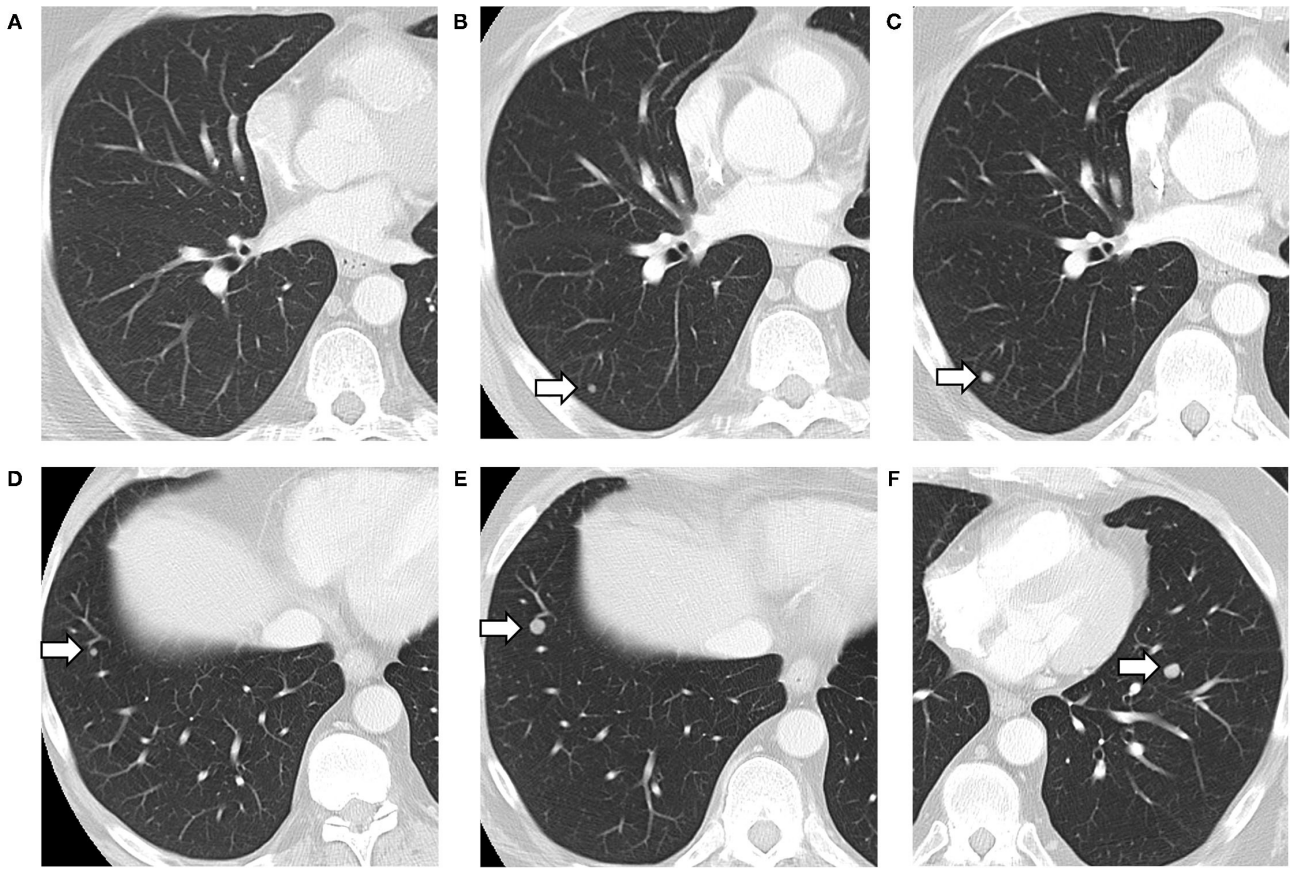


FIGURE 3 | Lung metastases developing on serial CT scans (arrows). Axial images of the lung from three consecutive CT scans, showing the development of a lung nodule, in the posterior right lower lobe **(A–C)**. A separate nodule in the anterior right lower lobe also grew between the second **(D)** and third CT scan **(E)**. A third nodule appeared in the left lower lobe on the third scan only **(F)**. The first CT was negative for metastasis **(A)**, with the text in the "Lungs" section of the findings reading "No suspicious findings." The model predicted correctly that there were no metastases described with 100% confidence. The second CT **(B,D)** was labeled as positive for metastases by the radiologist who had access to all three scans, but negative by the CNN (with a confidence of 99.60%) which only had reports for the first two. The third CT **(C,E,F)** was labeled as positive by both radiologists and the CNN (with 100% confidence).

make their predictions. This resulted in false negative errors (example provided in **Figure 3**), though the model was able to correct the prediction for later reports. The implications of this depend on the use-case of the model. In the case where "future" reports (with respect to the target report) are available, such as in a retrospective study of disease pattern, exposing the model to these future reports would be desired. However, if the use-case is predicting the presence of metastases in a patient currently undergoing treatment and all reports to-date are presented to the model, the model is not missing any information.

In conclusion, the multi-report NLP prediction models presented in this paper generate more reliable weak labels of radiology reports compared with a single-report prediction model. The success of digital cancer twins relies heavily on the access of high-resolution representations of individual cancer patients over time. The ability to automatically generate accurate labels of metastatic disease from radiology reports will improve the viability of these digital twins, while enabling recognition of disease progression patterns through the availability of such a large database of generated weak labels. This will allow for earlier detection of potential progression of disease in individual patients allowing for more successful intervention during disease management.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

KB, RD, LG, and AS: guarantors of integrity of entire study. RD, KB, KL, PC-A, FZ, and AS: literature research. RD and PC-A: clinical studies. KB, KL, LG, and AS: experimental studies.

KB, KL, PC-A, LG, FZ, and AS: statistical analysis. All authors study concepts, study design or data acquisition or data analysis, interpretation, manuscript drafting or manuscript revision for important intellectual content, approval of final version of submitted manuscript, agrees to ensure any questions related to the work are appropriately resolved, and manuscript editing, and contributed to the article and approved the submitted version.

FUNDING

Supported in part by National Institutes of Health/National Cancer Institute Cancer Center Support Grant P30 CA008748 and the New Frontiers in Research Fund Exploration Grant from Social Sciences and Humanities Research Council of Canada, and the Canada Research Chairs program.

REFERENCES

Abada, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems.* Available online at: https://www.tensorflow.org/ (accessed January 16, 2021).

Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., et al. (2020). Digital twins to personalize medicine. *Genome Med.* 12, 10–13. doi: 10.1186/s13073-019-0701-3

Bruynseels, K., Santoni de Sio, F., and van den Hoven, J. (2018). Digital twins in health care : ethical implications of an emerging engineering paradigm. *Front. Genet.* 9, 31. doi: 10.3389/fgene.2018.00031

Caballero Barajas, K. L., and Akella, R. (2015). "Dynamically modeling patient's health state from electronic medical records: a time series approach," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Sydney, NSW: Association for Computing Machinery) 69–78. doi: 10.1145/2783258.2783289

Chollet, F. (2015). Keras. *GitHub*. 2015.

Croatti, A., Gabellini, M., Montagna, S., and Ricci, A. (2020). On the integration of agents and digital twins in healthcare. *J. Med. Syst.* 44, 1–8. doi: 10.1007/s10916-020-01623-5

Do, R. K. G., Lupton, K., Causa Andrieu, P. I., Luthra, A., Taya, M., Batch, K., et al. (2021). Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT Radiology reports over a 10-year period. *Radiology* 301, 115–122. doi: 10.1148/radiol.2021210043

Gao, S., Alawad, M., Schaefferkoetter, N., Penberthy, L., Wu, X. C., Durbin, E. B., et al. (2020). Using case-level context to classify cancer pathology reports. *PLoS ONE* 15, e0232840. doi: 10.1371/JOURNAL.PONE.0232840

Ghannay, S., Favre, B., Esteve, Y., and Camelin, N. (2016). "Word embeddings evaluation and combination," in *10th edition of the Language Resources and Evaluation Conference (LREC 2016)* (PortoroŽ), 300–305.

Groot, O. Q., Bongers, M. E. R., Karhade, A. V., Kapoor, N. D., Fenn, B. P., Kim, J., et al. (2020). Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports. *Acta Oncol.* 59, 1455–1460. doi: 10.1080/0284186X.2020.1819563

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Huang, K., Altosaar, J., and Ranganath, R. (2020). ClinicalBERT: modeling clinical notes and predicting hospital readmission. *ArXiv*. arXiv:1904.05342.

Kehl, K. L., Elmarakeby, H., Nishino, M., Van Allen, E. M., Lepisto, E. M., Hassett, M. J., et al. (2019). Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* 5, 1421–1429. doi: 10.1001/jamaoncol.2019.1800

Khadanga, S., Aggarwal, K., Joty, S., and Srivastava, J. (2019). Using clinical notes with time series data for ICU management. *ArXiv*. arXiv:1909.09702.

National Institutes of Health. (2020). Help me understand precision medicine. *Medline Plus Genetics* (2020). Available online at: https://medlineplus.gov/genetics/understanding/precisionmedicine/ (accessed November 3, 2021).

Pons, E., Braun, L. M., M., Myriam Hunink, M. G., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology* 279, 329–343. doi: 10.1148/radiol.16142770

Renshaw, A. A., Mena-allauca, M., Gould, E. W., and Sirintrapun, S., J. (2018). Synoptic reporting : evidence-based review and future directions. *JCO Clin. Cancer Inform.* 2, 1–9. doi: 10.1200/CCI.17.00088

Senders, J. T., Karhade, A., Cote, D. J., Mehrtash, A., Lamba, N., DiRisio, A., et al. (2019). Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. *JCO Clin. Cancer Inform.* 3, 1–9. doi: 10.1200/CCI.18.00138

Verma, V. K., Pandey, M., Jain, T., and Tiwari, P. K. (2021). Dissecting word embeddings and language models in natural language processing. *J. Discr. Math. Sci. Cryptograph.* 24, 1509–1515. doi: 10.1080/09720529.2021.1968108

Zhao, R., and Mao, K. (2018). Fuzzy bag-of-words model for document. *IEEE Trans. Fuzzy Syst.* 26, 794–804. doi: 10.1109/TFUZZ.2017.2690222

Zuccon, G., Koopman, B., Bruza, P., and Azzopardi, L. (2015). "Integrating and evaluating neural word embeddings in information retrieval," in *Proceedings of the 20th Australasian Document Computing Symposium* (Parramatta, NSW: Association for Computing Machinery). doi: 10.1145/2838931.2838936

Conflict of Interest: LG activities not related to the present article; receives options for consultancy from Within Health.

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.

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 | Frontiers in Oncology

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/fonc.2022.790645

Precision Oncology: Artificial Intelligence and DNA Methylation Analysis of Circulating Cell-Free DNA for Lung Cancer Detection

Ray Bahado-Singh[1], *Kyriacos T. Vlachos*[2], *Buket Aydas*[3], *Juozas Gordevicius*[4], *Uppala Radhakrishna*[1] *and Sangeetha Vishweswaraiah*[5*]

[1] Department of Obstetrics and Gynecology, Oakland University William Beaumont School of Medicine, Royal Oak, MI, United States, [2] Department of Biomedical Sciences, Wayne State School of Medicine, Basic Medical Sciences, Detroit, MI, United States, [3] Department of Healthcare Analytics, Meridian Health Plans, Detroit, MI, United States, [4] Vugene, LLC, Grand Rapids, MI, United States, [5] Department of Obstetrics and Gynecology, Beaumont Research Institute, Royal Oak, MI, United States

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Prashanth N. Suravajhala, Amrita Vishwa Vidyapeetham University, India
Benson Babu, Northwell Health, United States

***Correspondence:**
Sangeetha Vishweswaraiah
sangeetha.vishweswaraiah@beaumont.org

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *07 October 2021*
Accepted: *04 April 2022*
Published: *04 May 2022*

Citation:
Bahado-Singh R, Vlachos KT, Aydas B, Gordevicius J, Radhakrishna U and Vishweswaraiah S (2022) Precision Oncology: Artificial Intelligence and DNA Methylation Analysis of Circulating Cell-Free DNA for Lung Cancer Detection. Front. Oncol. 12:790645. doi: 10.3389/fonc.2022.790645

Background: Lung cancer (LC) is a leading cause of cancer-deaths globally. Its lethality is due in large part to the paucity of accurate screening markers. Precision Medicine includes the use of omics technology and novel analytic approaches for biomarker development. We combined Artificial Intelligence (AI) and DNA methylation analysis of circulating cell-free tumor DNA (ctDNA), to identify putative biomarkers for and to elucidate the pathogenesis of LC.

Methods: Illumina Infinium MethylationEPIC BeadChip array analysis was used to measure cytosine (CpG) methylation changes across the genome in LC. Six different AI platforms including support vector machine (SVM) and Deep Learning (DL) were used to identify CpG biomarkers and for LC detection. Training set and validation sets were generated, and 10-fold cross validation performed. Gene enrichment analysis using g:profiler and GREAT enrichment was used to elucidate the LC pathogenesis.

Results: Using a stringent GWAS significance threshold, p-value $<5x10^{-8}$, we identified 4389 CpGs (cytosine methylation loci) in coding genes and 1812 CpGs in non-protein coding DNA regions that were differentially methylated in LC. SVM and three other AI platforms achieved an AUC=1.00; 95% CI (0.90-1.00) for LC detection. DL achieved an AUC=1.00; 95% CI (0.95-1.00) and 100% sensitivity and specificity. High diagnostic accuracies were achieved with only intragenic or only intergenic CpG loci. Gene enrichment analysis found dysregulation of molecular pathways involved in the development of small cell and non-small cell LC.

Conclusion: Using AI and DNA methylation analysis of ctDNA, high LC detection rates were achieved. Further, many of the genes that were epigenetically altered are known to be involved in the biology of neoplasms in general and lung cancer in particular.

Keywords: DNA methylation, lung cancer, artificial intelligence, machine learning, miRNAs

INTRODUCTION

Lung cancer (LC) is the leading cause of cancer deaths in the US and worldwide (1). There has been a dramatic rise in the incidence of this disorder over earlier decades largely due to smoking and more recently to environmental pollution among non-smokers. The 5-year survival rate is dismal at 4-17% (2) making LC the deadliest cancer in the USA. As per the International Agency for Research on Cancer (IARC) GLOBOCAN cancer statistics, 2.21 million cases of lung cancer cases were diagnosed in the year 2020 and 1.79 million deaths were registered worldwide (3). This high mortality is due principally to the late stage at which most cases are diagnosed highlighting the urgent need for the development of accurate biomarkers.

The US Preventative Services Task Force (USPTF) has recommended routine low-dose computed tomography (LDTC) LC screening of a defined population of high risk individuals (4). The USPTF however found that LDTC screening was associated with harms which included high false positive rates resulting in unnecessary tests and invasive procedures, incidental non-cancerous findings, overdiagnosis and radiation exposure. They therefore called for more research to develop biomarkers to improve the detection rate and lower the false positive rate of LDTC screening (4).

Significant focus has historically been placed on the role of gene mutations in the development of cancer. The extreme variability in the types of gene mutations in cancer however, has made it difficult to develop high sensitivity biomarkers for cancer diagnosis (5) using this approach. The stability and widespread nature of epigenomic changes in cancer has fueled its increasing study for understanding both the pathogenesis of cancer and for novel biomarker development. The best understood and most extensively studied epigenetic change is DNA methylation (6) which can alter gene expression.

Epigenetics and Cancer

Epigenetics is believed to play a key role in the neoplastic transformation of stem cells to form microscopic benign tumors (7), with extensive increase or decrease of methylation throughout the genome in most and possibly all tumors (8). Many studies have shown that tobacco smoke and other environmental exposures are important in LC pathogenesis, and induce significant epigenetic changes (9–12). Given the extensive degree of methylation changes throughout the genome and the likely role in neoplastic transformation, DNA methylation has great promise as an accurate and early potential biomarker for the detection of cancers.

Circulating Tumor DNA and LC

'Liquid biopsy' involves the harnessing of circulating tumor nucleic acids, such as tumor DNA (ctDNA), micro-RNA, exosomes, and tumor-educated platelets for LC (13) for cancer and other investigations. CtDNA describes cellular DNA released into the bloodstream and is present in higher amounts in cancer compared to normal cases. Several mechanisms such as necrosis and apoptosis induce this DNA release. Furthermore, it is known that newly synthesized DNA is periodically released even from viable intact cells. As a consequence, circulating tumor DNA (ctDNA) has gained increasing attention as a possible source of LC biomarkers (14) both for disease detection and real-time minimally invasive monitoring.

At its core, Precision Medicine deploys a combination of powerful biological approaches (e.g. genomics) and computational and bioinformatic tools for the detection and investigation of complex disorders. Precision Oncology is an established NIH priority (15). We have previously focused on the use of Machine Learning based Artificial Intelligence (AI) and 'omics' including epigenomics, metabolomics and proteomics for interrogation of disease mechanisms and the accurate detection of complex disorders (16, 17). Clinically validated DNA methylation markers currently do not exist for LC. In this study we used DNA methylation analysis of ctDNA to interrogate the molecular mechanisms of LC. Further, using multiple AI platforms combined with epigenomic markers, we accurately and minimally-invasively detected LC.

MATERIALS AND METHODS

Study Subjects and Sample Collection

This study was approved by Beaumont Institutional Review board (IRB#2018-306). Written patient consent was obtained. Blood samples were prospectively obtained from 10 LC cases and 20 controls in the present study. Only cases without any prior treatment for prior or current treatment for lung or other cancers were included in the study. Streck Cell-Free DNA BCT® tubes were used for collecting the blood samples from each study subjects. These tubes are designed to avoid the leukocyte genomic DNA contamination and thus minimizing the dilution and contamination of the cell-free (cf) DNA (18). Medical record numbers were removed, and unique study IDs were allocated to each sample for the purpose of de-identification of samples for laboratory analysis. All samples were processed within 24 hours of sample collection by centrifuging for 15 minutes at 3000 x g and aliquoting plasma into cryogenic vials. Samples were then stored at −80°C until further laboratory analysis (19). The **Figure 1** represents the overview of research methodology including downstream steps considered in the present study.

Sample Processing and Methylation Profiling

The cf-DNA was extracted using the QIAamp circulating nucleic acid kit (Qiagen Cat # 55114) manual vacuum manifold method. The samples were bisulfite converted using EZ DNA Methylation Kit (Zymo, USA) according to the standardized manufacturer's protocol. DNA methylation, analysis was performed using the Illumina Infinium MethylationEPIC BeadChip arrays (Illumina, Inc.). The array analyzes approximately 850,000 cytosine ('CpG' or 'cg') loci covering intragenic and extragenic regions of genome. The assay was performed based on the manufacturer's protocol, as described in detail previously (20).

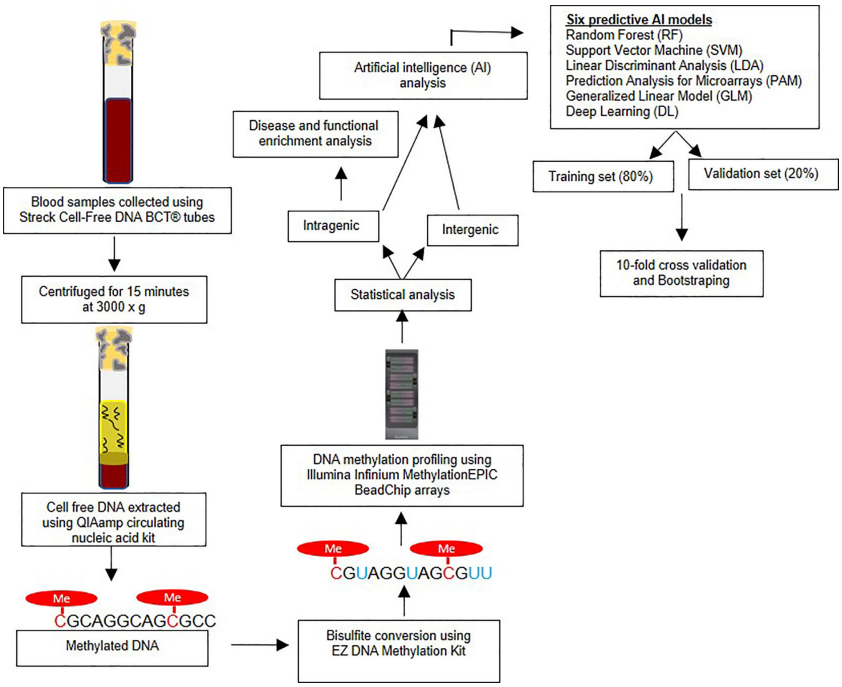


FIGURE 1 | Overview of research methodology – The figure outlines the sample collection, bisulfite conversion, methylation profiling followed by statistical and artificial intelligence analysis.

Statistical Analysis

The raw iDat files were analyzed using Illumina GenomeStudio software as described in our earlier studies (20). The β-values (methylation level at each cytosine locus) were measured and compared for statistical differences between the LC and control groups at each cytosine locus using the genome build hg37. To avoid gender bias, the CpG sites on the X and Y chromosomes were not considered in further analyses. Also, CpG loci within 10bp of any Single Nucleotide Polymorphism (SNPs) as observed on Single Nucleotide Polymorphism Database (dbSNP) were excluded as well to avoid genetic (e.g. mutations, single nucleotide polymorphisms) effects on methylation (21). For each CpG marker, the Area Under the Receiver Operating Characteristic (AUC) curve was computed using the R packages dplyr, reshape2 and ROCR. The genome-wide association studies (GWASs) significance p-value threshold $< 5x10^{-8}$ (22) was to designate significant CpG methylation change at each site.

Artificial Intelligence and Predictive Models for LC Detection

An important aim of our study was to test the performance of AI generated predictive algorithms, consistent with the objectives o Precision Oncology, for the detection of LC. AI ranked the top CpG markers in decreasing order of predictive ability. The top markers were then combined to generate the predictive algorithms for each AI platform. A total of six different AI algorithms were used to as previously reported (17, 23, 24). These platforms were: Random Forest (RF), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Prediction Analysis for Microarrays (PAM), Generalized Linear Model (GLM) and Deep Learning (DL). Each has relative strengths and limitations. The data was split into a training set (80% of subjects) and validation set (the remaining 20%) and 10-fold cross validation was performed. The splitting process was repeated ten times and the average area under the receiver operator characteristics curve (AUROC or AUC) and 95% confidence intervals was calculated for LC detection, along with sensitivity and specificity values (25). Bootstrapping using random sampling with replacement was also performed to optimize the accuracy of the estimates. The R package "Caret" was used to optimize predictions for five AI algorithms (RF, SVM, LDA, PAM and GLM) (https://cran.r-project.org/web/packages/caret/caret.pdf), and the package h2o was used to tune the parameters of DL algorithm (https://cran.r-project.org/web/packages/h2o/h2o.pdf) (26–28). The variable importance functions *varimp* in h2o and *varImp* in caret R packages were utilized to rank the models features in each of the predictive algorithms. We used pROC R package to compute the AUC, specificity and sensitivity values of the models (29). The detailed descriptions of AI algorithms, cross validation, bootstrapping, and feature ranking are provided in a **Supplementary Methods Section**.

Disease and Functional Enrichment Analysis

All analyses were performed using R programming language (v. 4.1.0). The EPIC array CpG loci were annotated using IlluminaHumanMethylationEPICanno.ilm10b4.hg19

Bioconductor package. For each CpG locus we determined the associated gene if any using the UCSC reference gene names (UCSC_RefGene_Name). When multiple genes were associated with a single CpG locus, the most frequently associated gene with that locus was used. Genomic Regions Enrichment of Annotations Tool (GREAT) was used to determine the number of CpGs associated with each gene and the distance of CpGs from the transcription start site (30). CpG methylation changes in transcription start site are more likely to be associated with altered gene expression and therefore to have an identifiable biological effect. g:profiler enrichment was performed using genes associated with statistically significant CpG loci as foreground and all annotated genes as background. R package gprofiler2 (v. 0.2.0) was used to make the enrichment API call with default parameters (31). miRNA enrichment analysis was performed by subjecting significant miRNAs to "miRNA Enrichment Analysis and Annotation Tool" (miEAA) v2.0 (32). We also searched for long non-coding RNA (lncRNA) using "LncExpDB" (33).

Principal Component Analysis

Given the large number of potential CpG epigenetic predictors generated, dimensionality reduction was performed using Principal Component Analysis (PCA). This approach reduces the number of predictors (dimensionality reduction) and thus simplifies and enhances the interpretability of the data. A visual display is generated showing whether with a limited number of CpG predictors the two groups (LC and controls) can be discriminated. We performed principal component analysis (PCA) MetaboAnalyst (v4.0) (34).

RESULTS

The demographic details of the study subjects are provided in **Table 1**. All study participants were of Caucasian race. The mean age between two groups was different (Mean age of cases is 64 years and controls were of 75 years, p-value < 0.01), BMI was also lower in LC cases. We therefore performed analysis adjusting for these confounders as well as gender. There were no differences between groups in the frequency of a positive family history for cancer. The histologic types and disease staging of the LC is also presented in **Supplementary Table S1**. Principal component analysis (PCA) showed very good visual the separation of LC and control groups (**Supplementary Figure 1**) using methylation markers. Using the GWAS significance threshold of p-value < $5x10^{-8}$ (22) we found a total of 4389 CpG loci (intragenic region) (3921 genes) that displayed significant methylation change in LC. Of the total of 4389 CpGs, 2906 were hyper-methylated (increased methylation) and 1483 CpGs were hypo methylated (decreased) in LC compared to control group (**Supplementary Table S2**).

TABLE 1 | The demographic characteristics of lung cancer cases and controls.

Parameter	Cases	Controls	p-value
Number of patients	10	20	–
Race - Caucasian	10	20	–
Age - Mean (Standard deviation)	63.9 (11.14)	74.85 (7.37)	0.01 (T)
Gender – n (%)			
Females	7 (70)	14 (70)	0.24 (W)
Males	3 (30)	6 (30)	
BMI - Mean (Standard deviation)	28.9 (3.4)	26.75 (5.3)	0.01 (T)
Family history of any cancer type – n (%)			
Yes	6 (60)	0 (0)	0.09 (W)
No	4 (40)	20 (100)	

T, T test; W, Wilcoxon Mann Whitney test.

We identified 1812 significantly differentially methylated CpGs in non-protein coding region of genome (intergenic region). Among them, 1067 CpGs were hyper methylated and 745 were hypo methylated CpGs (**Supplementary Table S3**). We found that 99% of these CpGs on both intra and intergenic CpGs showed methylation difference of greater than 5%. It should be noted that the higher the methylation difference the more likely is the epigenetic change to correlate with altered gene expression.

Artificial Intelligence and Lung Cancer Detection

A total of 19 individual CpGs among the intragenic CpGs and four among the intergenic CpGs had an excellent individual predictive value for LC detection based on AUC (AUC =1.00). We performed AI analysis using six different algorithms. Each AI platform was used to rank the CpG markers in decreasing order of predictive ability. We developed separate intragenic (within the gene) and intergenic (based on CpG markers) algorithms for LC detection. Using a 10-marker based algorithm, Five of the 6 AI algorithms using intragenic CpG markers achieved an excellent to outstanding diagnostic performance based on AUC (95% CI). These included SVM, GLM, RF and LDA with AUC=1.00 and 95% CI (0.90-1.00). DL had an AUC (95% CI) =1.00; (0.95-1.00) with 100% sensitivity and specificity, **Table 2**. Bootstrapping yielded excellent predictive accuracies, **Table 2A**. Equal or slightly lower detection rates were achieved when only 5 markers were used. For example, for SVM the AUC (95% CI) =1.00; (0.90-1.00) with 90% sensitivity and 100% specificity and for DL AUC (95% CI) =1.00; (0.95-1.00) with 100% sensitivity and 100% specificity. Likewise, when using 20 markers in the algorithm, the predictive accuracy was slightly higher but generally comparable to the 10-marker model. For example, for SVM the AUC (95% CI) =1.00; (0.90-1.00) with 94% sensitivity and 100% specificity and for DL AUC (95% CI) =1.00; (0.95-1.00) with 100% sensitivity and 100% specificity.

Likewise, using intergenic (non-coding region of the DNA) CpG markers, SVM, GLM, RF and LDA had excellent to outstanding diagnostic performance with AUC (95% CI) =1.00 (0.90-1.00) and DL performed with AUC (95% CI) =1.00 (0.95-1.00) and 100% sensitivity and specificity **Table 3**. Bootstrapping achieved similar detection performances **Table 3A**.

We identified 52 genes with at least 3 of their constituent CpGs significantly differentially methylated, 10294 genes had 2 CpGs and 5586 genes were found to have one CpG that had significant alteration in the methylation level in the ctDNA from LC versus normal group. The orientation of the CpGs from Transcription Start Site (TSS) and absolute distance from TSS are depicted on **Figure 2**. The closer the CpG locus is to the TSS the

TABLE 2 | Artificial Intelligence based prediction on methylation of cf-DNA Lung Cancer for the coding region CpGs (top 10 Variables).

	SVM	GLM	PAM	RF	LDA	DL
AUC 95% CI	1.0000 (0.9000-1)	1.0000 (0.9000-1)	0.9800 (0.8900-1)	1.0000 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9500-1)
Sensitivity	0.9400	0.9700	0.9600	0.9800	0.9200	1.0000
Specificity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

CpG predictors in decreasing order of contribution:
SVM: cg06829681 (TEAD1), cg24283889 (LOC102723701; ERLIN2), cg19403339 (DNAJC10), cg01430372 (TMEM99; KRT10), cg23280290 (HERPUD2), cg23178322 (FXR2; SHBG), cg15650170 (AGAP3), cg26864130 (MCAM), cg10299917 (LRP5L), cg25552416 (ZFP3).
GLM: cg10181281 (VWC2L), cg00941912 (KIAA1530), cg21722128 (MEIS3), cg15470857 (ZNF510), cg16267059 (MFAP1), cg16026813 (BTRC), cg25167447 (NAV1), cg16971745 (IFIH1), cg07401887 (DUXAP10), cg13390998 (NFKBIL2).
PAM: cg01430372 (TMEM99; KRT10), cg00071702 (CDH4), cg11149658 (MCPH1), cg07660991 (ZNF414), cg10299917 (LRP5L), cg08855953 (PRKACG), cg18227776 (NCOA2), cg14224170 (SAFB2), cg06270462 (EFHD1), cg00019091 (PTPN11).
RF: cg03871275 (DLK2), cg24847481 (SLC35A3), cg17094927 (ATP8B2), cg07199894 (ULK1), cg06831761 (SRPK2), cg18887033 (CMPK2), cg05398019 (COL27A1), cg24696183 (KCNQ1DN), cg06415550 (PTDSS2), cg16971745 (IFIH1).
LDA: cg26372202 (AK7), cg06819704 (CCNJL), cg10299917 (LRP5L), cg02401627 (LEKR1), cg26864130 (MCAM), cg11107657 (ODZ2), cg26024401 (DCDC2), cg11149658 (MCPH1), cg12282830 (AP1B1), cg25552416 (ZFP3).
DL: cg23496516 (USP36), cg07618979 (NFATC2), cg15684274 (NOC2L), cg06829681 (TEAD1), cg13302670 (CAMK2B), cg21466229 (SNTG1), cg23205538 (PARK2), cg14505733 (WNK2), cg25365034 (KLHL29), cg14364474 (GNAL).

TABLE 2A | Bootstrapping based on methylation of cf-DNA Lung Cancer for the coding region CpGs (top 10 Variables).

	SVM	GLM	PAM	RF	LDA	DL
AUC 95% CI	1.0000 (0.9000-1)	1.0000 (0.9000-1)	0.9822 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9500-1)
Sensitivity	0.9500	0.9700	0.9600	0.9800	0.9300	1.0000
Specificity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLE 3 | Artificial Intelligence based prediction on methylation of cf-DNA Lung Cancer for the non-coding region CpGs (top 10 Variables).

	SVM	GLM	PAM	RF	LDA	DL
AUC 95% CI	1.0000 (0.9000-1)	1.0000 (0.9000-1)	0.9900 (0.8900-1)	1.0000 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9500-1)
Sensitivity	0.9300	0.9600	0.9700	0.9800	0.9400	1.0000
Specificity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

CpG predictors in order of contribution:
SVM: cg16349277, cg10302285, cg21127580, cg07591229, cg15455979, cg13645106, cg05458412, cg00316520, cg14185604, cg02475408.
GLM: cg08090691, cg19319928, cg17373554, cg02821627, cg07099084, cg14852082, cg20802868, cg09853648, cg07877987, cg03388189.
PAM: cg05062489, cg09295542, cg06105068, cg24524245, cg19216204, cg03388189, cg06723904, cg13645106, cg12629103, cg02984449.
RF: cg07828654, cg02475408, cg14661028, cg03449513, cg22887498, cg10302285, cg26201011, cg08505243, cg20216928, cg04424605.
LDA: cg24196351, cg14071171, cg14559409, cg07892140, cg12629103, cg10430189, cg06723904, cg05909891, cg09295542, cg17001531.
DL: cg05458412, cg07652774, cg26399254, cg15398272, cg15125549, cg14852082, cg12629103, cg01076051, cg10086080, cg08852943.

TABLE 3A | Bootstrapping based on methylation of cf-DNA Lung Cancer for the non-coding region CpGs (top 10 Variables).

	SVM	GLM	PAM	RF	LDA	DL
AUC 95% CI	1.0000 (0.9000-1)	1.0000 (0.9000-1)	0.9910 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9000-1)	1.0000 (0.9500-1)
Sensitivity	0.9400	0.9650	0.9733	0.9800	0.9475	1.0000
Specificity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

greater is the likelihood that the methylation change will be biological significant i.e., result in altered gene expression.

Due to the difference in age group of cases and controls, we performed further analysis in which potential confounders such as age and gender were considered with CpG markers. A 50-marker algorithm did not find any of these potential confounders to contribute significantly to LC prediction. All 50 markers for each AI platform were CpG loci for both the intra- and extra-genic analyses (**Supplementary Table S4**).

Gene Enrichment Analysis

There were 4 significantly enriched terms associated with LC. (1) WikiPathways - Non-small cell lung cancer (NSCLC) (WP : WP4255, p=1.76e-7), (2) KEGG Non-small cell lung cancer (KEGG:05223, p=5.33e-7), (3) WikiPathways - Small cell lung cancer (WP : WP4658, p=0.0020) and (4) KEGG - Small cell lung cancer (KEGG:05222, p=0.0034). The constituent genes in these significantly enriched pathways that were found to be epigenetically altered are listed in **Supplementary Table S5**

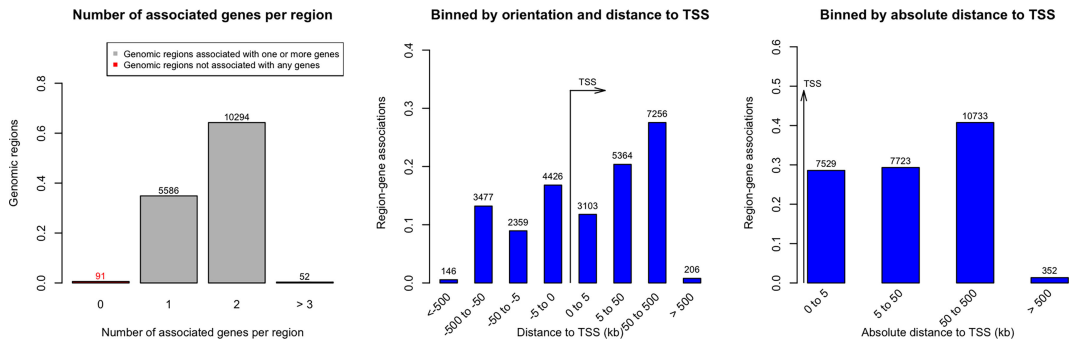


FIGURE 2 | Distance of significantly methylated CpGs' from Transcription Start Site (TSS) in Lung cancer.

along with their known or putative roles in LC and in neoplasms in general. Overall, these individual genes based on the quoted references, appear to have a significant role in LC and neoplastic transformation. The epigenetic dysregulation of known LC and cancer molecular pathways lends biological credibility to our findings and supports the argument for a significant role of DNA methylation changes in LC development.

Overall, miRNA genes (epigenetically altered in LC) were found to be enriched and was the top significant term with p-value of $<8.24e^{-239}$ based on g:profiler enrichment analysis. We observed 45 miRNA genes to be significantly differentially methylated in our study. The CpGs (49 CpGs) encompassing regions of these 45 miRNA genes are provided on **Supplementary Table S6** (This is a subset data of **Supplemental Table S2**) and their enrichment status relative to lung cancer is detailed (**Supplementary Table S7**). We also identified 70 CpGs from 66 lncRNA genes that were differentially methylated and associated with LC. The CpGs corresponding to lncRNAs are provided in the **Supplementary Table S8** (This is a subset data of **Supplementary Table S2**). A few of these differentially methylated lncRNAs were previously found to be associated with lung cancer as detailed in the **Supplementary Table S9**.

DISCUSSION

In 2017 the U.S. Food and Drug administration (FDA) established the Oncology Center of Excellence to promote Precision Medicine in oncology and for the development of new cancer therapies. Its writ included the development of biomarker-based treatments and is grounded in the advances made in our understanding of the genomics of cancer pathogenesis and propagation (35). As noted previously, key to the improvement of LC outcomes will be the development of accurate biomarkers. The potential therapeutic value of liquid biopsies including ctDNA in oncology, have been addressed in other reviews (36). These include cancer screening and diagnosis in asymptomatic populations, identifying individual patients for specific treatments, identifying evidence of residual disease after treatment, predicting the risk of relapse, detection of recurrence, distinguishing true from pseudo progression and reducing prolonged or unnecessary treatments in patients. We combined AI with the DNA methylation analysis of circulating tumor DNA to investigate both the mechanism and for the minimally-invasive detection of LC. We achieved highly accurate detection of LC using six different AI platforms with AUC = 0.90-1.0 and high sensitivity and specificity values. For example, Deep Learning achieved high performance with AUC (95% CI) =1.0, with 100% sensitivity and specificity in this preliminary study. High diagnostic accuracies were similarly achieved with algorithms based on combinations of smaller or larger numbers of individual CpG epigenetic markers. The excellent performance was also achieved when only intragenic or alternatively intergenic CpG loci were considered. In the present study, the classes are moderately imbalanced (i.e., no worse than 10:1). Hence, we did not perform analysis to limit the class imbalance which would otherwise have no huge benefit of considering either weighting or sampling techniques to limit the class imbalance. If there was a class imbalance, we would consider different methods to help improve classification performance. Some of the popular techniques to deal with class imbalances are: (i) Class weights: impose a heavier cost when errors are made in the minority class, (ii) Down-sampling: randomly remove instances in the majority class, (iii) Up-sampling: randomly replicate instances in the minority class and (iv) Synthetic minority sampling technique (SMOTE): down samples the majority class and synthesizes new minority instances by interpolating between existing ones.

We further found evidence of a significant role of epigenetic dysregulation in know molecular pathways involved in LC pathogenesis (discussed in more detail below). Confounders such as age and gender did not appear as independent predictors of cancer beyond the epigenetic markers when we adjusted for these confounders in AI analysis. This is likely due to the fact that these variables have an epigenetic impact which is already subsumed in the DNA data.

Freitas et al. (13) recently reviewed the literature on gene mutation analysis of ctDNA for LC detection. Overall, studies screening for multiple rather than a single cancer gene mutation in ctDNA appeared to have higher diagnostic performance. Gene mutation biomarker studies evaluating a combination from 3 to139 cancer related genes achieved a performance that varied

from 33% sensitivity and 100% specificity to a high of 85% sensitivity and 96% specificity.

We focused on DNA methylation given the burgeoning evidence of the centrality of epigenomics in tumorigenesis (37). Other studies have confirmed the feasibility of this approach. Using a combination of methylation markers in 6 cancer genes based on plasma ctDNA analysis, Hsu et al. (38) achieved an 73% sensitivity and 82% specificity for LC detection. Begum et al. (39) performed methylation analysis using serum cell-free DNA. Using a combination of five genes they reported an 75% sensitivity and 73% specificity for LC detection. Zheng et al. (40) achieved a sensitivity of 83.64% and a specificity of 74.0% using a combination methylation profiling of five genes from plasma ctDNA.

Methylation analysis may have a future advantage in facilitating new therapeutic approaches. Targeted alteration of epigenomic changes is emerging as a potentially highly impactful therapeutic approach in cancer. This involves the precise targeting of DNA sequences to reverse or introduce epigenetic marks. The CRISPR/Cas 9 system appears to be the most exciting though not the only such approach (41). The CRISPR/Cas-9 approach has been used for targeted reversal i.e. removal of DNA methylation (demethylation) leading to gene activation in cancer (42).

An important objective of "Precision Oncology" is deploying omics and AI to investigate disease pathogenesis. Recent advances in machine learning (a branch of AI) point to a significant potential for future impact on medical research and practice. It has been noted that AI methods could potentially make significant contributions in the medical field in the following areas: understanding "disease underlying architecture, perform early diagnosis of diseases, and disease progression prediction" (43).

We found alterations in molecular pathways that are involved in non-small cell lung cancer (NSCLC), small cell lung cancer development. Our findings provide further evidence in support of the importance of epigenetic dysregulation in LC. Further, the association with known or suspected LC cancer molecular pathways gives biological plausibility to our findings. The cancer related functions of the genes found to be epigenetically dysregulated in this study is further summarized. In **Supplementary Table S10**, we list the function of genes that were identified to be epigenetically altered and determined by AI to be LC markers, along with their known or suspected roles in LC and neoplastic transformation based on the published literature. Given what is known about their apparent roles of these genes in the neoplastic process, it is therefore not surprising that they emerged as significant markers for LC detection. Examples of epigenetically modified genes that were found in our study and are catalogued in **Supplementary Table S5** include *FHIT, FN1, FOXO3* and *GRB2.* They are thought to regulate epithelial-mesenchymal transition and/or metastasis and associated with LC. Also, *ITGA2, ITGA3* and *ITGA6* are integrin coding genes that participate in cell adhesion, proliferation, and differentiation and are known to have anti-cancer properties in LC. It should be pointed out however that the LC roles of a significant number of genes that were epigenetically altered in our study are currently unknown. Should our findings be subsequently validated, the function of the latter genes in cancer should be investigated. Also, the function of the constituent genes involved in the enrichment pathways reveal an important role in neoplasms in general. Overall, our results were generally enriched with many genes currently known or suspected to be involved in carcinogenesis, giving biological plausibility to our findings.

MicroRNA (miRNA) are small single stranded non-coding RNAs. They play an important role in gene expression through the post- translational regulation of multiple other genes. This is accomplished by binding of miRNA to and degradation of the mRNA of other genes and thus inhibiting their expression. MicroRNA is another well-known epigenetic mechanism. DNA methylation in turn is critical in regulating the expression of miRNA genes (44). miRNA is increasingly being recognized as playing an important role in lung cancer including in tumorigenesis, tumor suppression, with value as biomarkers and potential therapeutic roles among others (45). In the current study, miRNAs overall were found to be significantly enriched, p-value of $8.24e^{-249}$, in our gene enrichment analysis. We found a total of 45 miRNAs that were significantly differentially methylated and most of them were enriched in various LC phenotypes (**Supplementary Table S6**) signifying the complex regulation of miRNA *via* methylation and regulation of gene expression in LC. Further, we performed a literature review to determine whether our overrepresented miRNAs and their targets were previously identified as having a role in LC pathogenesis. These include miR-96-5p previously identified as an oncogene in lung adenocarcinoma (46), miR-126, miR-212, miR-330, miR-432, miR-563, miR-663a, miR-1238 are considered to be tumor suppressor miRNAs (47–53), miR-136 is significantly upregulated in human NSCLC primary tumors (54). Further, miR-141-3p appears to have prognostic value and is a tumor suppressor involved in the NSCLC progression (55), miR-346 promotes cell growth and metastasis and suppresses apoptosis in non-small cell lung cancer (56), miR-601 is associated with cell apoptosis in lung cancer (57), miR-2861 expression was found to be higher in lung cancer stem cells (58), miR-1307 promotes the proliferation of lung adenocarcinoma (59), the miR-1469 is an apoptosis enhancer that regulates lung cancer apoptosis (60) and miR-200c plays a significant role in suppressing Epithelial-mesenchymal transition in lung cancer (61).

Based on circulating miRNAs studies, the circulating miRNAs, miR-10b and miR141 were found to be elevated in lung cancer cases (62), while circulating miR-487a, miR-30b, miR-601 were found to be associated with NSCLC (63). The serum exosome miR-96 has been identified as a biomarker for lung cancer (64). We also identified lncRNA genes that were differentially methylated in lung cancer and a few of these lncRNAs were already identified in various lung cancer studies (**Supplementary Table S9**).

Although very encouraging, our study is not without limitations. As a proof-of-concept study, the sample sizes were

small. it is possible for example that more DNA methylation markers could be detected with the analysis of a larger sample cohort. Despite these limitations, high statistical significance was obtained. Due to the non-suitability of the circulating cell-free DNA for gene expression analyses, we were unable to assess gene expression associated with the methylation changes. We however searched databases based on two studies i.e. 65 (65) and 51 (66) that document gene expression changes in lung cancer tissue. We cross-matched their differentially expressed genes with our differentially methylated genes. We found the following genes to be both differentially expressed in LC tissue and differentially methylated in our study: *DSC3, MUC1, VSNL1, RORC, ACSL5, KRT6B* and *TP63*. Further, many of the CpGs that were epigenetically altered are located close to the gene transcription start site (TSS), which would indicate that methylation changes are likely to impact gene expression. Finally, there were many LC genes with methylation change ≥ 10%. This degree of methylation difference is generally associated with an increased likelihood of gene expression changes (67).

Conclusion

Using principles espoused in Precision Medicine, we report that a combination of DNA methylation analysis of circulating tumor DNA and AI achieved high LC detection rates based on this minimally invasive approach. High performances were observed with the analysis of either intragenic or intergenic areas of the DNA. In addition, many of the genes that were found to be differentially methylated in LC in our study are known or suspected, based on a search of the existing literature, to be involved in the mechanism of development, suppression, or growth of cancer in general including lung cancer. Larger confirmation studies will need to be performed in the future.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Beaumont Institutional Review Board (IRB#2018-306). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RB-S: conceptualization, overview of project, data analysis, and writing manuscript. KV: data analysis and manuscript writing. BA: artificial intelligence methodology. JG: data analysis and enrichment analysis. UR: sample processing, data analysis, and manuscript writing. SV: conceptualization, samples and array processing, data analysis, and writing manuscript. All authors have read and edited the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.790645/full#supplementary-material

Supplementary Figure 1 | Principal component analysis (PCA) showing separation of cases and control subjects based on methylation markers in Lung cancer.

Supplementary Table 2 | Significantly differentially methylated CpG markers (FDR-p values) for coding regions

Supplementary Table 3 | Significantly differentially methylated CpG markers (FDR-p values) for the non-coding DNA regions

Supplementary Table 6 | Significantly differentially methylated CpG markers (FDR-p values) for the miRNA coding genes (a subset of **Supplementary Table S2**).

Supplementary Table 7 | Over-representation of differentially methylated miRNA in lung cancer phenotypes based on multiple databases.

Supplementary Table 8 | Significantly differentially methylated CpG markers (FDR-p values) for the lncRNA coding genes (a subset of **Supplementary Table S2**).

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer Statistic. *CA: A Cancer J Clin* (2018) 68:7–30. doi: 10.3322/caac.21442
2. Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ Jr, Wu YL, et al. Lung Cancer: Current Therapies and New Targeted Treatments. *Lancet* (2017) 389:299–311. doi: 10.1016/S0140-6736(16)30958-8
3. Ferlay J, Colombet M, Soerjomataram I, Parkin DM. *Cancer Statistics for the Year 2020: An Overview*. (2021). Wiley Online Library.
4. Squires BS, Levitin R, Grills IS. The US Preventive Services Task Force Recommendation on Lung Cancer Screening. *JAMA* (2021) 326:440–1. doi: 10.1001/jama.2021.8240
5. Berdasco M, Esteller M. Clinical Epigenetics: Seizing Opportunities for Translation. *Nat Rev Genet* (2019) 20:109–27. doi: 10.1038/s41576-018-0074-2
6. Moore LD, Le T, Fan G. DNA Methylation and its Basic Function. *Neuropsychopharmacol: Off Publ Am Coll Neuropsychopharmacol* (2013) 38:23–38. doi: 10.1038/npp.2012.112
7. Scott RE, Wille JJ Jr, Wier ML. Mechanisms for the Initiation and Promotion of Carcinogenesis: A Review and a New Concept. *Mayo Clin Proc* (1984) 59:107–17. doi: 10.1016/S0025-6196(12)60244-4
8. Feinberg AP, Vogelstein B. Alterations in DNA Methylation in Human Colon Neoplasia. *Semin Surg Oncol* (1987) 3:149–51. doi: 10.1002/ssu.2980030304
9. Hong Y, Choi HM, Cheong HS, Shin HD, Choi CM, Kim WJ. Epigenome-Wide Association Analysis of Differentially Methylated Signals in Blood Samples of Patients With Non-Small-Cell Lung Cancer. *J Clin Med* (2019) 8:1307. doi: 10.3390/jcm8091307
10. Xu W, Lu J, Zhao Q, Wu J, Sun J, Han B, et al. Genome-Wide Plasma Cell-Free DNA Methylation Profiling Identifies Potential Biomarkers for Lung Cancer. *Dis Markers* (2019) 2019:4108474. doi: 10.1155/2019/4108474
11. Guo D, Yang L, Yang J, Shi K. Plasma Cell-Free DNA Methylation Combined With Tumor Mutation Detection in Prognostic Prediction of Patients With Non-Small Cell Lung Cancer (NSCLC). *Med (Baltimore)* (2020) 99:e20431. doi: 10.1097/MD.0000000000020431
12. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and Characterization of Lung Cancer Using Cell-Free DNA

Fragmentomes. *Nat Commun* (20215060) 12:1–14. doi: 10.1038/s41467-021-24994-w
13. Freitas C, Sousa C, Machado F, Serino M, Santos V, Cruz-Martins N, et al. The Role of Liquid Biopsy in Early Diagnosis of Lung Cancer. *Front Oncol* (2021) 11. doi: 10.3389/fonc.2021.634316
14. Lianidou E. Detection and Relevance of Epigenetic Markers on ctDNA: Recent Advances and Future Outlook. *Mol Oncol* (2021) 15:1683–700. doi: 10.1002/1878-0261.12978
15. Collins FS, Varmus H. A New Initiative on Precision Medicine. *New Engl J Med* (2015) 372:793–5. doi: 10.1056/NEJMp1500523
16. Turkoglu O, Citil A, Katar C, Mert I, Kumar P, Yilmaz A, et al. Metabolomic Identification of Novel Diagnostic Biomarkers in Ectopic Pregnancy. *Metabolomics* (2019) 15:143. doi: 10.1007/s11306-019-1607-1
17. Bahado-Singh RO, Vishweswaraiah S, Er A, Aydas B, Turkoglu O, Taskin BD, et al. Artificial Intelligence and the Detection of Pediatric Concussion Using Epigenomic Analysis. *Brain Res* (2020) 1726:146510. doi: 10.1016/j.brainres.2019.146510
18. Bartak BK, Kalmar A, Galamb O, Wichmann B, Nagy ZB, Tulassay Z, et al. Blood Collection and Cell-Free DNA Isolation Methods Influence the Sensitivity of Liquid Biopsy Analysis for Colorectal Cancer Detection. *Pathol Oncol Res* (2019) 25:915–23. doi: 10.1007/s12253-018-0382-z
19. Sheinerman KS, Toledo JB, Tsivinsky VG, Irwin D, Grossman M, Weintraub D, et al. Circulating Brain-Enriched microRNAs as Novel Biomarkers for Detection and Differentiation of Neurodegenerative Diseases. *Alzheimers Res Ther* (2017) 9:89. doi: 10.1186/s13195-017-0316-0
20. Bahado-Singh RO, Vishweswaraiah S, Aydas B, Yilmaz A, Metpally RP, Carey DJ, et al. Artificial Intelligence and Leukocyte Epigenomics: Evaluation and Prediction of Late-Onset Alzheimer's Disease. *PloS One* (2021) 16:e0248375. doi: 10.1371/journal.pone.0248375
21. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, et al. Review of Processing and Analysis Methods for DNA Methylation Array Data. *Br J Cancer* (2013) 109:1394–402. doi: 10.1038/bjc.2013.496
22. Jannot AS, Ehret G, Perneger T. P < 5 × 10(-8) has Emerged as a Standard of Statistical Significance for Genome-Wide Association Studies. *J Clin Epidemiol* (2015) 68:460–5. doi: 10.1016/j.jclinepi.2015.01.001
23. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. *Omics* (2018) 22:630–6. doi: 10.1089/omi.2018.0097
24. Dias R, Torkamani A. Artificial Intelligence in Clinical and Genomic Diagnostics. *Genome Med* (2019) 11:70. doi: 10.1186/s13073-019-0689-8
25. Gedeon TD. Data Mining of Inputs: Analysing Magnitude and Functional Measures. *Int J Neural Syst* (1997) 8:209–18. doi: 10.1142/S0129065797000227
26. Kuhn M. Building Predictive Models in R Using the Caret Package. *J Stat Softw Articles* (2008) 28:1–26. doi: 10.18637/jss.v028.i05
27. Alakwaa FM, Chaudhary K, Garmire LX. Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. *J Proteome Res* (2018) 17:337–47. doi: 10.1021/acs.jproteome.7b00595
28. Candel A, Parmar V, Ledell E, Arora A. *Deep Learning With H2O*. (2018).
29. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinf* (2011) 12:77. doi: 10.1186/1471-2105-12-77
30. Mclean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT Improves Functional Interpretation of Cis-Regulatory Regions. *Nat Biotechnol* (2010) 28:495–501. doi: 10.1038/nbt.1630
31. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G: Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene ListUpdate). *Nucleic Acids Res* (2019) 47:W191–8. doi: 10.1093/nar/gkz369
32. Kern F, Fehlmann T, Solomon J, Schwed L, Grammes N, Backes C, et al. miEAA 2.0: Integrating Multi-Species microRNA Enrichment Analysis and Workflow Management Systems. *Nucleic Acids Res* (2020) 48:W521–8. doi: 10.1093/nar/gkaa309
33. Li Z, Liu L, Jiang S, Li Q, Feng C, Du Q, et al. LncExpDB: An Expression Database of Human Long Non-Coding RNAs. *Nucleic Acids Res* (2020) 49:D962–8. doi: 10.1093/nar/gkaa850
34. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res* (2018) 46:W486–94. doi: 10.1093/nar/gky310
35. Goldberg KB, Blumenthal GM, Mckee AE, Pazdur R. The FDA Oncology Center of Excellence and Precision Medicine. *Exp Biol Med (Maywood)* (2018) 243:308–12. doi: 10.1177/1535370217740861
36. De Mattos-Arruda L, Siravegna G. How to Use Liquid Biopsies to Treat Patients With Cancer. *ESMO Open* (2021) 6:100060. doi: 10.1016/j.esmoop.2021.100060
37. Aran D, Hellman A. DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. *Cell* (2013) 154:11–3. doi: 10.1016/j.cell.2013.06.018
38. Hsu HS, Chen TP, Hung CH, Wen CK, Lin RK, Lee HC, et al. Characterization of a Multiple Epigenetic Marker Panel for Lung Cancer Detection and Risk Assessment in Plasma. *Cancer* (2007) 110:2019–26. doi: 10.1002/cncr.23001
39. Begum S, Brait M, Dasgupta S, Ostrow KL, Zahurak M, Carvalho AL, et al. An Epigenetic Marker Panel for Detection of Lung Cancer Using Cell-Free Serum DNA. *Clin Cancer Res* (2011) 17:4494–503. doi: 10.1158/1078-0432.CCR-10-3436
40. Zhang Y, Wang R, Song H, Huang G, Yi J, Zheng Y, et al. Methylation of Multiple Genes as a Candidate Biomarker in Non-Small Cell Lung Cancer. *Cancer Lett* (2011) 303:21–8. doi: 10.1016/j.canlet.2010.12.011
41. Gaj T, Gersbach CA, Barbas CF3rd. ZFN, TALEN, and CRISPR/Cas-Based Methods for Genome Engineering. *Trends Biotechnol* (2013) 31:397–405. doi: 10.1016/j.tibtech.2013.04.004
42. Choudhury SR, Cui Y, Lubecka K, Stefanska B, Irudayaraj J. CRISPR-Dcas9 Mediated TET1 Targeting for Selective DNA Demethylation at BRCA1 Promoter. *Oncotarget* (2016) 7:46545–56. doi: 10.18632/oncotarget.10234
43. Mi X, Zou B, Zou F, Hu J. Permutation-Based Identification of Important Biomarkers for Complex Diseases *via* Machine Learning Models. *Nat Commun* (20213008) 12:1–12. doi: 10.1038/s41467-021-22756-2
44. Fuso A, Raia T, Orticello M, Lucarelli M. The Complex Interplay Between DNA Methylation and miRNAs in Gene Expression Regulation. *Biochimie* (2020) 173:12–6. doi: 10.1016/j.biochi.2020.02.006
45. Iqbal MA, Arora S, Prakasam G, Calin GA, Syed MA. MicroRNA in Lung Cancer: Role, Mechanisms, Pathways and Therapeutic Relevance. *Mol Aspects Med* (2019) 70:3–20. doi: 10.1016/j.mam.2018.07.003
46. Liu Z, Cui Y, Wang S, Wu C, Mei F, Han E, et al. MiR-96-5p is an Oncogene in Lung Adenocarcinoma and Facilitates Tumor Progression Through ARHGAP6 Downregulation. *J Applied Genetic* (2021) 62:631–8. doi: 10.1007/s13353-021-00652-1
47. Incoronato M, Urso L, Portela A, Laukkanen MO, Soini Y, Quintavalle C, et al. Epigenetic Regulation of miR-212 Expression in Lung Cancer. *PloS One* (2011) 6:e27722. doi: 10.1371/journal.pone.0027722
48. Shi X, Zhan L, Xiao C, Lei Z, Yang H, Wang L, et al. miR-1238 Inhibits Cell Proliferation by Targeting LHX2 in Non-Small Cell Lung Cancer. *Oncotarget* (2015) 6:19043–54. doi: 10.18632/oncotarget.4232
49. Chen L, Kong G, Zhang C, Dong H, Yang C, Song G, et al. MicroRNA-432 Functions as a Tumor Suppressor Gene Through Targeting E2F3 and AXL in Lung Adenocarcinoma. *Oncotarget* (2016) 7:20041–53. doi: 10.18632/oncotarget.7884
50. Zhang Y, Xu X, Zhang M, Wang X, Bai X, Li H, et al. MicroRNA-663a is Downregulated in Non-Small Cell Lung Cancer and Inhibits Proliferation and Invasion by Targeting JunD. *BMC Cancer* (2016) 16:315. doi: 10.1186/s12885-016-2350-x
51. Zhang X, Li M, Sun G, Bai Y, Lv D, Liu C. MiR-563 Restrains Cell Proliferation *via* Targeting LIN28B in Human Lung Cancer. *Thorac Cancer* (2020) 11:55–61. doi: 10.1111/1759-7714.13257
52. Chen Q, Chen S, Zhao J, Zhou Y, Xu L. MicroRNA-126: A New and Promising Player in Lung Cancer. *Oncol Lett* (2021) 21:35–5. doi: 10.3892/ol.2020.12296
53. Mohammadi A, Mansoori B. Restoration of miR-330 Expression Suppresses Lung Cancer Cell Viability, Proliferation, and Migration. *J Cell Physiol* (2021) 236:273–83. doi: 10.1002/jcp.29840
54. Shen S, Yue H, Li Y, Qin J, Li K, Liu Y, et al. Upregulation of miR-136 in Human non-Small Cell Lung Cancer Cells Promotes Erk1/2 Activation by Targeting PPP2R2A. *Tumour Biol* (2014) 35:631–40. doi: 10.1007/s13277-013-1087-2

55. Li W, Cui Y, Wang D, Wang Y, Wang L. MiR-141-3p Functions as a Tumor Suppressor Through Directly Targeting ZFR in Non-Small Cell Lung Cancer. *Biochem Biophys Res Commun* (2019) 509:647–56. doi: 10.1016/j.bbrc.2018.12.089
56. Sun CC, Li SJ, Yuan ZP, Li DJ. MicroRNA-346 Facilitates Cell Growth and Metastasis, and Suppresses Cell Apoptosis in Human Non-Small Cell Lung Cancer by Regulation of XPC/ERK/Snail/E-Cadherin Pathway. *Aging (Albany NY)* (2016) 8:2509–24. doi: 10.18632/aging.101080
57. Ohdaira H, Nakagawa H, Yoshida K. Profiling of Molecular Pathways Regulated by microRNA 601. *Comput Biol Chem* (2009) 33:429–33. doi: 10.1016/j.compbiolchem.2009.09.003
58. Zhao M, Li L, Zhou J, Cui X, Tian Q, Jin Y, et al. MiR-2861 Behaves as a Biomarker of Lung Cancer Stem Cells and Regulates the HDAC5-ERK System Genes. *Cell Reprogram* (2018) 20:99–106. doi: 10.1089/cell.2017.0045
59. Du X, Wang S, Liu X, He T, Lin X, Wu S, et al. MiR-1307-5p Targeting TRAF3 Upregulates the MAPK/NF-κb Pathway and Promotes Lung Adenocarcinoma Proliferation. *Cancer Cell Int* (2020) 20:502. doi: 10.1186/s12935-020-01595-z
60. Xu C, Zhang L, Li H, Liu Z, Duan L, Lu C. MiRNA-1469 Promotes Lung Cancer Cells Apoptosis Through Targeting STAT5a. *Am J Cancer Res* (2015) 5:1180–9.
61. Liu C, Hu W, Li LL, Wang YX, Zhou Q, Zhang F, et al. Roles of miR-200 Family Members in Lung Cancer: More Than Tumor Suppressors. *Future Oncol* (2018) 14:2875–86. doi: 10.2217/fon-2018-0155
62. Roth C, Kasimir-Bauer S, Pantel K, Schwarzenbach H. Screening for Circulating Nucleic Acids and Caspase Activity in the Peripheral Blood as Potential Diagnostic Tools in Lung Cancer. *Mol Oncol* (2011) 5:281–91. doi: 10.1016/j.molonc.2011.02.002
63. Zhou C, Chen Z, Zhao L, Zhao W, Zhu Y, Liu J, et al. A Novel Circulating miRNA-Based Signature for the Early Diagnosis and Prognosis Prediction of Non-Small-Cell Lung Cancer. *J Clin Lab Anal* (2020) 34:e23505. doi: 10.1002/jcla.23505
64. Wu H, Zhou J, Mei S, Wu D, Mu Z, Chen B, et al. Circulating Exosomal microRNA-96 Promotes Cell Proliferation, Migration and Drug Resistance by Targeting LMO7. *J Cell Mol Med* (2017) 21:1228–36. doi: 10.1111/jcmm.13056
65. Karlsson A, Jönsson M, Lauss M, Brunnström H, Jönsson P, Borg Å., et al. Genome-Wide DNA Methylation Analysis of Lung Carcinoma Reveals One Neuroendocrine and Four Adenocarcinoma Epitypes Associated With Patient Outcome. *Clin Cancer Res* (2014) 20:6127–40. doi: 10.1158/1078-0432.CCR-14-1087
66. Zhang H, Jin Z, Cheng L, Zhang B. Integrative Analysis of Methylation and Gene Expression in Lung Adenocarcinoma and Squamous Cell Lung Carcinoma. *Front Bioengineering Biotechnol* (2020) 8. doi: 10.3389/fbioe.2020.00003
67. Leenen FA, Muller CP, Turner JD. DNA Methylation: Conducting the Orchestra From Exposure to Phenotype? *Clin Epigenet* (2016) 8:92. doi: 10.1186/s13148-016-0256-8

Conflict of Interest: Author BA was employed by Meridian Health Plans. Author JG was employed by Vugene, LLC.

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.

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 | Frontiers in Artificial Intelligence

BRIEF RESEARCH REPORT
published: 23 June 2022
doi: 10.3389/frai.2022.851841

A Straightforward HPV16 Lineage Classification Based on Machine Learning

Laura Asensio-Puig[1]*, *Laia Alemany*[1,2] and *Miquel Angel Pavón*[1,2]*

[1] Cancer Epidemiology Research Programme, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain, [2] Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

Human Papillomavirus (HPV) is the causal agent of 5% of cancers worldwide and the main cause of cervical cancer and it is also associated with a significant percentage of oropharyngeal and anogenital cancers. More than 60% of cervical cancers are caused by HPV16 genotype, which has been classified into lineages (A, B, C, and D). Lineages are related to the progression of cervical cancer and the current method to assess lineages is by building a Maximum Likelihood Tree (MLT); which is slow, it cannot assess poor sequenced samples, and annotation is done manually. In this study, we have developed a new model to assess HPV16 lineage using machine learning tools. A total of 645 HPV16 genomes were analyzed using Genome-Wide Association Study (GWAS), which identified 56 lineage-specific Single Nucleotide Polymorphisms (SNPs). From the SNPs found, training-test models were constructed using different algorithms such as Random Forest (RF), Support Vector Machine (SVM), and K-nearest neighbor (KNN). A distinct set of HPV16 sequences ($n = 1{,}028$), whose lineage was previously determined by MLT, was used for validation. The RF-based model allowed a precise assignment of HPV16 lineage, showing an accuracy of 99.5% in the known lineage samples. Moreover, the RF model could assess lineage to 273 samples that MLT could not determine. In terms of computer consuming time, the RF-based model was almost 40 times faster than MLT. Having a fast and efficient method for assigning HPV16 lineages, could facilitate the implementation of lineage classification as a triage or prognostic marker in the clinical setting.

Keywords: Human Papillomavirus (HPV), cancer, prognostic and predictive factors, classification, machine learning, HPV16 lineage

OPEN ACCESS

Edited by:
Mónica Hebe Vazquez-Levin, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Reviewed by:
Vishal Nayak, Centers for Disease Control and Prevention (CDC), United States
Carlos Riveros, University of Newcastle, Australia

***Correspondence:**
Laura Asensio-Puig
lasensio@idibell.cat
Miquel Angel Pavón
mpavon@iconcologia.net

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *10 January 2022*
Accepted: *05 May 2022*
Published: *23 June 2022*

Citation:
Asensio-Puig L, Alemany L and Pavón MA (2022) A Straightforward HPV16 Lineage Classification Based on Machine Learning. Front. Artif. Intell. 5:851841. doi: 10.3389/frai.2022.851841

INTRODUCTION

A total of 5% of worldwide cancers are caused by the Human Papillomavirus (HPV) being cervical cancer the fourth most common cancer in women (Arbyn et al., 2020). Although the incidence of cervical cancer has decreased over the last years (Arbyn et al., 2011; Van Dyne et al., 2018) due to the implementation of screening methods (Brisson et al., 2020) and it may decrease in the following years due to vaccination (Bruni et al., 2021; Falcaro et al., 2021), an estimated 570,000 women were diagnosed with cervical cancer worldwide in 2018 (Bray et al., 2018). Moreover, the incidence of non-cervical cancers has increased in recent years. While in cervical cancer HPV prevalence is close to 100%, in other HPV-associated anogenital cancers viral prevalence rates differ according to the anatomical site: anus (88%; Alemany et al., 2015), vagina (74%; Alemany et al., 2014), penis (33%; Alemany et al., 2016), vulva (29%; de Sanjosé et al., 2013), and oropharynx (29–70%; Stein et al., 2015).

HPV high-risk types (HR-HPV) include predominantly, alpha 9 (HPV 16/31/33/35/52/58), alpha 7 (HPV 18/39/45/59/68), alpha 6 (HPV 56/66), and alpha 5 (HPV 51) genus, but HPV16 is by far the most common HR-HPV type, which contributes to 70–75% of all cervical cancers and is found in 40–60% of cervical intraepithelial neoplasia 2 (CIN2+; Bzhalava et al., 2013). However, only 5% of persistent HPV16 infections will evolve to high-grade lesions, and from those, a small proportion will progress to invasive cancer. Although it remains unclear why some HPV16 infections progress while others are cleared spontaneously, viral genome variability has been described as a key factor that could play a crucial role in the progression toward high-grade lesion or invasive cancer risk (Cullen et al., 2015). HPV16 was classified accordingly to viral genome variability in different lineages (A, B, C, and D) and sublineages (A1-4, B1-3, C1, D1-3) by Burk et al. (2013). HPV16-A lineage is the most prevalent type worldwide, while HPV16-D is the most aggressive type associated with cervical cancer risk (Gheit et al., 2011; Mirabello et al., 2016; Clifford et al., 2019).

In the 90's, the HPV genotype and HPV16 variants were determined according to the L1—Open Reading Frame (ORF) region that was amplified and sequenced (Ho et al., 1991; Chen et al., 2005). The implementation of New Generation Sequencing (NGS) techniques allowed us to perform bulk experiments and obtain longer sequences beyond the L1 ORF. Full viral genome sequencing resulted in the discovery of more lineages and genome variants (Burk et al., 2013). High-throughput sequencing as Illumina or Ion Torrent (Cullen et al., 2015) methods leads us to read the full viral genome. Before estimating the similarity between genomes, sequence samples are aligned to the reference HPV16 sequence (NCBI genome IDs: NC_001526.4). Then, a Maximum Likelihood Tree (MLT) is built altogether with a set of known-lineage HPV genomes used as a reference to assign specific lineages (Smith et al., 2011). New samples are placed on the phylogenetic tree according to their similarity with the reference sequences. Finally, the researcher manually assigns a lineage for the sample of interest, looking at where the sample has been located on the phylogenetic tree.

However, since the current method uses the entire genome sequence, poor coverage samples and samples showing gaps or missing fragments are difficult to classify. Building a phylogenetic tree is a time-consuming method when the sample size is big, which may take a long time to process depending on the computer used and finally, the lineage assignment is done manually. As MLT classification is directly influenced by the operator's expertise, reproducibility and standardization of the method may vary. To improve the HPV16 lineage assessment, we propose a new model that uses a few positions on the HPV16 genome to assess lineage and it does not require visual control, which makes the process faster and reproducible.

In this study, we describe a new code that can be used to efficiently assign HPV16 lineages. Using a Genome-Wide Association Study (GWAS), we tested all the positions of the HPV16 genome that are known to be unique to a single lineage or sublineage. Then, using machine learning algorithms, we trained and tested different models using reference and known samples for these positions. The code has been developed with the R language and it has been validated with more than 4,000 HPV16 genomes. Having a fast and efficient method for assigning HPV linages will help clinics to provide better-informed prognoses and help to define screening and treatment decision strategies.

MATERIALS AND METHODS

Samples

HPV16 genome sequences were used to find the lineage-specific SNPs and to build the model to assess lineage. Reference samples were obtained from two different sets of known-lineage HPV16 genomes: one set was described by Burk ($n = 46$; Smith et al., 2011) and the other was obtained from the Papillomavirus genome database (PAVE) webpage ($n = 10$; **Supplementary Material 1**). To define the lineage-specific positions for HPV16A, HPV16B, HPV16C, and HPV16D and to build the training-test models we used the reference samples and all the complete HPV16 genomes from NCBI ($n = 588$), downloaded from NCBI nucleotide dataset by keyword search "txid333760 complete genome;" Species: Viruses; Molecular types: Genomic DNA/RNA; Sequence type: Nucleotide accessed on July 30, 2021.

Validation of the model was performed with two different sets of samples, the first set of 1,028 HPV16 samples collected and sequenced in our laboratory, and the second set of 3,898 samples (which included the complete genomes and other almost complete genomes) were downloaded from NCBI nucleotide dataset by keyword search txid333760; Species: Viruses; Molecular types: Genomic DNA/RNA; Sequence type: Nucleotide; Release Date: From 0000/01/01 to 2022/03/24; Sequence length: from 7,000 to 8,500; accessed on March 24, 2022.

All samples were aligned on the HPV16 reference genome (GenBank Accession code: K02718.1) with MAFFT (v7.475) software using "–add" and "–keeplength" options (Katoh et al., 2019). The HPV16 reference genome, which is the HPV16-A1 sublineage has been added to the reference sample set ($n = 57$).

Lineage Assessment

The HPV16 lineage was assigned to the 588 NCBI samples using the current lineage assignment process described by Burk (Burk et al., 2013; Cullen et al., 2015) based on phylogenetics, which we will henceforth call Maximum Likelihood Tree (MLT), as it is based on this the Maximum Likelihood algorithm. The process consists of building a phylogenetic tree with altogether known lineage sequences and samples of interest. Phylogenetic analysis was conducted using MEGAX (Tamura et al., 2021) (v10.2.4) using the 57 reference samples plus the 588 NCBI previously aligned samples. To build the phylogenetic tree, we first calculated the genomic variation in a group of sequences with the Maximum Likelihood statistical method applying the Tamura-Nei correction model for nucleotide substitution. The process was replicated 100 times with the bootstrapping method. Finally, a tree was built, and lineage was assigned to each sample accordingly to the closest reference sample and results were manually annotated. Not all samples were assigned to a

lineage, since some sequences were not placed in the main lineage branches, so they were classified as "n" or unknown lineage.

Detection of Main Nucleotides Related to Lineage

A Genomic Wide Association Study (GWAS; Manolio, 2010) was performed on the reference and NCBI sequences ($n = 645$) to find differences between lineages within the HPV16 genome. The 7,906 base pairs that make up the viral genome have been traced to detect mutations. Known positions with two or more alleles with a minimum variant frequency (MVF) of 0.05 and a call rate higher than 95% were called SNP candidates. A generalized linear model (GLM) with a binomial distribution and a logit function was used to test the relationship between each SNP candidate and the HPV16 lineage. *P*-values were adjusted by False Discovery Rate (FDR) and only SNPs with a *p*-value lower than 0.05 were considered significant.

Training a New Model to Assess Lineage

To assess lineage with the SNPs described in the previous step we opt for training-test models. Different algorithms had been used to train models: Random Forest (RF), Support Vector Machine (SVM) and K-nearest neighbor (KNN), and Classification and Regression Trees (CART). The model was built with a total of 646 samples, including the 588 NCBI complete HPV16 genomes, the 57 reference samples, and a new sample called the "n-sample." The n-sample had no information and was composed of 7,906 unknown nucleotides ("n"), to assign unknown lineage to those samples with poor coverage. The 80% ($n = 518$) of the samples had been used for training and testing the model, while the remaining 20% ($n = 128$) had been used for the validation. For a better estimation, samples have been randomly mixed 100 times creating different training and test groups with the k-fold cross-validation method, and the model has been trained and tested for each new dataset. Accuracy, Kappa constant, and the testing confusion matrixes have been used to compare models and to choose the best model for lineage assessment.

Validate the New Model

Finally, validation has been performed to test the new model with two datasets of samples. The model has been validated with 3,898 genomes downloaded from NCBI which included both complete genomes and almost complete genomes and with a dataset composed of 1,028 HPV16-positive samples, that were selected from the archive of HPV tumors collected for the RIS HPV TT, VVAP, and Head and Neck studies (De Sanjose et al., 2010) and coordinated by the Catalan Institute of Oncology (ICO). Formalin-fixed paraffin-embedded (FFPE) specimens were sequenced with the HPV16 assay designed for the Ion Torrent Sequencing platform, which covers more than 80% of the viral genome (Cullen et al., 2015). Therefore, this last step of validation has tried out the model with a set of incomplete genomes as the sequencing assay was designed to amplify low-quality archival DNA.

In both datasets, lineage was first assessed with MLT, to then compare the quality of the new lineage assessment done by the machine learning model. Both GWAS and the training-test model has been performed using R language under 3.6.3 version and the code is available on www.github/INCALAB-PREC/HPV16-linpred/.

RESULTS

A GWAS performed on 645 HPV16-reference genomes showed 56 SNPs that are unique for one or more HPV16-lineages: A, B, C, or D (**Figure 1A**). Significant SNPs were spread out into the full genome. Gene E1 had a total of 16 lineage-definers SNPs, followed by E2 (10 SNPs), L2 and URR (7 SNPs), E6 (4 SNPs), E5 and L1 (3 SNPs), E7 (1 SNP), and 5 SNPs were found in a non-codifying region. Most of the differences in nucleotides were found between A and D or C lineages.

The training-test models were built using the 80% ($n = 518$) of the HPV16 dataset randomly selected and considering only the 56 lineage-specific positions found in the GWAS. The 100 k-fold cross-validation method has been applied and the dataset has been resampled 100 times in train and test groups. Each new dataset group was trained and tested to improve the estimated values of the model. **Figure 1B** shows a comparison between the models used, revealing that the best model to assess HPV16 lineage was the Random Forest (RF) algorithm, with an accuracy of 0.99 (CI:95%), followed by Support Vector Machine (SVM) and K-nearest neighbor (KNN); with a mean accuracy of 0.98 (CI: 95%) for both. Validation of the models was performed with the remaining 20% of the dataset ($n = 128$). To build the confusion matrix, lineage was assessed using the three models (RF, SVM, and KNN) and individually compared with the lineage assessed by MLT. Random Forest was the model with less error since all the assessed lineages match with MLT and were selected for the next validation steps (**Figure 1C**). Despite the high accuracy of SVM and KNN models, both failed in one single sample.

Further validations were carried out with two independent set of samples, the first one included 1,028 HPV16 positive samples, whose genome was partially obtained from FFPE archive samples. Most of the high coverage samples were classified with the same lineage as the MLT method did, shown in green in the confusion matrix (**Figure 2A**). Only one sample was differently classified between models (in red). MLT lineage classification is a challenge in low coverage samples, since out of 1,028 samples only 569 (56.1%) could be evaluated. In contrast, RF model has been able to assess lineage in 943 (93.0%) of these sequences. Therefore, if the MLT model is considered as the reference method for assessing HPV16 lineage, the RF model has an error of 0.17%. A total of 375 samples with average coverage have been assessed for the first time (in blue). However, we have no way of confirming that these samples have been properly classified. Lineage could not be assessed in 84 samples by either method, which has been classified as "n" samples. The coverage of most of these samples is poor, although some samples with good coverage were found in the unclassified group.

To understand in which conditions the RF model can assign lineage, different statistical analyses had been performed. Lineage has been assessed with a median of 24 known SNPs out of the 56 lineage-specific SNPs in a single sample (in red), while the 84 sequences that no lineage could be assigned had <15 known

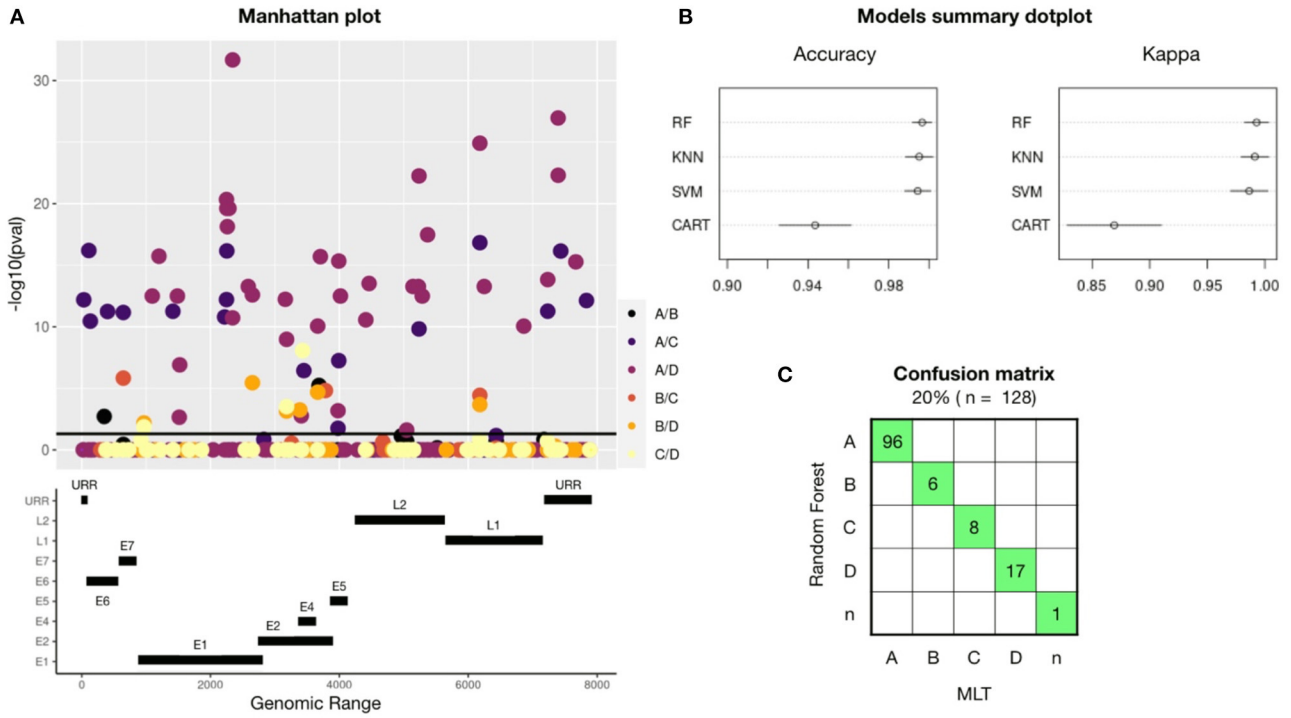


FIGURE 1 | A total of 56 lineage-definers SNPs were found in the GWAS, which were used to build four machine learning models, being RF the best model to predict HPV16 lineage. **(A)** Manhattan plot showing all the nucleotides in the HPV16 genome with any mutation (MAF > 0.05 and CR < 95%). SNP candidates have been colored differently according to which lineage they are specific for. Therefore, one SNP candidate can define two different lineages and it will be plotted twice. SNPs significantly related ($p < 0.05$) with any lineage are plotted above the black line. The genomic map under the Manhattan plot shows the positions of the HPV16 genes. **(B)** Accuracy and Kappa of the different training-test algorithms build to predict HPV16 lineage. Models used are Random Forest (RF), Support Vector Machine (SVM), K-nearest neighbor (KNN), and Command Assessment of Readiness and Training (CART). Kappa is a metric that compares an observed Accuracy with an expected Accuracy obtained by cross-validation (100-fold). When both, accuracy and kappa are 1, the model is perfect. **(C)** Confusion matrix for the Random Forest-based model using the validation set of samples. In green, samples that had been assessed equally with both methods: RF and Maximum Likelihood Tree (MLT).

SNPs (in blue; **Figure 2B**). Therefore, it must exist a minimum number of SNPs to successfully run the model. We have fixed a threshold at the intersection of the two density lines, which is 13 SNPs, and sequences with <13 out of the 56 lineage-definer SNPs will be directly assigned as unclassified lineage - "n." Applying the threshold, the confusion matrix slightly changes, losing a total of 102 samples that will be considered as "n" instead of the predicted lineage (**Figure 2C**). None of the samples equally assigned for both methods, RF and MLT, has been affected by the application of the threshold. After the correction, the percentage of lineage assignment decreased from 93.0 to 82.9%. Discarded samples included sequences of both good and bad coverage samples.

RF model has been validated with a second set that includes all the HPV16 genomes available in the NCBI dataset in March 2022. Lineage has been previously assessed by the MLT method and then has been assessed with the random-forest algorithm. The accuracy of the validation matrix is 98.9% ($p < 0.001$) and the error when assigning the lineage is <1.5%. However, a set of samples classified with the MLT method as HPV16-A lineage had been classified as B ($n = 28$) and D ($n = 12$) using the RF model. Discarding those who had <24 known SNPs the matrix improves, which indicates that the loss of certain SNPs after sequencing incomplete genomes, could influence the classification model accuracy. However, 22 samples are still classified as B instead of A (**Figure 2D**). This is probably due to a large number of HPV16-A samples included in the validation step compared to the other lineages. Although the error in lineage A classification is only 0.67%, most of the errors accumulate in B, which is the closest lineage to A, and overall, one of the less frequent lineages. In turn, all samples initially classified using MLT as B were well-classified as B using the RF model, which confirms that the model works to classify lineage B.

As the prevalence of HPV16-A lineage is higher in the world, for this reason, all the possible HPV16 datasets will have an important bias. We evaluate the model with a balanced dataset for each lineage. Sets of 200 lineage-balanced samples had been created randomly selecting 50 samples of each lineage from the full NCBI dataset ($n = 3{,}898$). The validation of the model, repeated with 10 different random sets shows an accuracy of 0.986 (95% CI: 0.958–0.997). The A-samples misclassification to B almost disappears (**Supplementary Material 2**).

In both pipelines, samples must be aligned to a reference genome. MAFFT takes an average of 2 min to align a total of 100 HPV16 genomes. It takes ∼40 min to calculate the

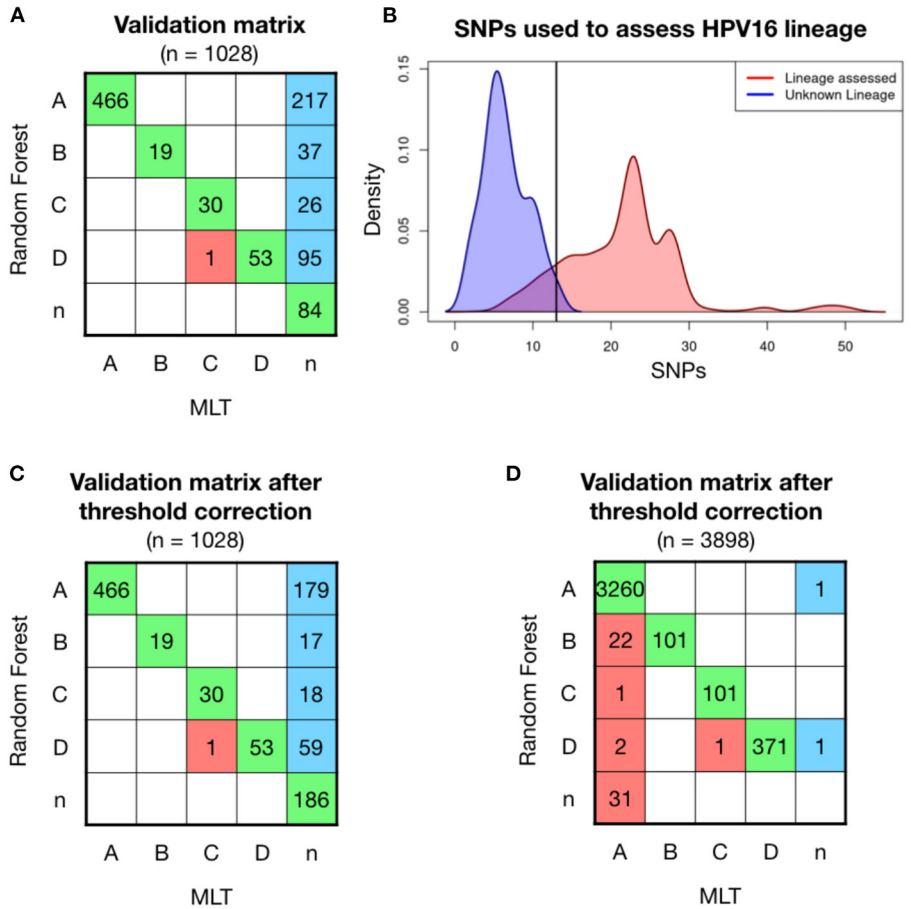


FIGURE 2 | Model validation on 1,028 patient' HPV16 sequences showed higher ratios of classification with Random Forest (RF) model than with Maximum Likelihood Tree (MLT). **(A)** RF validation was performed on 1,028 samples and shown in a confusion matrix. Samples receiving the same classification from both pipelines are colored in green, while samples that are not classified with the same lineage are shown in red. In the last column, sequences that RF-based model could assign for the first time and MLT could not. **(B)** Density curves of the number of covered lineage specific SNPs for each sample in function if their lineage has been assessed by RF (red) or not (blue), shows that the smaller number of known SNPs makes more difficult for the RF model to assess lineage. The black line corresponds at the intersection point between the two densities curves, where we had defined a threshold, where samples with <13 SNPs will be considered as the unclassified lineage or "n." **(C)** Validation matrix after threshold correction, discarding all the samples that have less than 13 known lineage-definers SNPs. Notice that the threshold only changes the blue column, increasing the n-samples from 87 to 182. **(D)** Validation matrix using 3,898 HPV16 genomes available in the nucleotide database from NCBI. Samples with <24 lineage-dependent SNPs had been classified as n-sample.

distances between samples with the MLT algorithm and to build a phylogenetic tree (bootstrapping samples 100 times) (**Table 1**). Followed by the annotation step, where the operator annotates manually the lineage by looking at the phylogenetic tree, which may take between 30 and 40 min depending on the skills of the worker. Using the developed code in this project, it only takes 0.97 s (SD = 0.43, repeated 25 times) to load the samples in Fasta format, assess lineage with the RF model and annotate lineage. For 100 samples, the new RF pipeline is almost 40 times faster than the current MLT pipeline. By increasing the number of samples to be tested, the difference between models becomes much larger. To assign lineage in our 1,028 HPV16 genomes dataset, the RF model was almost 40,000 times faster than MLT, since the process to build the MLT and annotating lineage lasted approximately up to 30 h, while the RF model took only 2,81 s (SD = 0.15, repeated 10 times).

Sublineage A

From the reference genome set ($n = 645$), a total of 481 HPV16-A samples had been selected, all of them assessed with A-lineage by both models, MLT and RF. Nucleotide differences between 0.5 and 1% of the complete genomes are used to define the sublineages (Burk et al., 2013), and HPV16-A lineage is classified in A1, A2, A3, and A4 groups. As HPV16-A1, A2, and A3 sublineages are more similar to each other and have a similar contribution on HPV-associated cancers than A4, we decided to cluster them into a single group called A123. A total of 67 positions were classified as SNP candidates (CR > 95% and

MAF > 0.05), but the GWAS only assessed 17 significantly SNPs associated with A123 or A4 sublineage.

An 80% of the samples were used to build the models, and from the five machine-learning models used in this study, RF and KNN were the models with better results to predict sublineage A. KNN model obtained an accuracy of 0.979 (95% CI: 0.926–0.997), which showed similar values than RF with an accuracy of 0.968 (CI: 0.911–0.993). After resampling and building the model 100 times, models were validated other 20% of the samples ($n = 96$). The validation matrix showed two mismatches between KNN and MLT (**Figure 3A**), instead of the three mismatches produced by the RF model, even showing the same accuracy values. A second validation was performed with the patient's sequenced HPV16-A samples ($n = 466$) obtained from the project led by ICO (**Figure 3B**). The accuracy of predicting sublineage A123 or A4 was 0.939 (95% CI: 0.914–0.959), being lower than the lineage model accuracy.

The training-tests with an accuracy higher than 95% (RF, KNN, and SVM) were ensembled by the majority vote method. The ensemble model did not improved the KNN prediction (**Figure 3C**).

TABLE 1 | HPV16 lineage classification is faster with the Random Forest pipeline.

100 HPV16 samples	Current pipeline (MLT)		New pipeline (RF)	
	Software/ method	Time (min)	Software/ method	Time (min)
Alignment	MAFFT	2	MAFFT	2
Algorithm	MEGAX/MLT	40	R/RF	0.97 s
Annotation	Manually*	30–40*	R	

The time for both pipelines was calculated on a set of 100 HPV16 sequences and tracked in a computer with the following features: UBUNTU 20.04 with 4 GHz Intel Core i7 and 16 GB of RAM.

For both pipelines, samples were aligned on the reference HPV16 genome with MAFFT software using "—keeplength" function. For the current pipeline we used MEGAX software to calculate the distance between sequences with the Maximum Likelihood tree (MLT) method and to build a phylogenetic tree. The new pipeline has been developed with R language and uses the Random Forest (RF) algorithm from the "caret" library. While the R code generates an output with the samples ID and the assigned lineage, the current pipeline requires manual annotation and the estimated time may depend on the operator's skills.*

DISCUSSION

The HPV16 lineage classification needs to be more efficient if we ever want to implement it as triage or prognostic marker in the clinical setting. Here we describe a faster and automated new model based on machine learning that efficiently classifies HPV16 sequences into lineages and requires lower sequence coverage if compared with the current method.

The current classification model calculates the similarity between samples and reference HPV16 genomes using the Maximum Likelihood estimation to classify the sampled sequence into a given lineage. To work, the MLT algorithm requires, as input, the whole HPV16 genome (7,906 base pairs), therefore, sampled sequences with large uncovered regions cannot be assigned to any lineage. We performed a genomic wide association study in which we identified 56 SNPs that are HPV16-lineage specific. The subset of SNPs included in the RF model is mainly lineage definers, our results are in agreement with previously described studies using phylogenetic reconstruction and classification to assign HPV16 variants to clinical sample (Ou et al., 2021).

Working with 56 SNPs instead of the full genome sequence, we can develop more efficient and faster models than the current model used for HPV16 lineage classification. Among different training-test models used to assess lineage based on the 56 SNPs, Random Forest was the best one, with an accuracy close to 100%. Using the RF to classify more than 1,000 samples we could assign a lineage to 93% of the samples, whereas using MLT we assigned a lineage to 56.1%. If the MLT model is considered as the reference

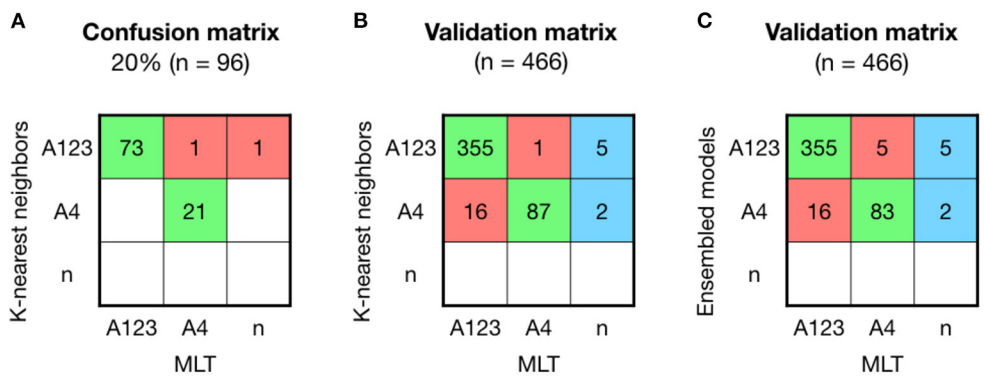


FIGURE 3 | Comparison between KNN, MLT and ensembled models to assign sublineage A shows good results but with higher error than the lineage model. **(A)** K-nearest neighbors (KNN) confusion matrix on the 20% of the HPV16-A reference sequences that were not used to build the model. **(B)** KNN validation was performed on 466 HPV16 patient's sequences and shown in a confusion matrix. Samples receiving the same classification from both pipelines are colored in green, while samples that are not classified with the same sublineage are shown in red. In the last column, sequences that KNN-based model could assign for the first time and Maximum Likelihood Tree (MLT) could not. **(C)** Ensembled model by majority vote was validated on the 466 HPV16-A patient's genomes.

method for HPV16 lineage classification since it does not exist another method, the new RF-based model would have an error between 0.17 and 1.4% according to both validation matrices. Therefore, from the 273 samples of first-time lineage assessed by RF in the 1,028 patient samples, we may assume that the error is similar, so there would be between 1 or 4 misclassified samples in this group.

Not all the SNPs are equally related to the lineage. A total of 20 out of 56 SNPs used in the model show higher Odds Ratio (OR) values when the relation between nucleotide and lineage is tested, thus lineage assessment could also work with a smaller set of SNPs in each sample. The density histogram showed that at least 13 SNPs must be known to assess lineage with the RF model, in consequence, samples with <13 known SNPs will be considered non-classified samples to avoid errors in low coverage samples. Besides the reduction of data required, if compared to the MLT pipeline, the RF model also allows a much faster process that does not require manual annotation. The RF model is 40 times faster than the MLT model.

Sequencing is becoming affordable to most laboratories, and consequently becoming a part of the clinical setting; however, it generates large amounts of data that may be difficult to analyze, besides being time-consuming. The new model we present here allows a straightforward assignment of HPV16 sequence alignment of virtually all sampled sequences.

The main limitation of this study is that we did not test our model for all sublineages, the training-test models could be only applied for A sublineage. Further studies should investigate the mismatched samples in order to unveil any potential limitation of the RF model for assigning HPV16 lineages. Our model can be implemented to classify HPV genotypes and other HPV lineages. Thus, samples from cervical and anogenital sites that are positive for any HPV type could be assigned to a specific lineage.

Having a fast and efficient method for assigning HPV linages may allow better-informed prognosis and may better guide doctors on the best course for women showing an HPV16 positive test or individuals with HPV positive pre-neoplastic lesions and high-grade lesions. Most of the current screening algorithms, using HPV as a primary test, define that HPV16 positive women should be referred directly to colposcopy, while more than 95% of these infections will be cleared spontaneously during the next 12 months. The identification of HPV16-positive women with a high risk of progression is a key point to develop new diagnostic tools for improving screening or diagnostic specificity avoiding unnecessary methods.

In addition, the computational model described in this work would be easily implementable in a user-friendly software or web interface, which will make easier the introduction of HPV16 lineage classification in the clinical setting.

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 authors.

ETHICS STATEMENT

The study was approved by the Ethics Committee of Hospital Universitari de Bellvitge. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LA-P and MP conceived of the presented idea. LA-P developed the theory and performed the computations. LA contributed to the conceptualization of the work. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by a grant from the Instituto de Salud Carlos III (Spanish Government) through the projects PI17/00123 (Co-funded by European Regional Development Fund. ERDF, a way to build Europe) and CIBERESP CB06/02/0073, and the Secretariat for Universities and Research of the Department of Business and Knowledge of the Government of Catalonia grants to support the activities of research groups 2017SGR1085. We thank the CERCA Programme/Generalitat de Catalunya for institutional support. None of these entities played a role in data collection, data analysis, data interpretation, or report writing. All authors had full access to all data in the study and had final responsibility for the decision to submit for publication.

ACKNOWLEDGMENTS

We would like to thank L. Mirabello and M. Schiffman from NCBI to sequence the HPV16 genomes used for the model validation and to Ana Esteban and Marleny Vergara to process all the samples from the RIS HPV TT, VVAP, and Head and Neck studies. I would also like to thank Marcia Triunfol at Publicase or her help with manuscript drafting.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.851841/full#supplementary-material

REFERENCES

Alemany, L., Cubilla, A., Halec, G., Kasamatsu, E., Quirós, B., Masferrer, E., et al. (2016). Role of human papillomavirus in penile carcinomas worldwide. *Eur. Urol.* 69, 953–961. doi: 10.1016/j.eururo.2015.12.007

Alemany, L., Saunier, M., Alvarado-Cabrero, I., Quirós, B., Salmeron, J., Shin, H.-R., et al. (2015). Human papillomavirus dna prevalence and type distribution in anal carcinomas worldwide. *Int. J. Cancer* 136, 98–107. doi: 10.1002/ijc.28963

Alemany, L., Saunier, M., Tinoco, L., Quirós, B., Alvarado-Cabrero, I., Alejo, M., et al. (2014). Large contribution of human papillomavirus in vaginal neoplastic

lesions: a worldwide study in 597 samples. *Eur. J. Cancer* 50, 2846–2854. doi: 10.1016/j.ejca.2014.07.018

Arbyn, M., Castellsagué, X., de Sanjosé, S., Bruni, L., Saraiya, M., Bray, F., et al. (2011). Worldwide burden of cervical cancer in 2008. *Ann. Oncol.* 22, 2675–2686. doi: 10.1093/annonc/mdr015

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., et al. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob. Health* 8, e191–e203. doi: 10.1016/S2214-109X(19)30482-6

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. *Cancer J. Clinicians* 68, 394–424. doi: 10.3322/caac.21492

Brisson, M., Kim, J. J., Canfell, K., Drolet, M., Gingras, G., Burger, E. A., et al. (2020). Impact of hpv vaccination and cervical screening on cervical cancer elimination: a comparative modelling analysis in 78 low-income and lower-middle-income countries. *Lancet* 395, 575–590. doi: 10.1016/S0140-6736(20)30068-4

Bruni, L., Saura-Lázaro, A., Montoliu, A., Brotons, M., Alemany, L., Diallo, M. S., et al. (2021). Hpv vaccination introduction worldwide and who and unicef estimates of national hpv immunization coverage 2010–2019. *Prev. Med.* 144:106399. doi: 10.1016/j.ypmed.2020.106399

Burk, R. D., Harari, A., and Chen, Z. (2013). Human papillomavirus genome variants. *Virology* 445, 232–243. doi: 10.1016/j.virol.2013.07.018

Bzhalava, D., Guan, P., Franceschi, S., Dillner, J., and Clifford, G. (2013). A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. *Virology* 445, 224–231. doi: 10.1016/j.virol.2013.07.015

Chen, Z., Terai, M., Fu, L., Herrero, R., DeSalle, R., and Burk, R. D. (2005). Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. *J. Virol.* 79, 7014–7023. doi: 10.1128/JVI.79.11.7014-7023.2005

Clifford, G. M., Tenet, V., Georges, D., Alemany, L., Pavón, M. A., Chen, Z., et al. (2019). Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: whole viral genome sequences from 7116 hpv16-positive women. *Papillomavirus Res.* 7, 67–74. doi: 10.1016/j.pvr.2019.02.001

Cullen, M., Boland, J. F., Schiffman, M., Zhang, X., Wentzensen, N., Yang, Q., et al. (2015). Deep sequencing of hpv16 genomes: a new high-throughput tool for exploring the carcinogenicity and natural history of hpv16 infection. *Papillomavirus Res.* 1, 3–11. doi: 10.1016/j.pvr.2015.05.004

de Sanjosé, S., Alemany, L., Ordi, J., Tous, S., Alejo, M., Bigby, S. M., et al. (2013). Worldwide human papillomavirus genotype attribution in over 2000 cases of intraepithelial and invasive lesions of the vulva. *Eur. J. Cancer* 49, 3450–3461. doi: 10.1016/j.ejca.2013.06.033

De Sanjose, S., Quint, W. G., Alemany, L., Geraets, D. T., Klaustermeier, J. E., Lloveras, B., et al. (2010). Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* 11, 1048–1056. doi: 10.1016/S1470-2045(10)70230-8

Falcaro, M., Castañon, A., Ndlela, B., Checchi, M., Soldan, K., Lopez-Bernal, J., et al. (2021). The effects of the national hpv vaccination programme in england, UK, on cervical cancer and grade 3 cervical intraepithelial neoplasia incidence: a register-based observational study. *Lancet* 398, 2084–2092. doi: 10.1016/S0140-6736(21)02178-4

Gheit, T., Cornet, I., Clifford, G. M., Iftner, T., Munk, C., Tommasino, M., et al. (2011). Risks for persistence and progression by human papilloma virus type 16 variant lineages among a population-based sample of danish women. *Cancer Epidemiol. Prev. Biomark.* 20, 1315–1321. doi: 10.1158/1055-9965.EPI-10-1187

Ho, L., Chan, S., Chow, V., Chong, T., Tay, S., Villa, L. L., et al. (1991). Sequence variants of human papillomavirus type 16 in clinical samples permit verification and extension of epidemiological studies and construction of a phylogenetic tree. *J. Clin. Microbiol.* 29, 1765–1772. doi: 10.1128/jcm.29.9.1765-1772.1991

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformat.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Manolio, T. A. (2010). Genome wide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176. doi: 10.1056/NEJMra0905980

Mirabello, L., Yeager, M., Cullen, M., Boland, J. F., Chen, Z., Wentzensen, N., et al. (2016). Hpv16 sublineage associations with histology-specific cancer risk using hpv whole-genome sequences in 3200 women. *J. Natl. Cancer Instit.* 2016:108:djw100. doi: 10.1093/jnci/djw100

Ou, Z., Chen, Z., Zhao, Y., Lu, H., Liu, W., Li, W., et al. (2021). Genetic signatures for lineage/sublineage classification of HPV16, 18, 52 and 58 variants. *Virology* 553, 62–69. doi: 10.1016/j.virol.2020.11.003

Smith, B., Chen, Z., Reimers, L., Van Doorslaer, K., Schiffman, M., DeSalle, R., et al. (2011). Sequence imputation of hpv16 genomes for genetic association studies. *PLoS ONE* 6:e21375. doi: 10.1371/journal.pone.0021375

Stein, A. P., Saha, S., Kraninger, J. L., Swick, A. D., Yu, M., Lambertg, P. F., et al. (2015). Prevalence of human papillomavirus in oropharyngeal cancer: a systematic review. *Cancer J.* 21:138. doi: 10.1097/PPO.0000000000000115

Tamura, K., Stecher, G., and Kumar, S. (2021). Mega11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Van Dyne, E. A., Henley, S. J., Saraiya, M., Thomas, C. C., Markowitz, L. E., and Benard, V. B. (2018). Trends in human papillomavirus–associated cancers?united states, 1999–2015. *Morbidity Mortality Weekly Rep.* 67:918. doi: 10.15585/mmwr.mm6733a2

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 | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 11 August 2022
doi: 10.3389/frai.2022.876100

Application of Artificial Intelligence to Plasma Metabolomics Profiles to Predict Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer

Ehsan Irajizad[1], Ranran Wu[2], Jody Vykoukal[2], Eunice Murage[2], Rachelle Spencer[2], Jennifer B. Dennison[2], Stacy Moulder[3], Elizabeth Ravenberg[3], Bora Lim[4], Jennifer Litton[3], Debu Tripathym[3], Vicente Valero[3], Senthil Damodaran[3], Gaiane M. Rauch[5], Beatriz Adrada[6], Rosalind Candelaria[6], Jason B. White[3], Abenaa Brewster[2], Banu Arun[3], James P. Long[1], Kim Anh Do[1], Sam Hanash[2] and Johannes F. Fahrmann[2]**

[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2] Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [3] Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [4] Breast Cancer Research Program, Baylor College of Medicine, Houston, TX, United States, [5] Department of Abdominal Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [6] Department of Breast Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

There is a need to identify biomarkers predictive of response to neoadjuvant chemotherapy (NACT) in triple-negative breast cancer (TNBC). We previously obtained evidence that a polyamine signature in the blood is associated with TNBC development and progression. In this study, we evaluated whether plasma polyamines and other metabolites may identify TNBC patients who are less likely to respond to NACT. Pre-treatment plasma levels of acetylated polyamines were elevated in TNBC patients that had moderate to extensive tumor burden (RCB-II/III) following NACT compared to those that achieved a complete pathological response (pCR/RCB-0) or had minimal residual disease (RCB-I). We further applied artificial intelligence to comprehensive metabolic profiles to identify additional metabolites associated with treatment response. Using a deep learning model (DLM), a metabolite panel consisting of two polyamines as well as nine additional metabolites was developed for improved prediction of RCB-II/III. The DLM has potential clinical value for identifying TNBC patients who are unlikely to respond to NACT and who may benefit from other treatment modalities.

Keywords: triple-negative breast cancer, biomarkers, artificial intelligence, deep-learning model, neoadjuvant chemotherapy, prediction

OPEN ACCESS

Edited by:
Jaume Reventos, Institut d'Investigacio Biomedica de Bellvitge (IDIBELL), Spain

Reviewed by:
Miguel Abal, Health Research Institute of Santiago de Compostela (IDIS), Spain
David Sarrio, Centro de Investigación Biomédica en Red del Cáncer (CIBERONC), Spain

***Correspondence:**
Sam Hanash
shanash@mdanderson.org
Johannes F. Fahrmann
jffahrmann@mdanderson.org

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *15 February 2022*
Accepted: *03 May 2022*
Published: *11 August 2022*

Citation:
Irajizad E, Wu R, Vykoukal J, Murage E, Spencer R, Dennison JB, Moulder S, Ravenberg E, Lim B, Litton J, Tripathym D, Valero V, Damodaran S, Rauch GM, Adrada B, Candelaria R, White JB, Brewster A, Arun B, Long JP, Do KA, Hanash S and Fahrmann JF (2022) Application of Artificial Intelligence to Plasma Metabolomics Profiles to Predict Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. Front. Artif. Intell. 5:876100. doi: 10.3389/frai.2022.876100

INTRODUCTION

Triple-negative breast cancer (TNBC) accounts for ∼15–20% of breast cancers and represents a heterogeneous subtype characterized by high pathological grade, strong invasiveness, local recurrence, high metastasis rate, and poor prognosis (Foulkes et al., 2010). TNBCs are defined based on the lack expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal

growth factor receptor type 2 (HER2) and are thus not amenable to endocrine therapy or therapies targeted to the HER2 receptor type (Foulkes et al., 2010). Chemotherapy remains the mainstay of systemic treatment, typically consisting of anthracycline and taxane-based chemotherapy regimens (Foulkes et al., 2010; Bianchini et al., 2022). Platinum-based neoadjuvant chemotherapy has been shown to increase pathological complete response (pCR) rates compared to platinum-free neoadjuvant chemotherapy. However, platinum-based treatment is associated with higher rates of toxicity and treatment discontinuation, and the optimal integration of platinum-based agents remains controversial (Poggio et al., 2018). The addition of immunotherapy has shown promise with recent Phase III clinical trials demonstrating that the addition of the anti-PD-L1 inhibitor atezolizumab or the anti-PD1 inhibitor pembrolizumab with chemotherapy improved pCR compared to chemotherapy alone in patients with TNBC (Schmid et al., 2018, 2020).

In the curative neoadjuvant setting, a pCR after neoadjuvant chemotherapy (NACT) in TNBC is associated with improved long-term survival yielding estimated 10-year relapse survival rates of 86% (Symmans et al., 2017). However, up to 60% of patients will have residual disease after receiving standard NACT and are at an elevated risk of poor outcome, with reported 10-year estimated relapse survival rates of 81, 55, and 23% for TNBC patients with a residual cancer burden (RCB) index of I, II, and III, respectively (Huober et al., 2010; Symmans et al., 2017; Schmid et al., 2020). Currently, there is a paucity of biomarkers that can reliably identify TNBC patients that will have poor response to NACT.

Polyamines, including putrescine, spermidine, and spermine, are polycationic alkylamines that are essential for eukaryotic cell growth. Dysregulation of polyamine metabolism is frequent in cancer and polyamines have been reported to play central roles in neoplastic transformation and tumor progression (Park and Igarashi, 2013; Casero et al., 2018; Chia et al., 2022). We previously obtained evidence that increased plasma levels of the acetylated polyamine diacetylspermine (DAS) in TNBC was prognostic for poor progression-free survival and overall survival. Specifically, we found that elevated levels of plasma DAS to be prognostic for worse 5-year metastasis free survival and poor 5-year overall survival in newly-diagnosed treatment naïve TNBC patients (Fahrmann et al., 2020).

Here, we tested the utility of plasma polyamines for identifying subjects who will be insensitive to NACT as part of a comprehensive plasma metabolomics profiling. We further applied artificial intelligence to plasma metabolic profiles and, using a deep-learning model (DLM), established a metabolite biomarker panel consisting of two polyamines as well as nine additional metabolites for prediction of response to NACT.

MATERIALS AND METHODS

Specimen Sets

Patients with stage I–III TNBC enrolled in the prospective, Institutional Review Board (IRB)-approved, clinical study, "A Robust TNBC Evaluation framework to Improve Survival" (ARTEMIS, NCT02276443), were included in this study. Briefly, the ARTEMIS trial included treatment-naïve patients with localized TNBC (stage I-III) that underwent a pre-treatment ultrasound with biopsy following by 4 cycles of Adriamycin-cyclophosphamide (AC) chemotherapy. The outcome of the molecular characterization from the pre-treatment biopsy in combination with response assessment (clinical exam/diagnostic imaging, after 4 cycles of AC) were used to identify chemotherapy-insensitive disease and to inform the second phase of neoadjuvant therapy. Patients deemed to have chemo-sensitive disease after 4 cycles of AC (≥70% volumetric reduction by ultrasound after 4 cycles of AC) were recommended to undergo standard paclitaxel-based chemotherapy as the second phase of their NACT consisting of 4 cycles or weekly for 12 doses. Patients with TNBC predicted to be chemo-insensitive (≤70% volumetric reduction by ultrasound after 4 cycles of AC) were offered therapy on clinical trials using targeted therapy in combination with chemotherapy based on the specific molecular characteristics of their tumor as the second phase of their therapy with dose regimens varying depending on therapy. Response to neoadjuvant therapy was determined using the residual cancer burden (RCB) index (Symmans et al., 2007). The specimen set consisted of pre-treatment EDTA plasma from 88 patients who received standard-of-care NACT; 62 of the 88 patients had a second plasma sample available after four cycles of AC. Detailed patient and tumor characteristics are provided in **Table 1**.

TABLE 1 | Patient and tumor characteristics.

	TNBC[†] cases	Controls
N	88	167
Age, mean +/– SD	50 +/– 11	58 +/– 9
Stage, *N* (%)		
I	9 (10)	–
II	64 (73)	–
III	15 (17)	–
RCB status, *N* (%)		
0	48 (55)	–
I	14 (16)	–
II	21 (24)	–
III	5 (6)	–

[†] *All TNBC patients received NACT; plasma samples were collected pre-treatment. TNBC, triple-negative breast cancer; RCB, Residual Cancer Burden.*

EDTA plasma from cancer-free women ($n = 167$) were obtained from the MD Anderson Cancer Center (MDACC) Longitudinal High-Risk Cohort initiated September 1st, 2011, for the prospective follow-up of cancer-free high-risk women seen in the MDACC Cancer Prevention Center (IRB protocol LAB07-0086).

Immunohistochemistry

Immunohistochemical (IHC) staining for Ki-67 was performed on unstained 4-μm-thick tissue sections that had been prepared from a representative paraffin block of tumor in each case. IHC

staining for Ki-67 was performed using the polymeric biotin-free horseradish peroxidase method on the Leica Microsystems Bond III autostainer (Leica Microsystems, Buffalo Grove, IL, USA). The slides were incubated at 60°C for 25 min. Following heat-induced epitope retrieval with Tris-EDTA buffer for 20 min at 100°C, slides were incubated with mouse monoclonal antibody to Ki-67 (clone MIB-1, Dako; 1:100). The Refine Polymer Detection kit was used to detect bound antibody, with 3,3-diaminobenzidine serving as the chromogen (Leica Microsystems). For Ki-67, the percentage of any nuclear staining of any intensity in the tumor cells was recorded.

Metabolomic Analysis

Sample Preparation

Plasma metabolites were extracted from pre-aliquoted biospecimens (15 μL) with 45 μL of LCMS grade methanol (ThermoFisher) in a 96-well microplate (Eppendorf). Plates were heat sealed, vortexed for 5 min at 750 rpm, and centrifuged at 2,000 × g for 10 mins at room temperature. The supernatant (30 μL) was transferred to a 96-well plate, leaving behind the precipitated protein. The supernatant was further diluted with 60 μL of 100 mm ammonium formate, pH3 (Fisher Scientific). For Hydrophilic Interaction Liquid Chromatography (HILIC) positive ion analysis, 15 μL of the supernatant and ammonium formate mix were diluted with 195 μL of 1:3:8:144 water (GenPure ultrapure water system, Thermofisher): LCMS grade methanol (ThermoFisher): 100 mm ammonium formate, pH3 (Fisher Scientific): LCMS grade acetonitrile (ThermoFisher). For C18 analysis, 15 μL of the supernatant and ammonium formate mix were diluted with 90 μL water (GenPure ultrapure water system, ThermoFisher) for positive ion mode. Each sample solution was transferred to 384-well microplate (Eppendorf) for LCMS analysis.

Untargeted Analysis of Primary Metabolites and Biogenic Amines

Untargeted metabolomics analysis was conducted on Waters Acquity™ UPLC system with 2D column regeneration configuration (I-class and H-class) coupled to a Xevo G2-XS quadrupole time-of-flight (qTOF) mass spectrometer as previously described (Fahrmann et al., 2019, 2020, 2021a,b). Chromatographic separation was performed using HILIC (Acquity™ UPLC BEH amide, 100 Å, 1.7 μm 2.1 × 100 mm, Waters Corporation, Milford, U.S.A) and C18 (Acquity™ UPLC HSS T3, 100 Å, 1.8 μm, 2.1 × 100 mm, Water Corporation, Milford, U.S.A) columns at 45°C.

Quaternary solvent system mobile phases were (A) 0.1% formic acid in water, (B) 0.1% formic acid in acetonitrile and (D) 100 mm ammonium formate, pH 3. Samples were separated on the HILIC using the following gradient profile at 0.4 mL/min flow rate: (95% B, 5% D) linear change to (70% A, 25% B and 5% D) over 5 min; 100% A for 1 min; and 100% A for 1 min. For C18 separation, the chromatography gradient was as follows at 0.4 mL/min flow rate: 100% A with a linear change to (5% A, 95% B) over 5 min; (95% B, 5% D) for 1 min; and 1 min at (95% B, 5% D).

A binary pump was used for column regeneration and equilibration. The solvent system mobile phases were (A1) 100 mm ammonium formate, pH 3, (A2) 0.1% formic in 2-propanol and (B1) 0.1% formic acid in acetonitrile. The HILIC column was stripped using 90% A2 for 5 min at 0.25 mL/min flow rate, followed by a 2 min equilibration using 100% B1 at 0.3 mL/min flow rate. Reverse phase C18 column regeneration was performed using 95% A1, 5% B1 for 2 min followed by column equilibration using 5% A1, 95% B1 for 5 min at 0.4 mL/min flow rate.

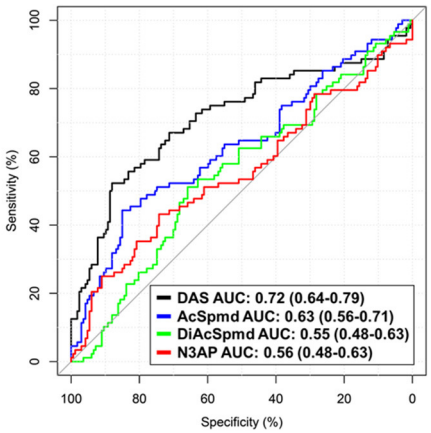
Mass Spectrometry Data Acquisition

Mass spectrometry data was acquired using 'sensitivity' mode in positive and negative electrospray ionization mode within 50–800 Da range. For the electrospray acquisition, the capillary voltage was set at 1.5 kV (positive), sample cone voltage 30 V, source temperature at 120°C, cone gas flow 50 L/h and desolvation gas flow rate of 800 L/h with scan time of 0.5 sec in continuum mode. Leucine Enkephalin; 556.2771 Da (positive) was used for lockspray correction and scans were performed at 0.5 sec. The injection volume for each sample was 6 μL. The acquisition was carried out with instrument auto gain control to optimize instrument sensitivity over the samples acquisition time.

Data were processed using Progenesis QI (Non-linear, Waters). Peak picking and retention time alignment of LC-MS and MSe data were performed using Progenesis QI software (Non-linear, Waters). Data processing and peak annotations were performed using an in-house automated pipeline as previously described (Fahrmann et al., 2019, 2020, 2021a; Vykoukal et al., 2020). Annotations were determined by matching accurate mass and retention times using customized libraries created from authentic standards and by matching experimental tandem mass spectrometry data against the NIST MSMS, LipidBlast or HMDB v3 theoretical fragmentations. To correct for injection order drift, each feature was normalized using data from repeat injections of quality control samples collected every 10 injections throughout the run sequence. Measurement data were smoothed by Locally Weighted Scatterplot Smoothing (LOESS) signal correction (QC-RLSC) as previously described. Values are reported as ratios relative to the median of historical quality control reference samples run with every analytical batch for the given analyte (Fahrmann et al., 2019, 2020, 2021a; Vykoukal et al., 2020).

Statistical Analysis

A deep learning algorithm employing all quantified metabolites with tuned hyperparameters using the grid search approach (Candel et al., 2016) was run 20 times, and the relevance importance score for each metabolite was calculated using the Gedeon method (Gedeon, 1997). Metabolites were prioritized based on consistently exhibiting a relative importance score >0.5. Ten models, including deep learning, random forest, ensemble learning and gradient boosting method algorithms, incorporating eleven metabolites were assessed for distinguishing responder/partial responders (RCB-0/I) from



Metabolite	Cases	Controls	AUC (95% CI)	P-value	Sens @ 95% Spec	Spec @ 95% Sens
DAS	88	167	0.72 (0.64-0.79)	< 0.0001	23.9	6.0
AcSpmd	88	167	0.63 (0.56-0.71)	0.0006	18.2	6.0
DiAcSpmd	88	167	0.55 (0.48-0.63)	0.18	1.1	6.6
N3AP	88	167	0.56 (0.48-0.63)	0.15	9.1	0.0

FIGURE 1 | Predictive performance of individual polyamines for distinguishing treatment-naïve TNBC cases from healthy controls. Table beneath shows AUC (95% CI), Wilcoxon rank sum test 2-sided *P*-values as well as sensitivity and specificity estimates at 95% specificity/sensitivity thresholds of individual polyamines for distinguishing TNBC cases from healthy controls.

non-responders (RCB-II/III). The predictability, reliability, and stability of the models in the training set was evaluated using a 5-fold cross validation as well as through introducing perturbations (e.g., random sample selection with replacement) to the dataset.

Model discrimination was assessed based on receiver operating characteristic curve (ROC), as well as sensitivity and specificity estimates. The 95% confidence intervals (CI) for AUCs were estimated using the Delong method (DeLong et al., 1988). All modeling was performed using the H_2O package and R statistical program (Candel et al., 2016).

RESULTS

Plasma Polyamine Levels in Triple-Negative Breast Cancer

Using mass spectrometry, we first assessed polyamines levels in plasmas from 88 newly diagnosed treatment-naïve TNBC cases and 167 cancer-free women enrolled in the MDACC Longitudinal High-Risk Cohort (**Table 1**). A total of four polyamines, acetylspermidine (AcSpmd), diacetylspermidine (DiAcSpmd), diacetylspermine (DAS), and N-(3-acetamidopropyl)pyrrolidin-2-one (N3AP) were detected and quantified. Of these, AcSpmd and DAS were statistically significantly elevated (Wilcoxon rank sum test 2-sided $p < 0.01$) in case plasmas compared to controls (**Figure 1**). DAS exhibited the highest discrimination performance for distinguishing all cases from controls with an AUC of 0.72 (95% C.I.: 0.64–0.79) (**Figure 1**).

Association of Polyamines With RCB Status

All 88 TNBC patients were treated with AC in the neoadjuvant setting. A subset of 62 (70.5%) had a complete pathological response (pCR/RCB-0) or minimal residual disease (RCB-I) following NACT, whereas 26 (29.5%) had a moderate to extensive tumor burden (RCB-II and III) (**Table 1**). Pathological response tended to be associated with tumor stage and % tumoral Ki-67 staining positivity, albeit not statistically significant (**Supplementary Figure 1**).

Elevated pre-treatment plasma levels of AcSpmd, N3AP, DiAcSpmd and DAS were associated with higher odds of RCB-II/III following NACT [adjusted ORs of 1.24 (95% CI: 0.76–2.04), 1.33 (95% CI: 0.79–2.46), 1.15 (95% CI: 0.71–1.85) and 1.26 (95% CI: 0.71–1.91) per standard deviation increase, respectively] (**Table 2**).

Applying Artificial Intelligence to Metabolic Profiles to Develop a Combination Rule for Prediction of RCB-II/III

Complementary to the four polyamines, untargeted metabolomics analyses of these plasmas yielded an additional 82 uniquely annotated metabolite features (**Supplementary Table 1**). To prioritize metabolites associated with response to NACT for model building, relative importance scores were calculated using the Gedeon method (Gedeon, 1997) and metabolites were selected that constantly showed a relative important score of > 0.5 (*see Methods*). This approach resulted in 11 cancer-related metabolites, consisting of two polyamines, two lipids, three amino acids, a purine catabolite, and two indole-derivatives (**Supplementary Table 2**). Spearman correlation analyses indicated low to moderate associations between these metabolites (**Supplementary Figure 2**).

We next sought to develop a machine learning algorithm that incorporated the eleven metabolites for predicting RCB-II/III. For model building, we tested 10 different machine learning algorithms (**Table 3**). Of these, a deep learning model (DLM) with 3 hidden layers and 20 nodes in each layer achieved the highest predictive performance with an AUC of 0.97 (95% CI: 0.93–1.00) with 85% sensitivity at 95% specificity for identifying RCB-II/III (**Figure 2**). Notably, the DLM yielded an AUC of 0.76 (95% CI: 0.65–0.87) with 48% sensitivity at 95% specificity for distinguishing TNBC cases with residual disease (RCB-I/II/III) from those that achieved a pCR (**Supplementary Table 3**). To assess model reproducibility and stability, we introduced perturbation into the dataset (e.g., random selection with replacement) and re-evaluated model performance, the results of which showed that the DLM was robust (**Supplementary Table 4**).

We additionally assessed the predictive performance of the DLM model using plasma samples collected during NACT from a subset of TNBC patients ($n = 62$). The DLM model showed an AUC of 0.74 (95% CI: 0.62–0.87) with 21% sensitivity at 95% specificity for RCB-II/III (**Figure 3**).

TABLE 2 | Performance estimates of polyamines for distinguishing RCB-II/III from RCB-0/I.

Polyamines	AUC (95% CI)	Sensitivity @ 95% sen	Specificity @ 95% spec	Odds ratio	Adjusted Odds ratio[†]
AcSpmd	0.59 (0.46–0.71)	0.12 (0.00–0.23)	0.19 (0.10–0.31)	1.34 (0.85–2.10)	1.24 (0.76–2.04)
N3AP	0.55 (0.42–0.68)	0.12 (0.00–0.27)	0.10 (0.02–0.32)	1.34 (0.86–2.26)	1.33 (0.79–2.46)
DiAcSpmd	0.54 (0.40–0.67)	0.08 (0.00–0.23)	0.08 (0.00–0.26)	1.15 (0.72–1.80)	1.15 (0.71–1.85)
DAS	0.58 (0.46–0.71)	0.15 (0.00–0.31)	0.24 (0.10–0.39)	1.39 (0.89–2.23)	1.26 (0.77–2.10)

Area under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, odds ratios, and adjusted odds ratios estimates and corresponding 95% confidence intervals of individual polyamines are shown. AcSpmd, acetylspermidine; N3AP, N-(3-acetamidopropyl)pyrrolidin-2-one; DiAcSpmd, diacetylspermidine; DAS, diacetylspermine. [†] *age and stage were included as covariables in adjusted odd ratios.*

TABLE 3 | Performance of the different learning models in the training set.

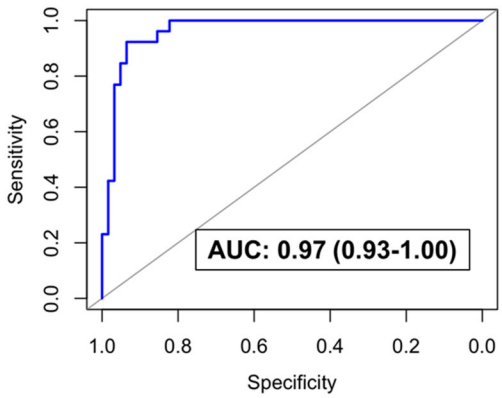
Model	Hyper parameters	AUC	Log loss	AUCpr	Mean per class error	RMSE
Deep learning model	Activation: Maxout, hidden layers:3, number of nodes in each layer: 20	0.97	0.396	0.62	0.249	0.339
Deep learning model	Activation: Maxout, hidden layers:2, number of nodes in each layer = 1	0.86	0.412	0.61	0.268	0.385
Deep learning model	Activation: Tanh, hidden layers: 1, number of nodes in each layer = 3	0.78	0.429	0.60	0.283	0.393
Deep learning model	Activation: Tanh hidden layers:1, number of nodes in each layer: 1	0.72	0.438	0.60	0.297	0.399
GLM	Family: Binomial	0.68	0.585	0.53	0.331	0.47
Gradient boosting method	Number of tree: 50, Maximum depth:6	0.61	0.692	0.53	0.342	0.499
Distributed random forest (DRF)	–	0.55	0.709	0.51	0.49	0.507
Extremely randomized trees (XRT)	–	0.53	0.787	0.45	0.429	0.537
StackedEnsemble	Ensemble models (best of each family): GLM, Deep Learning, Random Forest, Gradient Boost Method	0.53	2.274	0.46	0.421	0.671
Extreme gradient boosting	–	0.52	4.198	0.47	0.481	0.66

AUC, Area under the ROC curve; AUCpr, Area under precision-recall curve; RMSE, root-mean-square deviation; GLM, generalized linear model; DRF, Distributed Random Forest; XRT, Extremely Randomized Trees.

DISCUSSION

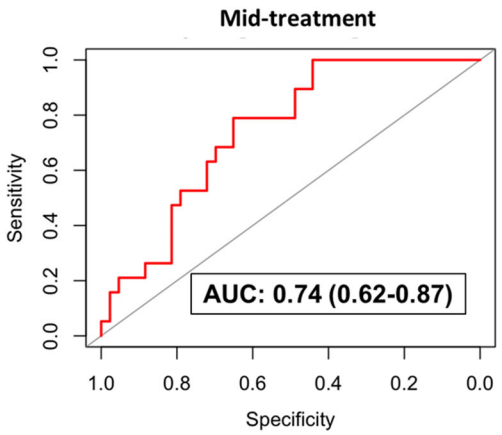
The heterogeneity of TNBC results in a spectrum of responses to NACT with pCR being achieved in only a subset of patients (Sikov et al., 2015; Gamucci et al., 2018). Several methods have been used to measure and predict residual disease during course of treatment, including ultrasound, MRI scans, histopathology; however, none have yet achieved adequate performance to predict response to NACT (Croshaw et al., 2011; Shin et al., 2011; Ono et al., 2012; De Los Santos et al., 2013; Leon-Ferre et al., 2018). Here, we applied artificial intelligence to metabolomic profiles of TNBC patient plasmas obtained prior to NACT and, using a DLM, establish a blood-based polyamine-centric metabolite panel that is predictive of non-response to NACT. The metabolite panel may be implemented in the clinical setting to stratify TNBC patients who are at high-risk of being non-responsive to NACT and who may benefit from alternate treatment modalities. Conversely, TNBC patients who are likely to be responsive to NACT may potentially benefit from dose de-escalation, thereby permitting management of treatment-associated toxicity.

The metabolite panel consisted of several cancer-relevant metabolites including the acetylated polyamines DAS and AcSpmd, which were found to be elevated in TNBC patients who were less likely to respond to NACT. Elevated levels of acetylated polyamines in various biofluids including urine, plasma, and serum, have been shown to report on cancer status (Park and Igarashi, 2013; Wikoff et al., 2015; Fahrmann et al., 2019, 2020, 2021b). Targeting of cancer cell polyamine



	DLM performance
	Baseline
AUC (95% CI)	0.97 (0.93 - 1.00)
Sensitivity @ 95% spec	0.85 (0.27 - 1.00)
Specificity @ 95% sen	0.87 (0.77 - 0.98)
AUC (Stage I/II)	0.97 (0.94 - 1.00)
AUC (Stage III)	0.96 (0.88 - 1.00)
Odds Ratio	20.94 (6.94 - 92.42)
Adjusted Odds Ratio†	42.08 (10.2 - 339.79)

FIGURE 2 | ROC curve for the DLM for distinguishing TNBC patients that went on to have RCB-II/III following NACT from those that had RCB-0/I. Table provides tabulated performance estimates of the DLM. † Age and stage were included as covariables in adjusted odd ratios.



	DLM performance
	Mid-treatment
AUC	0.74 (0.62 - 0.87)
Sensitivity @ 95% spec	0.21 (0.00 - 0.42)
Specificity @ 95% sen	0.44 (0.28 - 0.63)
AUC (Stage I/II)	0.74 (0.60 - 0.88)
AUC (Stage III)	0.74 (0.43 - 1.00)
Odds Ratio	1.92 (1.12 - 3.44)
Adjusted Odds Ratio	2.10 (1.19 - 3.96)

FIGURE 3 | ROC curve for the DLM for distinguishing TNBC patients that went on to have RCB-II/III following NACT from those that had RCB-0/I using plasmas collected after four cycles of AC. Table provides tabulated performance estimates of the DLM. † Age and stage were included as covariables in adjusted odd ratios.

metabolism *via* small molecule inhibitors has been proposed for anti-cancer therapy for cancer, including TNBC (Casero et al., 2018; Geck et al., 2020; Capellen et al., 2021). Acetylation of polyamines is mediated by spermidine/spermine N1-acetyltransferase 1 (SAT1) (Pegg, 2013; Fahrmann et al., 2020). Our prior investigations demonstrated that oncogenic MYC regulates transcription of polyamine metabolizing enzymes ornithine decarboxylase (ODC1), spermidine synthase (SRM), and spermine synthase (SMS) in TNBC, and that elevated intracellular polyamine levels induce expression of SAT1 (Pegg, 2013) resulting in elevated cancer cell biosynthesis and secretion of acetylated polyamines (Fahrmann et al., 2020). We further reported that plasma polyamines, particularly DAS, are associated with TNBC development and progression (Fahrmann et al., 2020). Given our prior findings, we posit that the elevation in polyamines may underly an aggressive subtype of TNBC (Fahrmann et al., 2020) that is less likely to respond to NACT.

Elevated serum levels of urate, a purine catabolite, are also reported to be prognostic for TNBC recurrence and poor overall survival (Ackermann and Tardito, 2019; Gong et al., 2021). Lysophosphatidylethanolamines and lauroylcarnitine are associated with cancer metabolic plasticity and fatty acid oxidation (Melone et al., 2018). We have previously reported that JAK/STAT3-mediated fatty acid oxidation promotes chemoresistance in TNBC (Chakraborty et al., 2016; Wang et al., 2018). Methylhistidine has been shown to be elevated in serum of TNBC patients who had an cPR following NACT (He et al., 2021). TNBC cells are reported to exhibit a glutamine-dependent phenotype; promoting survival advantage as well as chemo-resistance (Kung et al., 2011; Lampa et al., 2017).

Remarkably, among the metabolites in the metabolite panel were two microbiome-related metabolites, indoleacrylic acid (IAA) and indole-acetylaldehyde (IAALD). IAA and IAALD are produced through the catabolism of tryptophan by the gut microbes (Vujkovic-Cvijin et al., 2013). Increasing evidence implicates that alterations in the microbiome influence resistance to anticancer treatment, including conventional chemotherapy, immunotherapy, radiotherapy, and surgery (Pryor et al., 2020; Garajová et al., 2021; Pandey and Umar, 2021). The relationship

between changes in the microbiome and response to NACT warrants further investigation.

On balance, limitations to our study include limited sample availability and lack of external validation. To assess for potential overfitting, we tested the model by introducing perturbation (e.g., random selection and replacement) to the dataset and re-evaluated performance, the results of which demonstrated that our model was robust. We performed further validation using available samples and found that the metabolite panel provided good classifier performance for distinguishing individuals with RCB-II/III following four cycles of NACT from those with RCB-0/I, thus providing an independent validation. We note that attenuation of model performance after four cycles of NACT could be attributable to elevations in plasma metabolites consistent with chemotherapy-induced cancer cell death and turnover. Additionally, the relative cost-effectiveness of using the metabolite panel for risk-based prediction of non-responsiveness to NACT needs to be considered compared to other clinical predictors (Croshaw et al., 2011; Shin et al., 2011; Ono et al., 2012; De Los Santos et al., 2013; Leon-Ferre et al., 2018).

In conclusion, using a deep learning model, we developed a blood-based metabolite panel and that offers potential utility for identifying TNBC patients who are at high-risk of being non-responsive to NACT and who may benefit from more personalized treatment modalities.

DATA AVAILABILITY STATEMENT

The data presented in the study is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, where it has been assigned Study ID ST002235. The data can be accessed directly via its Project DOI: http://dx.doi.org/10.21228/M8HX4C.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by patients with stage I–III TNBC enrolled in the prospective, Institutional Review Board (IRB)-approved, clinical study, A Robust TNBC Evaluation framework to Improve Survival (ARTEMIS, NCT02276443), were included in this study. Cancer-free women ($n = 167$) were obtained from the MD Anderson Cancer Center (MDACC) Longitudinal High-Risk Cohort initiated September 1st, 2011, for the prospective follow-up of cancer-free high-risk women seen in the MDACC Cancer Prevention Center (IRB protocol LAB07-0086). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EI, SH, and JF: conceptualization. EI, RW, EM, JD, JLo, KD, and JF: methodology. EI and JF: validation, visualization, writing-original draft preparation, and formal analysis. ER, BL, JL, DT, VV, SD, GR, BAd, BAr, RC, JW, and AB: resources. RW, EM, JD, JW, and JF: data curation. RW, JV, EM, RS, JD, SM, ER, BL, JLi, DT, VV, SD, GR, BAd, RC, JW, AB, BAr, JLo, KD, and SH: writing-review and editing. SH and JF: supervision. JLo, KD, and SH: funding acquisition.

FUNDING

The work presented was supported by Cancer Prevention & Research Institute of Texas (CPRIT) (RP180505) (SH), the generous philanthropic contributions to the University of Texas MD Anderson Cancer Center Moon Shots Program™, the Little Green Book Foundation, and the Still Water Foundation. KD was partially supported by a Cancer Center Support Grant NCI Grant P30 CA016672, NIH grants UL1TR003167, 5R01GM122775. JLo was partially supported by the National Cancer Institute and the National Center for Advancing Translational Sciences of the NIH (P30CA016672 and CCTS UL1TR003167).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.876100/full#supplementary-material

REFERENCES

Ackermann, T., and Tardito, S. (2019). Cell culture medium formulation and its implications in cancer metabolism. *Trends Cancer*. 5, 329–332. doi: 10.1016/j.trecan.2019.05.004

Bianchini, G., De Angelis, C., Licata, L., and Gianni, L. (2022). Treatment landscape of triple-negative breast cancer - expanded options, evolving needs. *Nat. Rev. Clin. Oncol.* 19, 91–113. doi: 10.1038/s41571-021-00565-2

Candel, A., Parmar, V., LeDell, E., Arora, A. (2016). Deep learning with H2O. *H2O ai Inc.* 2016, 1–21. Available online at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf

Capellen, C. C., Ortega-Rodas, J., Morwitzer, M. J., Tofilau, H. M. N., Dunworth, M., Casero, R. A., et al. (2021). Hyperglycemic conditions proliferate triple negative breast cancer cells: role of ornithine decarboxylase. *Breast Cancer Res. Treat.* 190, 255–264. doi: 10.1007/s10549-021-06388-0

Casero, R. A., Murray Stewart, T., and Pegg, A. E. (2018). Polyamine metabolism and cancer: treatments, challenges and opportunities. *Nat. Rev. Cancer*. 18, 681–695. doi: 10.1038/s41568-018-0050-3

Chakraborty, S., Ghosh, S., Banerjee, B., Santra, A., Adhikary, A., Misra, A. K., et al. (2016). Phemindole, a synthetic di-indole derivative maneuvers the store operated calcium entry (SOCE) to induce potent anti-carcinogenic activity in human triple negative breast cancer cells. *Front. Pharmacol.* 7, 114. doi: 10.3389/fphar.2016.00114

Chia, T. Y., Zolp, A., and Miska, J. (2022). Polyamine immunometabolism: central regulators of inflammation, cancer and autoimmunity. *Cells*. 11, 5. doi: 10.3390/cells11050896

Croshaw, R., Shapiro-Wright, H., Svensson, E., Erb, K., and Julian, T. (2011). Accuracy of clinical examination, digital mammogram, ultrasound, and MRI in determining postneoadjuvant pathologic tumor response in operable breast cancer patients. *Annal. Surgic. Oncol.* 18, 3160–3163. doi: 10.1245/s10434-011-1919-5

De Los Santos, J. F., Cantor, A., Amos, K. D., Forero, A., Golshan, M., Horton, J. K., et al. (2013). Magnetic resonance imaging as a predictor of pathologic response in patients treated with neoadjuvant systemic treatment for operable breast cancer: translational breast cancer research consortium trial 017. *Cancer*. 119, 1776–1783. doi: 10.1002/cncr.27995

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. doi: 10.2307/2531595

Fahrmann, J. F., Bantis, L. E., Capello, M., Scelo, G., Dennison, J. B., Patel, N., et al. (2019). A Plasma-derived protein-metabolite multiplexed panel for early-stage pancreatic cancer. *J. Natl. Cancer Inst*. 111, 372–379. doi: 10.1093/jnci/djy126

Fahrmann, J. F., Irajizad, E., Kobayashi, M., Vykoukal, J., Dennison, J. B., Murage, E., et al. (2021a). A MYC-driven plasma polyamine signature for early detection of ovarian cancer. *Cancers*. 13, 4. doi: 10.3390/cancers13040913

Fahrmann, J. F., Vykoukal, J., Fleury, A., Tripathi, S., Dennison, J. B., Murage, E., et al. (2020). Association between plasma diacetylspermine and tumor spermine synthase with outcome in triple-negative breast cancer. *J. Natl. Cancer Inst*. 112, 607–616. doi: 10.1093/jnci/djz182

Fahrmann, J. F., Wasylishen, A. R., Pieterman, C. R. C., Irajizad, E., Vykoukal, J., Murage, E., et al. (2021b). A blood-based polyamine signature associated with men1 duodenopancreatic neuroendocrine tumor progression. *J. Clin. Endocrinol. Metab*. 106, e4969–e4980.doi: 10.1142/s0129065797000227

Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *New Engl. J. Med*. 363, 1938–1948. doi: 10.1056/NEJMra1001389

Gamucci, T., Pizzuti, L., Sperduti, I., Mentuccia, L., Vaccaro, A., Moscetti, L., et al. (2018). Neoadjuvant chemotherapy in triple-negative breast cancer: A multicentric retrospective observational study in real-life setting. *J. Cell. Physiol*. 233, 2313–2323. doi: 10.1002/jcp.26103

Garajová I., Balsano, R., Wang, H., Leonardi, F., Giovannetti, E., Deng, D., et al. (2021). The role of the microbiome in drug resistance in gastrointestinal cancers. *Exper. Rev. Anticancer Therapy*. 21, 165–176. doi: 10.1080/14737140.2021.1844007

Geck, R. C., Foley, J. R., Murray Stewart, T., Asara, J. M., Casero, R. A. Jr., and Toker, A. (2020). Inhibition of the polyamine synthesis enzyme ornithine decarboxylase sensitizes triple-negative breast cancer cells to cytotoxic chemotherapy. *J. Biol. Chem*. 295, 6263–6277. doi: 10.1074/jbc.RA119.012376

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *Int. J. Neural Syst*. 8, 209–218. doi: 10.1142/S0129065797000227

Gong, Y., Ji, P., Yang, Y.-S., Xie, S., Yu, T.-J., Xiao, Y., et al. (2021). Metabolic-pathway-based subtyping of triple-negative breast cancer reveals potential therapeutic targets. *Cell Metabol*. 33, 51–64. doi: 10.1016/j.cmet.2020.10.012

He, X., Gu, J., Zou, D., Yang, H., Zhang, Y., Ding, Y., et al. (2021). NMR-based metabolomics analysis predicts response to neoadjuvant chemotherapy for triple-negative breast cancer. *Front. Mol. Biosci*. 8, 21. doi: 10.3389/fmolb.2021.708052

Huober, J., von Minckwitz, G., Denkert, C., Tesch, H., Weiss, E., Zahm, D. M., et al. (2010). Effect of neoadjuvant anthracycline–taxane-based chemotherapy in different biological breast cancer phenotypes: overall results from the GeparTrio study. *Breast Cancer Res. Treat*. 124, 133–140. doi: 10.1007/s10549-010-1103-9

Kung, H. N., Marks, J. R., and Chi, J. T. (2011). Glutamine synthetase is a genetic determinant of cell type-specific glutamine independence in breast epithelia. *PLoS Genet*. 7, e1002229. doi: 10.1371/journal.pgen.1002229

Lampa, M., Arlt, H., He, T., Ospina, B., Reeves, J., Zhang, B., et al. (2017). Glutaminase is essential for the growth of triple-negative breast cancer cells with a deregulated glutamine metabolism pathway and its suppression synergizes with mTOR inhibition. *PLoS One*. 12, e0185092. doi: 10.1371/journal.pone.0185092

Leon-Ferre, R. A., Polley, M.-Y., Liu, H., Gilbert, J. A., Cafourek, V., Hillman, D. W., et al. (2018). Impact of histopathology, tumor-infiltrating lymphocytes, and adjuvant chemotherapy on prognosis of triple-negative breast cancer. *Breast Cancer Res. Treat*. 167, 89–99. doi: 10.1007/s10549-017-4499-7

Melone, M. A. B., Valentino, A., Margarucci, S., Galderisi, U., Giordano, A., and Peluso, G. (2018). The carnitine system and cancer metabolic plasticity. *Cell Death Dis*. 9, 1–12. doi: 10.1038/s41419-018-0313-7

Ono, M., Tsuda, H., Shimizu, C., Yamamoto, S., Shibata, T., Yamamoto, H., et al. (2012). Tumor-infiltrating lymphocytes are correlated with response to neoadjuvant chemotherapy in triple-negative breast cancer. *Breast Cancer Res. Treat*. 132, 793–805. doi: 10.1007/s10549-011-1554-7

Pandey, K., and Umar, S. (2021). Microbiome in drug resistance to colon cancer. *Curr. Opin. Physiol*. 23, 100472. doi: 10.1016/j.cophys.2021.100472

Park, M. H., and Igarashi, K. (2013). Polyamines and their metabolites as diagnostic markers of human diseases. *Biomol. Ther. (Seoul)*. 21, 1–9. doi: 10.4062/biomolther.2012.097

Pegg, A. E. (2013). Toxicity of polyamines and their metabolic products. *Chem. Res. Toxicol*. 26, 1782–1800. doi: 10.1021/tx400316s

Poggio, F., Bruzzone, M., Ceppi, M., Pond,é N. F., La Valle, G., Del Mastro, L., et al. (2018). Platinum-based neoadjuvant chemotherapy in triple-negative breast cancer: a systematic review and meta-analysis. *Annal. Oncol*. 29, 1497–1508. doi: 10.1093/annonc/mdy127

Pryor, R., Martinez-Martinez, D., Quintaneiro, L., and Cabreiro, F. (2020). The role of the microbiome in drug response. *Ann. Rev. Pharmacol. Toxicol*. 60, 417–435. doi: 10.1146/annurev-pharmtox-010919-023612

Schmid, P., Adams, S., Rugo, H. S., Schneeweiss, A., Barrios, C. H., Iwata, H., et al. (2018). Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *New Engl. J. Med*. 379, 2108–2121. doi: 10.1056/NEJMoa1809615

Schmid, P., Salgado, R., Park, Y. H., Muñoz-Couselo, E., Kim, S. B., Sohn, J., et al. (2020). Pembrolizumab plus chemotherapy as neoadjuvant treatment of high-risk, early-stage triple-negative breast cancer: results from the phase 1b open-label, multicohort KEYNOTE-173 study. *Annal. Oncol*. 31, 569–581. doi: 10.1016/j.annonc.2020.01.072

Shin, H., Kim, H., Ahn, J., Kim, S., Jung, K., Gong, G., et al. (2011).Comparison of mammography, sonography, MRI and clinical examination in patients with locally advanced or inflammatory breast cancer who underwent neoadjuvant chemotherapy. *Br. J. Radiol*. 84, 612–620. doi: 10.1259/bjr/74430952

Sikov, W. M., Berry, D. A., Perou, C. M., Singh, B., Cirrincione, C. T., Tolaney, S. M., et al. (2015). Impact of the addition of carboplatin and/or bevacizumab to neoadjuvant once-per-week paclitaxel followed by dose-dense doxorubicin and cyclophosphamide on pathologic complete response rates in stage II to III triple-negative breast cancer: CALGB 40603 (Alliance). *J. clinic. oncol*. 33, 13. doi: 10.1200/JCO.2014.57.0572

Symmans, W. F., Peintinger, F., Hatzis, C., Rajan, R., Kuerer, H., Valero, V., et al. (2007). Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J. Clinic. Oncol*. 25, 4414–4422. doi: 10.1200/JCO.2007.10.6823

Symmans, W. F., Wei, C., Gould, R., Yu, X., Zhang, Y., Liu, M., et al. (2017). Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J. Clinic. Oncol*. 35, 1049–1060. doi: 10.1200/JCO.2015.63.1010

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., et al. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Sci. Translat. Med*. 5, 1931. doi: 10.1126/scitranslmed.3006438

Vykoukal, J., Fahrmann, J. F., Gregg, J. R., Tang, Z., Basourakos, S., Irajizad, E., et al. (2020). Caveolin-1-mediated sphingolipid oncometabolism underlies a metabolic vulnerability of prostate cancer. *Nat Commun*. 11, 4279. doi: 10.1038/s41467-020-17645-z

Wang, T., Fahrmann, J. F., Lee, H., Li, Y. J., Tripathi, S. C., Yue, C., et al. (2018). JAK/STAT3-Regulated fatty acid β-oxidation is critical for breast cancer stem cell self-renewal and chemoresistance. *Cell Metab*. 27, 136–150. doi: 10.1016/j.cmet.2017.11.001

Wikoff, W. R., Hanash, S., DeFelice, B., Miyamoto, S., Barnett, M., Zhao, Y., et al. (2015). Diacetylspermine is a novel prediagnostic serum biomarker for non-small-cell lung cancer and has additive performance with pro-surfactant protein B. *J. Clin. Oncol*. 33, 3880–3886. doi: 10.1200/JCO.2015.61.7779

Conflict of Interest: Author SM declares honoraria from Novartis, Pfizer and research funding from Oncothyreon (Inst), Pfizer (Inst), Novartis (Inst), Genentech (Inst), Takeda (Inst), Bayer (Inst), EMD Serono (Inst), Genentech (Inst); travel, accommodations, expenses: Novartis, Pfizer. Author JLi declares consulting or advisory role: Pfizer, AstraZeneca, Medivation/Pfizer; speakers' bureau from Physician Education Resource, UpToDate, Med Learning, Group, Medscape; research funding from Novartis (Inst), Bristol-Myers Squibb (Inst), Genentech (Inst), Pfizer (Inst), EMD Serono (Inst), Jounce Therapeutics (Inst), GlaxoSmithKline (Inst), Medivation/Pfizer (Inst); patents, royalties,

other intellectual property: UpToDate; travel, accommodations, expenses: Physician Education Resource, Med Learning Group, Medscape. Author DT declares consulting or advisory role: AstraZeneca, Genomic Health, Gilead Sciences Inc, GlaxoSmithKline, Novartis Pharma, OncoPep, Pfizer; research funding: Pfizer, Novartis Pharma. Author VV declares honoraria: Genentec; consulting or advisory role: Genentech; travel, accommodations, expenses: Genentech. Author SD declares honoraria: Novartis; consulting or advisory role: Tempus, Taiho Pharmaceutical, Pfizer; research funding: EMD Serono, Guardant Health; travel, accommodations; expenses: Phillips Gilmore Oncology Communications. Author BAd declares consulting or advisory role: Bright Pink, AbbVie; research funding: AbbVie (Inst), PharmaMar (Inst), AstraZeneca (Inst), InVitae (Inst); travel, accommodations, expenses: AstraZeneca.

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.

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 | Frontiers in Oncology

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fonc.2022.895515

OPEN ACCESS

EDITED BY
Mónica Hebe Vazquez-Levin,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

REVIEWED BY
Zhendong Jin,
Second Military Medical University, China
Hui Jiang,
Naval Medical University, China

*CORRESPONDENCE
Carlos H. F. Chan
carloshfchan@gmail.com

SPECIALTY SECTION
This article was submitted to Gastrointestinal Cancers: Hepato Pancreatic Biliary Cancers, a section of the journal Frontiers in Oncology

RECEIVED 13 March 2022
ACCEPTED 09 November 2022
PUBLISHED 08 December 2022

CITATION
Chang J, Liu Y, Saey SA, Chang KC, Shrader HR, Steckly KL, Rajput M, Sonka M and Chan CHF (2022) Machine-learning based investigation of prognostic indicators for oncological outcome of pancreatic ductal adenocarcinoma. *Front. Oncol.* 12:895515. doi: 10.3389/fonc.2022.895515

Machine-learning based investigation of prognostic indicators for oncological outcome of pancreatic ductal adenocarcinoma

Jeremy Chang[1], Yanan Liu[2], Stephanie A. Saey[1], Kevin C. Chang[1], Hannah R. Shrader[1,3], Kelsey L. Steckly[3], Maheen Rajput[4], Milan Sonka[2,5] and Carlos H. F. Chan[1,3]*

[1]Department of Surgery, University of Iowa Hospitals and Clinics, Iowa City, IA, United States, [2]Iowa Initiative for Artificial Intelligence, University of Iowa, Iowa City, IA, United States, [3]Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, United States, [4]Department of Radiology, University of Iowa Hospitals and Clinics, Iowa City, IA, United States, [5]Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA, United States

Introduction: Pancreatic ductal adenocarcinoma (PDAC) is an aggressive malignancy with a poor prognosis. Surgical resection remains the only potential curative treatment option for early-stage resectable PDAC. Patients with locally advanced or micrometastatic disease should ideally undergo neoadjuvant therapy prior to surgical resection for an optimal treatment outcome. Computerized tomography (CT) scan is the most common imaging modality obtained prior to surgery. However, the ability of CT scans to assess the nodal status and resectability remains suboptimal and depends heavily on physician experience. Improved preoperative radiographic tumor staging with the prediction of postoperative margin and the lymph node status could have important implications in treatment sequencing. This paper proposes a novel machine learning predictive model, utilizing a three-dimensional convoluted neural network (3D-CNN), to reliably predict the presence of lymph node metastasis and the postoperative positive margin status based on preoperative CT scans.

Methods: A total of 881 CT scans were obtained from 110 patients with PDAC. Patients and images were separated into training and validation groups for both lymph node and margin prediction studies. Per-scan analysis and per-patient analysis (utilizing majority voting method) were performed.

Results: For a lymph node prediction 3D-CNN model, accuracy was 90% for per-patient analysis and 75% for per-scan analysis. For a postoperative margin prediction 3D-CNN model, accuracy was 81% for per-patient analysis and 76% for per-scan analysis.

Discussion: This paper provides a proof of concept that utilizing radiomics and the 3D-CNN deep learning framework may be used preoperatively to improve the prediction of positive resection margins as well as the presence of lymph node metastatic disease. Further investigations should be performed with larger cohorts to increase the generalizability of this model; however, there is a great promise in the use of convoluted neural networks to assist clinicians with treatment selection for patients with PDAC.

KEYWORDS

machine learning, neural network, pancreatectomy, pancreatic cancer, surgical outcome, radiomics

Introduction

Pancreatic cancer is currently the third leading cause of cancer-related death in Western societies with an average annual incidence rate of 12.9 cases per 100,000 but a disproportionately high mortality rate of 10.9 deaths per 100,000 (1). Pancreatic ductal adenocarcinoma (PDAC) is the most common type of pancreatic cancer. At the time of diagnosis, only ~10% of PDAC are localized since small early cancers are often asymptomatic and left undiagnosed (2). Although surgery is the only curative treatment for PDAC, only 15%–20% of patients are candidates for surgical resection due to late presentation (2). The decision for upfront surgical resection followed by adjuvant chemotherapy vs. neoadjuvant treatment followed by surgical resection is based on both the anatomy of the tumor (i.e., vascular involvement) and risk stratification/prognostic features including the health condition, blood tumor markers, and lymph node involvement on imaging studies (3). Currently, computerized tomography (CT) scan is the most utilized modality for the evaluation of PDAC with a specificity and sensitivity of ~89% and ~90% (2). The nodal status is a well-established prognostic indicator for both overall survival and disease recurrence (4–6). Although CT image resolution has increased dramatically over the last two decades, the ability of a CT scan to assess both vascular invasion (sensitivity and specificity of 60% and 94%) and the nodal status remains suboptimal (positive predictive value and negative predictive value are 68% and 43.1%, respectively), as it may heavily depend on physician experience (1, 7). An automated prediction model for the presence of lymph node metastatic disease may preoperatively aid in clinical decision-making.

The choice to undergo the upfront surgical treatment of PDAC is determined by the preoperative CT stratification of resectability that is dependent on tumor proximity to the surrounding vessels (portal vein, superior mesenteric vein, superior mesenteric artery, and celiac artery) (8). The impact of the R1 resection status (i.e., presence of microscopic disease), defined as the distance of a tumor from the resection margin of less than or equal to 1 mm, on overall survival and recurrence-free survival is controversial (9–11). However, recent studies have suggested that the presence of microscopic disease within 1 mm is associated with decreased overall survival and decreased disease-free survival in PDAC in comparison to R0 resection (i.e., free of cancer cells at the resection margin) (12). Hong et al. demonstrated that of patients with the designation of a "resectable" tumor based on preoperative CT imaging, only 73% of patients had postoperative R0 resection on pathology (13). Preoperative CT appeared to overpredict resectability in tumors with any level of portomesenteric vein abutment and for larger tumors greater than 4 cm (12). An enhanced preoperative prediction of the surgical margin status would allow for improved patient selection for upfront curative intent surgery and importantly direct patients with tumors more likely to have postoperative R1 or R2 resection to neoadjuvant chemotherapy.

Radiomics is a novel approach to medical imaging that abstracts vast amounts of qualitative imaging features using data-characterizing algorithms, converting medical images into big data (14). The basis of the application of radiomics is that distinct imaging features between disease forms may be used to predict a prognosis and a therapeutic response (15). With radiomics exponentially increasing the data obtained from medical imaging, there has been growing interest with utilizing artificial intelligence or machines learning models to provide techniques to analyze these image data (16). One such model demonstrating great utility is the convoluted neural network (CNN). CNNs contain multiple interconnected layers of artificial neurons whereby each neuron can take an input, perform a computation, and produce output, while learning increases its higher-level functions (17). CNNs have been utilized to investigate a number of medical imaging questions including segmentation (i.e., tumor vs. normal tissue (18)), disease classification (19), detection and localization (i.e., identification of cerebral microbleeds in MRI (20)), and

registration (i.e., integrating multiple scans of same patient (21)). Some examples include the following: Huang et al. have described that specific radiomic signatures differed between normal lymph nodes and lymph nodes with metastatic disease and that these differences allowed the creation of a nomogram for the prediction of the lymph node status in colorectal cancer (22). Chen et al. created hybrid many-objective radiomics and a three-dimensional CNN (3D-CNN) model to evaluate lymph node metastasis in head and neck cancers (23).

This paper proposes a novel machine learning predictive model, utilizing a 3D-CNN, to reliably predict the presence of lymph node metastasis and the postoperative positive margin status based on preoperative CT scans. This is the first deep learning predictive model for both lymph node disease in pancreatic cancer and the margin status based on preoperative imaging. Manual image segmentation was not performed allowing for an unbiased approach and a potential generalizability of the model to other abdominal/gastrointestinal cancers.

Materials and methods

Study population

The Biospecimen Procurement and Molecular Epidemiology Core (BioMER) is a shared core resource at the University of Iowa Holden Comprehensive Cancer Center that prospectively enrolls cancer patients into disease-specific MER patient cohorts annotated with clinicopathological, treatment, and outcome data. Within the gastrointestinal cancer cohort of the BioMER (GIMER), 462 patients were enrolled from 2015 to 2021. Study inclusion criteria included 1) having a diagnosis of pancreatic ductal adenocarcinoma by pathology, 2) receiving curative intent surgery, 3) available CT images prior to surgical intervention, and 4) available surgical pathologic data regarding the tumor margin status and lymph nodes. Positive margin was defined by the presence of cancer cells found within 1 mm from the inked resection margin. CT images and clinical and pathologic data were obtained from 110 patients (Table 1). A total of 881 CT scans were obtained. A patient's CT scan from a particular date may contain the images of arterial, venous, and delayed phases with different resolutions. For the purposes of subgrouping, the images from each individual phase are classified as "one" scan. Due to small patient numbers, each scan was treated independently. The patient cohort was divided into two groups, one for training and one for validation for each study algorithm, margin study, and lymph node study. The training vs. validation split was 59 patients (340 scans) vs. 20 patients (140 scans) for the lymph node study and 83 patients (629 scans) vs. 27 patients (252 scans) for the margin study. For the margin study, additional PDAC patients with surgeon-determined unresectable locally advanced disease on preoperative CT were included to provide additional control cases with positive margin to improve study power.

Development of machine learning algorithm

In collaboration with the Iowa Initiative for Artificial intelligence (IIAI), a 3D-CNN was developed for the purpose of image classification based on the lymph node disease or the margin status. In a basic sense, the CNN involves creating a scaffolding of computational "layers" stacked on one another whereby the outputs of terminal layers are built upon the inputs of the previous. The specific structuring of the number of layers and the type of layer (i.e., convolutional, pooling, and fully connected) based on the research question is where nuance arises. The goal was to learn a

TABLE 1 Study population characteristics.

	Lymph node study		Margin study	
	Training group (n=59)	Validation group (n=20)	Training group (n=83)*	Validation group (n=27)§
Age	66.1 [63.6–68.7]	62.5 [58.6–66.3]	65.8 [63.6–68.0]	64.2 [60.5–67.8]
Gender				
Male	27 (45.8%)	10 (50%)	44 (53%)	13 (48.1%)
Female	32 (54.2%)	10 (50%)	39 (47%)	14 (51.9%)
Pathological Stage				
Stage 0	0	1 (5%)	0	1 (3.7%)
Stage I	5 (8.5%)	2 (10%)	5 (6.0%)	2 (7.4%)
Stage II	54 (91.5%)	17 (85%)	55 (66.3%)	16 (59.3%)
Stage III	0	0	16 (19.3%)	6 (22.2%)
Stage IV	0	0	7 (8.4%)	2 (7.4%)
Positive Margin	12 (20.3%)	4 (20%)	36 (44.4%)	11 (40.7%)
Number of Images	340	140	629	252

*Includes 23 unresectable cases (these cases would yield a positive resection margin if they have undergone surgery).
§Includes 8 unresectable cases.

discriminative function, f ∈ {0, 1}, where 1 indicates lymph node metastasis or a positive margin and 0 otherwise.

The 3D-CNN utilized was modeled after that described by Zunair et al. (24). Like Zunair et al, this study utilized a 17-layer 3D-CNN. Four 3D convolutional layers are used with each convolutional layer followed by a max-pooling layer and a subsequent batch normalization layer creating a CON-MAXPOOL-BN module (24). The subsequent output runs through a global average pooling layer and then a dense layer. An effective dropout rate of 60% was utilized. A second dense layer was used to produce output consistent with the binary classification problem (Figure 1). The binary cross-entropy loss function was utilized within model learning to optimize the performance of the classification model. A total of 1,351,813 learnable parameters were present in this study. All codes were written and run utilizing Python (Python Software Foundation, Delaware, USA).

Data preparation

To decrease computational time, slice selection was performed. Each axial CT scan was analyzed, and the slices between an anatomical boundary of superior to the celiac artery takeoff to inferior to the renal vein were identified. Subsequently, each image was resized to a resolution of 128 × 128 × 64 pixels. Image intensity and parameters were normalized to a scale of (0, 1). The initial input for the first layer of the 3D-CNN model was resized CT scan.

Statistical analysis

The sensitivity, specificity, positive and negative predictive values, and accuracy of the model were evaluated on training and validation datasets. With the use of the Wilson–Brown Method with GraphPad Prism8 software, 95% confidence intervals (17) were determined. Receiver operating characteristic (25) curves were plotted for the per-patient analysis using the different cutoff values of percent-positive scan from per-scan analysis for each patient, and area-under-the-curve (26) analysis was performed using GraphPad Prism8 software. Algorithm prediction accuracy was displayed in the confusion matrix as appropriate.

Results

The clinical characteristics of study population are summarized in Table 1.

Lymph node metastasis predictive model

The training group consisted of 37 patients with lymph node metastasis and 22 patients without (total of 340 scans), and the validation group consisted of 15 patients with lymph node metastasis and 5 patients without (total of 140 scans). In per-scan analysis, the 3D-CNN model achieved a sensitivity of 93% (95%CI: 86%–97%) and a specificity of 42% (95%CI: 29%–56%) with an accuracy of prediction at 75% and a positive and negative predictive value of 74% (95%CI: 66%–82%) and 78% (95%CI: 59%–89%), respectively (Table 2). Using majority voting strategy in per-patient analysis, the 3D-CNN model achieved a sensitivity of 100% (95%CI: 80%–100%) with a specificity of 60% (95%CI: 23%–93%) with an accuracy of 90% and a positive and negative value of 88% (95%CI: 66%–98%) and 100% (95%CI: 44%–100%), respectively (Table 2). Using various cutoff values in per-patient analysis, an ROC curve was constructed with an AUC of 0.786 (95%CI: 0.510–1.000) (Figure 2A) and the best cutoff value was indeed the same as the major voting strategy (i.e., >50% of scans predicted to be positive).

Postoperative positive-margin predictive model

The training group consisted of 83 patients (total of 629 scans) with 36 patients having a positive margin. The validation

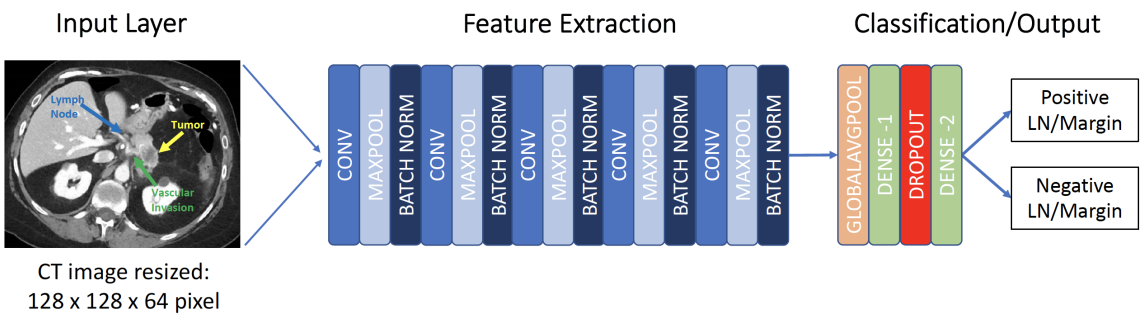


FIGURE 1
Framework for 3D convolutional neural network. CONV, convolutional layer; MAXPOOL, max pooling; LN, Lymph node.

TABLE 2 Confusion matrix for lymph node study.

Type of analysis	True positive	True negative
Per-patient analysis (n=20)		
Predicted Positive	15	2
Predicted Negative	0	3
Per-scan analysis (n=140)		
Predicted Positive	84	29
Predicted Negative	6	21

group for the margin model consisted of 27 patients (total of 252 scans), 11 of whom had a positive margin. In per-scan analysis, the 3D-CNN model achieved a sensitivity of 67% (95%CI: 59%–74%) and a specificity of 89% (95%CI: 81%–93%) with an accuracy of 76% and a positive and negative predictive value of 89% (95%CI: 82%–94%) and 65% (95%CI: 57%–73%), respectively (Table 3). Using majority voting strategy in per-patient analysis, the 3D-CNN model achieved a sensitivity of 73% (95%CI: 43%–90%) and a specificity of 88% (95%CI: 64%–98%) with accuracy of 81% and a positive and negative predictive value of 80% (95%CI: 49%–96%) and 82% (95%CI: 59%–94%), respectively (Table 3). Using various cutoff values in per-patient analysis, an ROC curve was constructed with an AUC of 0.852 (95%CI: 0.670–1.000) (Figure 2B) and the best cutoff values were between 40% and 60%.

Discussion

The purpose of this study is to provide a proof of concept that 3D CNN-based algorithms can predict lymph node metastasis and the postoperative margin status with clinically relevant levels of accuracy. The CT scans of 110 patients from a single tertiary care institution were utilized without segmentation. The lymph node prediction model achieved an accuracy of 75% in per-scan analysis and 90% in per-patient analysis using majority voting, while the postoperative margin prediction model achieved an accuracy of 76% in per-scan analysis and 81% in per-patient analysis using majority voting. This is the first study to utilize a 3D-CNN for the prediction of postoperative margins and the first study to utilize a 3D-CNN to predict the lymph node status in pancreatic cancer.

The most promising type of machine learning model for radiomic analysis has been the CNN (16). CNNs were developed in the late 1970s and saw their first application into medical imaging analysis in the 1990s (27). CNNs became more widely recognized after the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, whereby algorithms were tasked with classifying over 1.2 million high-resolution images from 22,000 categories into 1,000 classes (28). AlexNet, the winning model, was highly efficient and accurate and provided a framework for the future iterations of CNNs (29). CNNs are a more popular option in comparison to other types of machine learning algorithms, such as the random forest model or decision trees for radiomic data. They are superior in modeling non-linear relationships in seemingly unrelated data to achieve a result (30). In contrast to random

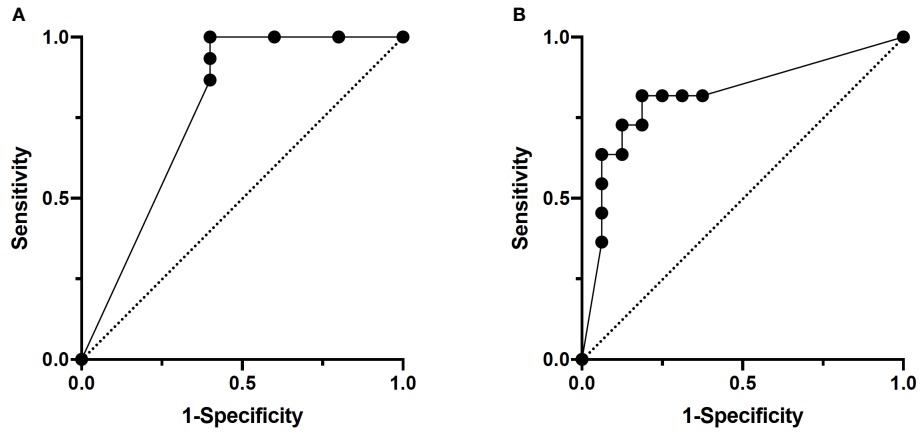


FIGURE 2
ROC curves of per-patient analysis. **(A)** Lymph node study. AUC: 0.79. **(B)** Margin study. AUC: 0.85.

TABLE 3 Confusion matrix for margin study.

Type of analysis	True positive	True negative
Per-patient analysis (n=27)		
Predicted Positive	8	2
Predicted Negative	3	14
Per-scan analysis (n=252)		
Predicted Positive	98	12
Predicted Negative	49	93

forest models, CNNs lack the interpretability of individual features and focus on solving a specific problem (16, 31). CNNs have been applied to a wide range of medical problems with over 300 papers published in the last few years (17). All kinds of medical imaging including X-ray, CT, MRI, and ultrasound have been utilized with CNNs (17). For example, in pancreas imaging, studies have been performed looking to use 3D-CNNs for the diagnosis of pancreatic cystic neoplasms, neuroendocrine tumors, and additional segmentation of the pancreas. Recently in 2020, a 3D-CNN model was described for the classification of pancreatic cancer from initial diagnostic CT scans that demonstrated a sensitivity of 99% and an accuracy of 99%. Another study used CNNs to measure pancreas volumes in patients with type 1 diabetes (32). The pancreas is an inherently more difficult organ to evaluate than the liver or kidney due to its variable shape, size, and proximity to numerous structures. Studies utilizing pre-analysis segmentation to isolate the pancreas from neighboring structures have yielded improved accuracy in comparison to non-segmentation studies (33).

Lymph node metastasis is a significant prognostic factor in pancreatic cancer survival; however, preoperative lymph node identification remains a challenge in the diagnostic radiology of pancreatic and other abdominal cancers with sensitivities ranging from 40% to 87% and specificities ranging from 64% to 100% for CT and MRI (26, 34). Radiologists are limited to looking at the size, shape, and contour of lymph nodes on CT scans. Specifically, CT and MRI techniques are limited in the ability to detect metastatic disease in normal-sized or minimally enlarged lymph nodes. Based on tumor morphology, the incidence of metastatic disease within normal-sized nodes may occur anywhere from 10% to 90% of cases (35). In pancreatic cancer, the size of ≥1cm was only 44.2% sensitive to the identification of lymph node metastasis (34). While there have been attempts to utilize CNNs in CT segmentation to identify metastatic lymph nodes, the first use of CNNs to evaluate for potentially metastatic lymph nodes was performed in head and neck cancers by Chen et al. (36). Utilizing many-objective radiomics in conjunction with the 3D-CNN framework, the group created a model to predict three classes of lymph nodes: normal, suspicious, or involved. Segmentation was used to identify specific nodes for analysis. The accuracy of the model was 0.88. Additional machine learning models have been created for the identification of lymph node metastasis in cervical cancer (37) and the prediction of lymph node metastasis in gastric cancer (38) and prostate cancer (39). This is the first study investigating the pancreas. The 3D-CNN proposed by this paper offers a different approach as this model does not utilize segmentation and imaging studies were at a different anatomical location likely involving different radiomic parameters. An acceptable accuracy of 90% was achieved in per-patient analysis.

The accuracy of CT imaging for predicting resectability is approximately 70% and is prone to overestimation (40). The ability to improve preoperative patient selection for such a substantial surgical procedure could be vital in improving overall clinical outcomes. Patients who are deemed to be high risk for R1 resection even though their tumors are classified as technically resectable based on current clinical and radiological guidelines may warrant a consideration for neoadjuvant therapy. It remains controversial whether the postoperative positive microscopic margin (R1 resection) has an impact on survival postoperatively since the probability of the recurrence-free survival and overall survival of these patients depends on multiple factors including underlying medical conditions, the postoperative course, the choice of systemic treatment, the treatment response, pathological and molecular subtypes, and the stage of disease. The rates of R1 resection in the literature may range widely from as low as 16% to >75% with some studies also noting an association with poorer clinical outcomes in comparison to R0 resection but not others (41). This discrepancy was due to a lack of standardization with the pathologic evaluation of resection specimens and definitions (9). The Royal College of Pathologists define R1 resection as a "microscopic evidence of tumor within 1 mm of a resection margin" (42, 43). The adoption of the standardized definitions of R1 resection as well as the circumferential resection margin has led to increase in the literature-reported incidence rates of R1 resection (44). Recent meta-analysis data have shown that R1 resection is associated with decreased overall survival and disease-free survival in PDAC patients after pancreaticoduodenectomy (Whipple procedure) (9, 12). One thinking is that R1 resection following curative intent surgery may indicate the presence of micrometastatic disease unable to be identified preoperatively. A

developing paradigm shift in the management of PDAC is the usage of neoadjuvant therapy in the cases of resectable or borderline resectable cancer (45). The Preoperative radiochemotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer (PREOPANC-1) randomized phase III trial, comparing neoadjuvant with gemcitabine and chemoradiation vs. adjuvant gemcitabine in resectable or borderline resectable tumors did not identify any difference in overall survival; however, there was improvement in the secondary endpoints of disease-free survival and the R0 resection rate (45). In subgroup analysis, borderline resectable but not resectable tumors demonstrated an improvement in overall survival (46). In this study population, survival analysis supports the notion that a positive resection margin is associated with worse overall survival and recurrence-free survival, as well as worse local and distant recurrence-free survival in Kaplan–Meier and univariate Cox hazard ratio analyses (Supplemental Figure 1 and Supplemental Tables 1, 2). The lymph node status was only associated with overall survival (Supplemental Figure 2 and Supplemental Tables 1, 2). In multivariate Cox hazard ratio analysis, a positive margin remains associated with recurrence-free survival (Supplemental Table 2). This suggests that the margin status may act as a surrogate marker of recurrence. It is important to note that in this study population, only 16 out of 110 patients received neoadjuvant therapy, with the majority of borderline resectable tumors receiving upfront surgery. The purpose of performing preoperative margin prediction is to potentially assist in clinical decision-making for these types of tumors, where patients predicted to have positive margin should probably consider neoadjuvant treatment instead of upfront surgery.

Margin studies are difficult to accomplish specifically in the pancreas due to the need for the CNN to understand and evaluate proximity to a "weighted" group of vital structures. There have been no machine-learning models trained to identify the postoperative margin status from preoperative images. A study performed by Halicek et al. described the use of CNNs in patients with squamous cell carcinoma in their oral cavity to identify residual disease on postresection imaging studies (47). The model proposed in this paper utilized a simplistic approach to provide a proof of concept with subsequent fine-tuning available in future iterations. Without the segmentation of images, the model learns the pancreas, auto-segments the tumor from the normal pancreas, and attempts to classify the characteristics of surrounding pixels to trained binary outcomes. A future iteration of the model should look to identify specific radiomic parameters investigated in order to compare the radiomic differences between high- and low-risk tumors for a postoperative positive margin.

A major limitation of this study is the small sample size for respective training and validation groups. In our algorithm, this was attempted to be mediated by additional per-scan analysis to increase sample size as well as well as limiting the number of trainable parameters, which demonstrated worse accuracy in comparison to majority voting in per-patient analysis. The concern when utilizing smaller datasets is whether the model is specifically learning features that would ideally distinguish from the testing criteria or just overfitting for some features in the given dataset. Additionally, the design of this study is such that the outcome is a binary yes or no to the question posed. There is no distinction to which the lymph node station or margin is the one predicted to be positive nor if the features of the true-positive lymph node or margin are sampled. A consideration for future modification to this 3D-CNN model would be to use postoperative lymph node pathology with preoperative image segmentation for individual lymph node stations and tumor boundaries in the training group. Additionally, future CNN models on larger datasets should seek to perform iterations with the optimization of overall survival and recurrence-free survival with propensity- matched cases to alleviate confounding characteristics. Lastly, a small dataset of 110 patients, majority (98%) Caucasian, could mean that training CT images may not be representative of the generalizable population of PDAC tumors (48). Additional diversity should be included in additional training groups for CNNs.

Medical outcome modeling for treatment planning is a novel application of convolutional neural networks that warrants additional investigation.

Conclusion

In conclusion, this study provides a proof of concept that utilizing radiomics, the 3D-CNN deep learning framework may be used to improve the preoperative prediction of positive resection margins as well as the presence of lymph node metastatic disease. Further investigations should be performed with larger cohorts to increase the generalizability of this model; however, there is great promise in the use of CNNs to assist clinicians with treatment selection for patients with PDAC.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board, University of Iowa. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JC performed clinical data abstraction, defined the boundaries of all CT scans, performed data analysis, and wrote the manuscript. YL prepared the image files, built the 3D-CNN models, and performed data analysis. SS and KC performed clinical data abstraction. HS and KS performed the abstraction of clinical and imaging data. MR offered radiology insights and provided supervision/guidance on defining boundaries on CT scans. MS offered insight and guidance on imaging analysis. CC conceptualized the overall study, supervised the study, provided clinical and surgical insights, obtained research funding, performed data analysis, interpreted experimental data, and edited the manuscript. All authors contributed to the manuscript and approved for the manuscript submission.

Funding

This study was supported by the Carver College of Medicine/Iowa Initiative for Artificial Intelligence Pilot grant, a subaward from the Carnegie-Mellon University through funds from the National Institute of Health under award number OT2OD026675, and the Holden Comprehensive Cancer Center through funds from the National Cancer Institute of the National Institutes of Health under award number P30CA086862 for supporting the Biospecimen Procurement and Molecular Epidemiology Resource Core.

Acknowledgments

We would like to thank all the current and past members of the BioMER team for consenting and enrolling all the pancreatic cancer patients, maintaining the GIMER prospective patient database, and taking care of all regulatory aspects of the project. We would also like to thank the staff members in the Department of Radiology who facilitated the procurement of deidentified CT scan images.

Conflict of interest

CC received research support from Checkmate Pharmaceuticals, Angiodynamics, and Optimum Therapeutics for clinical trials and research projects unrelated to this study.

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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.895515/full#supplementary-material

References

1. McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG, McCain RS. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World J Gastroenterol* (2018) 24(43):4846–61. doi: 10.3748/wjg.v24.i43.4846

2. Zhang L, Sanagapalli S, Stoita A. Challenges in diagnosis of pancreatic cancer. *World J Gastroenterol* (2018) 24(19):2047–60. doi: 10.3748/wjg.v24.i19.2047

3. Wei K, Hackert T. Surgical treatment of pancreatic ductal adenocarcinoma. *Cancers (Basel)* (2021) 13(8):1971. doi: 10.3390/cancers13081971

4. Morales-Oyarvide V, Rubinson DA, Dunne RF, Kozak MM, Bui JL, Yuan C, et al. Lymph node metastases in resected pancreatic ductal adenocarcinoma: Predictors of disease recurrence and survival. *Br J Cancer* (2017) 117(12):1874–82. doi: 10.1038/bjc.2017.349

5. Slidell MB, Chang DC, Cameron JL, Wolfgang C, Herman JM, Schulick RD, et al. Impact of total lymph node count and lymph node ratio on staging and survival after pancreatectomy for pancreatic adenocarcinoma: A large, population-based analysis. *Ann Surg Oncol* (2008) 15(1):165–74. doi: 10.1245/s10434-007-9587-1

6. You MS, Lee SH, Choi YH, Shin BS, Paik WH, Ryu JK, et al. Lymph node ratio as valuable predictor in pancreatic cancer treated with R0 resection and adjuvant treatment. *BMC Cancer* (2019) 19(1):952. doi: 10.1186/s12885-019-6193-0

7. Khalvati F, Zhang Y, Baig S, Lobo-Mueller EM, Karanicolas P, Gallinger S, et al. Prognostic value of CT radiomic features in resectable pancreatic ductal adenocarcinoma. *Sci Rep* (2019) 9(1):5449. doi: 10.1038/s41598-019-41728-7

8. Garces-Descovich A, Beker K, Jaramillo-Cardoso A, James Moser A, Mortele KJ. Applicability of current NCCN guidelines for pancreatic adenocarcinoma resectability: Analysis and pitfalls. *Abdom Radiol (NY)* (2018) 43(2):314–22. doi: 10.1007/s00261-018-1459-6

9. Rau BM, Moritz K, Schuschan S, Alsfasser G, Prall F, Klar E, et al. R1 resection in pancreatic cancer has significant impact on long-term outcome in standardized pathology modified for routine use. *Surgery* (2012) 152(3 Suppl 1): S103–11. doi: 10.1016/j.surg.2012.05.015

10. Teske C, Stimpel R, Distler M, Merkel S, Grützmann R, Bolm L, et al. Impact of resection margin status on survival in advanced n stage pancreatic cancer - a multi-institutional analysis. *Langenbecks Arch Surg* (2021) 406(5):1481–9. doi: 10.1007/s00423-021-02138-4

11. Strobel O, Hank T, Hinz U, Bergmann F, Schneider L, Springfeld C, et al. Pancreatic cancer surgery: The new r-status counts. *Ann Surg* (2017) 265(3):565–73. doi: 10.1097/SLA.0000000000001731

12. Demir IE, Jäger C, Schlitter AM, Konukiewitz B, Stecher L, Schorn S, et al. R0 versus R1 resection matters after pancreaticoduodenectomy, and less after distal or total pancreatectomy for pancreatic cancer. *Ann Surg* (2018) 268(6):1058–68. doi: 10.1097/SLA.0000000000002345

13. Hong SB, Lee SS, Kim JH, Kim HJ, Byun JH, Hong SM, et al. Pancreatic cancer CT: Prediction of resectability according to NCCN criteria. *Radiology* (2018) 289(3):710–8. doi: 10.1148/radiol.2018180628

14. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (2017) 14(12):749–62. doi: 10.1038/nrclinonc.2017.141

15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology* (2016) 278(2):563–77. doi: 10.1148/radiol.2015151169

16. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D deep learning on medical images: A review. *Sensors (Basel)* (2020) 20(18):5097. doi: 10.3390/s20185097

17. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

18. Lin M, Momin S, Lei Y, Wang H, Curran WJ, Liu T, et al. Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-net. *Med Phys* (2021) 48(8):4365–74. doi: 10.1002/mp.15032

19. Wegmayr V AS, Buhmann J. Classification of brain MRI with big data and deep 3D convolutional neural networks. *Proc SPIE* (2018) 10575:105751S. doi: 10.1117/12.2293719

20. Lu S, Liu S, Wang SH, Zhang YD. Cerebral microbleed detection *via* convolutional neural network and extreme learning machine. *Front Comput Neurosci* (2021) 15:738885. doi: 10.3389/fncom.2021.738885

21. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X, et al. Deep learning in medical image registration: A review. *Phys Med Biol* (2020) 65(20):20TR01. doi: 10.1088/1361-6560/ab843e

22. Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *J Clin Oncol* (2016) 34(18):2157–64. doi: 10.1200/JCO.2015.65.9128

23. Zhou Z, Chen L, Sher D, Zhang Q, Shah J, Pham NL, et al. Predicting lymph node metastasis in head and neck cancer by combining many-objective radiomics and 3-dimensioal convolutional neural network through evidential reasoning. *Annu Int Conf IEEE Eng Med Biol Soc 2018.* (2018) p:1–4. doi: 10.1109/EMBC.2018.8513070

24. Zunair H RA, Mohammed N, Cohen JP. (2020). Uniformizing techniques to process CT scans with 3D CNNs for tuberculosis prediction, in: Rekik I, Adeli E, Park SH, Valdés Hernández MdC . (eds). *Predictive Intelligence in Medicine. PRIME 2020. Lecture Notes in Computer Science* (Springer, Cham),12329. doi: 10.1007/978-3-030-59354-4_15

25. Wu L, Holbrook C, Zaborina O, Ploplys E, Rocha F, Pelham D, et al. Pseudomonas aeruginosa expresses a lethal virulence determinant, the PA-I lectin/adhesin, in the intestinal tract of a stressed host: The role of epithelia cell contact and molecules of the quorum sensing signaling system. *Ann Surg* (2003) 238 (5):754–64. doi: 10.1097/01.sla.0000094551.88143.f8

26. Ozaki H, Hiraoka T, Mizumoto R, Matsuno S, Matsumoto Y, Nakayama T, et al. The prognostic significance of lymph node metastasis and intrapancreatic perineural invasion in pancreatic cancer after curative resection. *Surg Today* (1999) 29(1):16–22. doi: 10.1007/BF02482964

27. Lo SB LS, Lin JS, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans Med Imaging* (1995) 14(4):711–8. doi: 10.1109/42.476112

28. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev BioMed Eng* (2017) 19:221–48. doi: 10.1146/annurev-bioeng-071516-044442

29. Krizhevsky A SI, Hinton G. (2012). Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Proceedings of the Advances in Neural Information Processing Systems, 2012* (Curran Associates, Red Hook, NY), 29:1097–105.

30. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* (2018) 172(5):1122–1131.e9. doi: 10.1016/j.cell.2018.02.010

31. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J Big Data* (2021) 8(1):53. doi: 10.1186/s40537-021-00444-8

32. Roger R, Hilmes MA, Williams JM, Moore DJ, Powers AC, Craddock RC, et al. Deep learning-based pancreas volume assessment in individuals with type 1 diabetes. *BMC Med Imaging* (2022) 22(1):5. doi: 10.1186/s12880-021-00729-7

33. Liu KL, Wu T, Chen PT, Tsai YM, Roth H, Wu MS, et al. Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: A retrospective study with cross-racial external validation. *Lancet Digit Health* (2020) 2(6):e303–13. doi: 10.1016/S2589-7500(20)30078-9

34. Loch FN, Asbach P, Haas M, Seeliger H, Beyer K, Schineis C, et al. Accuracy of various criteria for lymph node staging in ductal adenocarcinoma of the pancreatic head by computed tomography and magnetic resonance imaging. *World J Surg Oncol* (2020) 18(1):213. doi: 10.1186/s12957-020-01951-3

35. Ganeshalingam S, Koh DM. Nodal staging. *Cancer Imaging* (2009) 9:104–11.

36. Chen L, Zhou Z, Sher D, Zhang Q, Shah J, Pham NL, et al. Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Phys Med Biol* (2019) 64(7):075011. doi: 10.1088/1361-6560/ab083a

37. Wu Q, Wang S, Zhang S, Wang M, Ding Y, Fang J, et al. Development of a deep learning model to identify lymph node metastasis on magnetic resonance imaging in patients with cervical cancer. *JAMA Netw Open* (2020) 3(7):e2011625. doi: 10.1001/jamanetworkopen.2020.11625

38. Zhou CM, Wang Y, Ye HT, Yan S, Ji M, Liu P, et al. Machine learning predicts lymph node metastasis of poorly differentiated-type intramucosal gastric cancer. *Sci Rep* (2021) 11(1):1300. doi: 10.1038/s41598-020-80582-w

39. Wessels F, Schmitt M, Krieghoff-Henning E, Jutzi T, Worst TS, Waldbillig F, et al. Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer. *BJU Int* (2021) 128(3):352–60. doi: 10.1111/bju.15386

40. Bluemke DA, Cameron J, Hruban R, Pitt H, Siegelman S, Soyer P, et al. Potentially resectable pancreatic adenocarcinoma: Spiral CT assessment with surgical and pathologic correlation. *Radiology* (1995) 197(2):381–5. doi: 10.1148/radiology.197.2.7480681

41. Verbeke CS. Resection margins and R1 rates in pancreatic cancer–are we there yet? *Histopathology* (2008) 52(7):787–96. doi: 10.1111/j.1365-2559.2007.02935.x

42. Campbell F, Smith RA, Whelan P, Sutton R, Raraty M, Neoptolemos JP, et al. Classification of R1 resections for pancreatic cancer: The prognostic relevance of tumour involvement within 1 mm of a resection margin. *Histopathology* (2009) 55(3):277–83. doi: 10.1111/j.1365-2559.2009.03376.x

43. Verbeke CS, Leitch D, Menon KV, McMahon MJ, Guillou PJ, Anthoney A, et al. Redefining the R1 resection in pancreatic cancer. *Br J Surg* (2006) 93 (10):1232–7. doi: 10.1002/bjs.5397

44. Esposito I, Kleeff J, Bergmann F, Reiser C, Herpel E, Friess H, et al. Most pancreatic cancer resections are R1 resections. *Ann Surg Oncol* (2008) 15(6):1651–60. doi: 10.1245/s10434-008-9839-8

45. Chawla A, Ferrone CR. Neoadjuvant therapy for resectable pancreatic cancer: An evolving paradigm shift. *Front Oncol* (2019) 9:1085. doi: 10.3389/fonc.2019.01085

46. Versteinjne E SM, Versteijne E, Suker M, Groothuis K, Akkermans-Vogelaar JM, Besselink MG, Bonsing BA, et al, et al. Preoperative chemoradiotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer: Results of the Dutch randomized phase III PREOPANC trial. *J Clin Oncol* (2020) 38(16):1763–73. doi: 10.1200/JCO.19.02274

47. Halicek M, Little JV, Wang X, Patel M, Griffith CC, Chen AY, et al. Tumor margin classification of head and neck cancer using hyperspectral imaging and convolutional neural networks. *Proc SPIE Int Soc Opt Eng* (2018) 10576:1057605. doi: 10.1117/12.2293167

48. Tavakkoli A, Singal AG, Waljee AK, Elmunzer BJ, Pruitt SL, McKey T, et al. Racial disparities and trends in pancreatic cancer incidence and mortality in the united states. *Clin Gastroenterol Hepatol* (2020) 18(1):171–178.e10. doi: 10.1016/j.cgh.2019.05.059

Frontiers in Oncology

Advances knowledge of carcinogenesis and tumor progression for better treatment and management

The third most-cited oncology journal, which highlights research in carcinogenesis and tumor progression, bridging the gap between basic research and applications to imrpove diagnosis, therapeutics and management strategies.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact