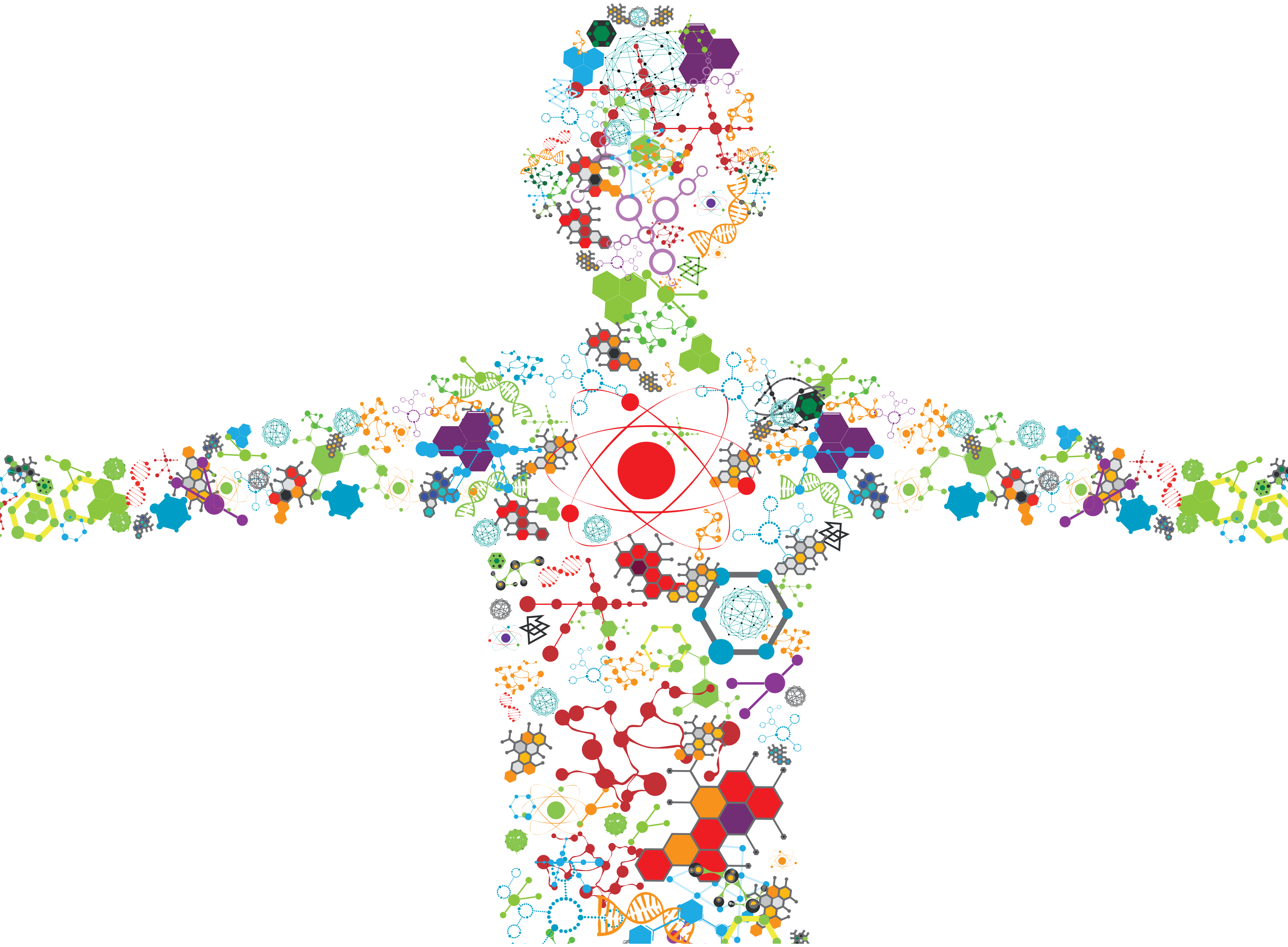
MACHINE LEARNING USED IN BIOMEDICAL COMPUTING AND INTELLIGENCE HEALTHCARE, VOLUME I

EDITED BY: Honghao Gao, Ying Li, Zijian Zhang and Wenbing Zhao

PUBLISHED IN: Frontiers in Bioengineering and Biotechnology, Frontiers in Genetics and Frontiers in Public Health



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-932-5
DOI 10.3389/978-2-88966-932-5

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

MACHINE LEARNING USED IN BIOMEDICAL COMPUTING AND INTELLIGENCE HEALTHCARE, VOLUME I

Topic Editors:
Honghao Gao, Shanghai University, China
Ying Li, Zhejiang University, China
Zijian Zhang, The University of Auckland, New Zealand
Wenbing Zhao, Cleveland State University, United States

Citation: Gao, H., Li, Y., Zhang, Z., Zhao, W., eds. (2021). Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume I.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-932-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 03 May 2021
doi: 10.3389/fgene.2021.678140

Editorial: Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume I

Honghao Gao[1]*, *Ying Li*[2], *Zijian Zhang*[3] and *Wenbing Zhao*[4]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China, [2] School of Computer, Zhejiang University, Hangzhou, China, [3] The School of Computer Science, The University of Auckland, Auckland, New Zealand, [4] College of Engineering, Cleveland State University, Cleveland, OH, United States

Keywords: machine learning, biomedical computing, intelligence, healthcare, precision medicine

Editorial on the Research Topic

Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume I

In recent years, the development of biomedical imaging techniques, integrative sensors, and artificial intelligence has brought many benefits to the protection of health. We can collect, measure, and analyze vast volumes of health-related data using the technologies of computing and networking, leading to tremendous opportunities for the health and biomedical community. Biomedical intelligence, especially precision medicine, is considered one of the most promising directions for healthcare development. Meanwhile, these technologies have also brought new challenges and issues.

This Research Topic was supported by Frontiers and includes three collaborating journals: Frontiers in Genetics, Frontiers in Public Health, and Frontiers in Computer Science. We accepted 10 papers from 21 open submissions. The summaries of these papers are outlined below.

In the article entitled "Development, Validation and Comparison of Artificial Neural Network Models and Logistic Regression Models Predicting Survival of Unresectable Pancreatic Cancer" by Tong et al. the authors developed Artificial Neural Network (ANN) models based on 3, 7, and 32 basic features, predicting the survival of unresectable pancreatic cancer patients over 8 months. These models might help to optimize personalized patient management.

In the article entitled "*P*-Wave Area Predicts New Onset Atrial Fibrillation in Mitral Stenosis: A Machine Learning Approach" by Tse et al. the authors studied Chinese patients diagnosed with mitral stenosis in sinus rhythm at baseline between November 2009 and October 2016. They concluded that atrial electrophysiological alterations in mitral stenosis could be detected using electrocardiograms and based on age, and systolic blood pressure and the *P*-wave area in V3 could predict new onset atrial fibrillation (AF). They proposed a decision tree learning model, which significantly improves outcome prediction.

In the article entitled "Detection and Severity Assessment of Peripheral Occlusive Artery Disease via Deep Learning Analysis of Arterial Pulse Waveforms: Proof-of-Concept and Potential Challenges" by Kim et al. the authors demonstrated the deep learning-based arterial pulse waveform analysis contributes to the PAD screening, and they presented challenges that must be addressed for real-world clinical applications.

In the article entitled "Develop and Evaluate a New and Effective Approach for Predicting Dyslipidemia in Steel Workers" by Wu et al. the authors collected the physical examination information of thousands of steelworkers and screened out the risk factors of dyslipidemia in steelworkers. Then, based on the data characteristics, they employed the convolutional neural network to predict the risk of dyslipidemia in steelworkers.

OPEN ACCESS

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

****Correspondence:***
Honghao Gao
gaohonghao@shu.edu.cn

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

Received: *09 March 2021*
Accepted: *29 March 2021*
Published: *03 May 2021*

Citation:
Gao HH, Li Y, Zhang ZJ and Zhao WB (2021) Editorial: Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume I. Front. Genet. 12:678140. doi: 10.3389/fgene.2021.678140

In the article entitled "Automated Detection of Acute Lymphoblastic Leukemia From Microscopic Images Based on Human Visual Perception" by Bodzas et al. the authors proposed a novel approach based on conventional digital image processing techniques and machine learning algorithms. The traditional machine learning classifiers, the artificial neural network and the support vector machine, were used to automatically identify acute lymphoblastic leukemia from peripheral blood smear images.

In the article entitled "Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA" by Yang et al. the authors introduced the development process of sequencing technology. They analyzed the basic process of data mining, summary several major machine learning algorithms, and pointed out the challenges faced by machine learning algorithms in the mining of biological sequence data and possible solutions. They also reviewed four typical applications of machine learning in Deoxyribonucleic acid (DNA) sequence data.

In the article entitled "Parkinson's Disease in Teneurin Transmembrane Protein 4 (TENM4) Mutation Carriers" by Pu et al. the authors investigated clinical and genetic manifestations in four unrelated pedigrees with both essential tremor (ET) and Parkinson's disease (PD) in which TENM4 variants were identified. They discussed whether TENM4 variants contributed to the risk of developing PD. Thus, the frequency of TENM4 variants was evaluated from four PD pedigrees and other 407 subjects.

In the article entitled "Deep Learning in Head and Neck Tumor Multiomics Diagnosis and Analysis: Review of the Literature" by Wang and Li the authors reviewed the multiomics image analysis of head and neck tumors using convolution neural network (CNN) and other Deep learning (DL) neural networks. They evaluated its application in early tumor detection, classification, prognosis/metastasis prediction, and the signing out of the reports.

In the article entitled "Alzheimer's Disease Classification with a Cascade Neural Network" by You et al. the authors proposed a cascade neural network with two steps to achieve a faster and more accurate Alzheimer's Disease (AD) classification by exploiting gait and electroencephalogram (EEG) data simultaneously. They collected gait and EEG data from 35 cognitively healthy controls, 35 mild cognitive impairment (MCI), and 17 AD patients to demonstrate their proposed method.

In the article entitled "Application of Structural and Functional Connectome Mismatch for Classification and Individualized Therapy in Alzheimer Disease" by Ren et al. the authors performed a preliminary exploration into a set of Alzheimer disease data to improve the personalized approach in order to understand individual connectomes in an actionable manner. They found that there were consistent patterns of white matter fiber loss, mainly the Default Mode Network (DMN) and Deep Subcortical Structures (DSS), which were present in nearly all patients with clinical Alzheimer's disease.

In conclusion, we would like to thank all the authors who submitted their research articles to our Research Topic. We highly appreciate the contributions of the reviewers for their constructive comments and suggestions. We also would like to acknowledge the guidance from the Editor-in-Chief and staff members of Frontiers.

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.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 13 March 2020
doi: 10.3389/fbioe.2020.00196

Development, Validation and Comparison of Artificial Neural Network Models and Logistic Regression Models Predicting Survival of Unresectable Pancreatic Cancer

Zhou Tong[1†], *Yu Liu*[1†], *Hongtao Ma*[2], *Jindi Zhang*[2], *Bo Lin*[2], *Xuanwen Bao*[3], *Xiaoting Xu*[4], *Changhao Gu*[5], *Yi Zheng*[1], *Lulu Liu*[1], *Weijia Fang*[1,6], *Shuiguang Deng*[2] *and Peng Zhao*[1*]

[1] Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China, [3] Technical University Munich (TUM), Munich, Germany, [4] Department of Medical Oncology, Tai He People's Hospital, Fuyang, China, [5] Internal Medicine, Cangnan Traditional Chinese Medicine Hospital, Wenzhou, China, [6] Zhejiang Provincial Key Laboratory of Pancreatic Disease, Hangzhou, China

Background: Prediction models for the overall survival of pancreatic cancer remain unsatisfactory. We aimed to explore artificial neural networks (ANNs) modeling to predict the survival of unresectable pancreatic cancer patients.

Methods: Thirty-two clinical parameters were collected from 221 unresectable pancreatic cancer patients, and their prognostic ability was evaluated using univariate and multivariate logistic regression. ANN and logistic regression (LR) models were developed on a training group (168 patients), and the area under the ROC curve (AUC) was used for comparison of the ANN and LR models. The models were further tested on the testing group (53 patients), and k-statistics were used for accuracy comparison.

Results: We built three ANN models, based on 3, 7, and 32 basic features, to predict 8 month survival. All 3 ANN models showed better performance, with AUCs significantly higher than those from the respective LR models (0.811 vs. 0.680, 0.844 vs. 0.722, 0.921 vs. 0.849, all $p < 0.05$). The ability of the ANN models to discriminate 8 month survival with higher accuracy than the respective LR models was further confirmed in 53 consecutive patients.

Conclusion: We developed ANN models predicting the 8 month survival of unresectable pancreatic cancer patients. These models may help to optimize personalized patient management.

Keywords: artificial neural network, logistic regression, unresectable pancreatic cancer, survival, prognosis

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Dan Cao,
West China Hospital, Sichuan University, China
Kun Wang,
Peking University Cancer Hospital, China
Xiaoxian Yang,
Shanghai Second Polytechnic University, China

***Correspondence:**
Peng Zhao
zhaop@zju.edu.cn

†*These authors have contributed equally to this work*

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

Received: *29 January 2020*
Accepted: *27 February 2020*
Published: *13 March 2020*

Citation:
Tong Z, Liu Y, Ma H, Zhang J, Lin B, Bao X, Xu X, Gu C, Zheng Y, Liu L, Fang W, Deng S and Zhao P (2020) Development, Validation and Comparison of Artificial Neural Network Models and Logistic Regression Models Predicting Survival of Unresectable Pancreatic Cancer. Front. Bioeng. Biotechnol. 8:196. doi: 10.3389/fbioe.2020.00196

INTRODUCTION

Pancreatic cancer is one of the leading causes of cancer-related mortality worldwide (Ferlay et al., 2015). Most patients present with few specific symptoms and are diagnosed at an advanced stage. Despite the development of surgical techniques, radiotherapy and chemotherapy, the prognosis of pancreatic cancer is dismal (Hidalgo et al., 2015). In most cases, the disease itself leads to the patients' short survival time, and treatment rarely achieves cure, although some patients achieve remissions lasting several years (Kuhlmann et al., 2004; Cress et al., 2006; Bradley, 2008). Given that life expectancy is relatively short, even in the face of optimal treatment, doctors must weigh the potential survival benefits with the potential impact of treatment complications on patients' quality of life.

Different predictive evaluation systems or risk scores have been developed for decision-making, including perioperative mortality risk (Are et al., 2009), post-surgery complications (Braga et al., 2011) and survival prediction (Miura et al., 2014; Dasari et al., 2016). Survival prediction models help doctors make appropriate recommendations for the most suitable treatment option, thus maximizing the survival benefit. In addition, proper and uniform prediction models can facilitate more accurate enrolment in clinical trials. Nevertheless, current options to predict overall survival remain unsatisfying. The TNM classification developed by the American Joint Committee on Cancer has been used to estimate the prognosis of cancer. However, there are different prognoses in pancreatic cancer patients whose TNM stages are similar (Xu et al., 2017). Previous clinical research has shown the predictive effect of clinical pathological biomarkers such as tumor heterogeneity, main vessel invasion, and complexity at the genomic, epigenetic, and metabolic levels in patients with pancreatic cancer (Kleeff et al., 2016; Neoptolemos et al., 2018; Naito et al., 2019). However, these predictive biomarkers still have many limitations. Additional reliable prognostic indicators are urgently needed.

Artificial neural networks (ANNs), a commonly used method of machine learning, work in a non-linear mode and model a biological neural system both structurally and functionally (Cucchetti et al., 2010). In addition to its application in the field of computer engineering, ANN modeling emerges as a potential useful tool for projecting clinical outcomes (Penny and Frost, 1996). Many clinical studies have compared the predictive power of ANN models with logistic regression (LR) models and have shown ANNs to have better performance (Hanai et al., 2003; Ghoshal and Das, 2008). A systemic review showed an increase in the benefit of ANNs over existing statistics in healthcare provision (Lisboa and Taktak, 2006). However, few studies have compared the performance of ANN with LR in the field of pancreatic cancer.

In our study, we aimed to explore possible prognostic indicators for unresectable pancreatic cancer on the basis of clinical and radiological variables and investigate the diagnostic accuracy of these two methodologies (LR, ANN) in predicting overall survival. The performance of the ANN and logistic regression models were validated externally using a different data set.

MATERIALS AND METHODS

Patients

We retrospectively reviewed 221 cases of unresectable pancreatic cancer registered between May 2010 and December 2018 at the First Affiliated Hospital of Zhejiang University. Taking January 2018 as the dividing point, patients were classified into two groups: 168 patients were used as a training dataset, and 53 patients were used as an independent validation dataset. The inclusion criteria for patients were as follows: (i) patients were histologically confirmed adenocarcinoma of the pancreas; (ii) resectability status were evaluated as unresectable according to the Pancreatic Adenocarcinoma NCCN Guidelines; (iii) patients were $\geq$18 years of age and had a Eastern Cooperative Oncology Group (ECOG) score 0–2; (iv) patients had adequate hematologic, hepatic, and renal function before treatment; (v) Complete clinical imaging data and biochemical data 2 weeks before chemotherapy and survival data were available. The exclusion criteria were: (i) patients received prior chemotherapy or surgery; (ii) recurrent pancreatic cancer. The study followed the international and national regulations in accordance with the Declaration of Helsinki and was approved by the ethics committee of the First Affiliated Hospital, Zhejiang University School of Medicine. The following clinical and biochemical data were collected before the patient received chemotherapy: age, sex, main vascular invasion (celiac axis, superior mesenteric artery, common hepatic artery), clinical TNM staging, metastasis (including retroperitoneal lymph node, liver, lung and peritoneum), ascites, size of the largest tumor in the pancreas and liver, tumor position in the pancreas, stomach invasion, duodenum invasion, liver metastasis number, carcinoembryonic antigen (CEA), carbohydrate antigen 199 (CA199), albumin-to-globulin ratio (AGR), alanine transaminase (ALT), aspartate transaminase (AST), creatinine, total bilirubin, direct bilirubin, indirect bilirubin, haemoglobulin, neutrophil/lymphocyte ratio, platelet/lymphocyte ratio, hepatitis B virus, and white blood cell (WBC) count. Pancreatic tumor or metastatic lesions directly invading stomach was defined as stomach invasion which was diagnosed based on patients' imaging, according to pancreatic ductal adenocarcinoma radiology reporting template (Al-Hawary et al., 2014). Progression-free survival (PFS), overall survival (OS), and chemotherapy regimen were recorded. All patients underwent primary palliative chemotherapy. TNM staging was adopted according to the NCCN Guidelines (version 1. 2019) for pancreatic cancer. The number of tumors, size of the largest tumor (cm), tumor position, and metastasis or invasion organs were defined for all patients on the basis of the CT scan or MRI.

Follow-Up

Patients were followed by outpatient clinics or phone calls until September 2019. These follow-ups were conducted at 3 month intervals. OS was defined as the number of months from the date of diagnosis to the date of death or the date of last follow-up. PFS was defined as the number of months from the date of diagnosis to the date of identification of disease progression. In this study, the median follow-up duration was 9 (range 3–36) months.

Statistical Analysis

All patient characteristics in the training and testing groups were compared. Continuous variables with parametric distributions were evaluated by t-test. Categorical variables were evaluated by χ^2-test (or Fisher's exact test, if appropriate). OS was estimated using the Kaplan–Meier method. The association of the baseline parameters with 8 month survival was assessed using univariate logistic regression analyses, and those with $p < 0.05$ were entered into multivariate logistic regression analyses. Significantly skewed continuous variables (CEA, CA199, ALT, AST, total bilirubin, direct bilirubin, indirect bilirubin, haemoglobulin, the neutrophil/lymphocyte ratio, the platelet/lymphocyte ratio, and WBC count) were normalized by logarithmic transformation. The violin plot was generated using the Python (version 3.7.5) seaborn library.

Development of the Logistic Regression Models

In the training set of 168 patients, variables found to be significantly related to 8 month survival in the multivariate analysis and univariate analysis were entered into logistic regression models 1 and 2, respectively. All 32 variables were entered into logistic regression model 3. A total of 168 patients in the training group were selected to train the logistic regression model, and the remaining 53 patients were used for testing. Logistic regression is a predictive linear model that can be used to predict the causality relationship between a dependent binary variable and one or more independent variables. The formula for logistic regression can be simply presented in linear algebra terms as $Y = A^T X + b$., where Y is the output of our model and X is the input. Both A and b are parameters to be learned from training data. The learned parameter A can be interpreted as the relative importance of each factor in the survival of the patient. Our logistic regression models were built using the Python scikit-learn library.

Development of the Artificial Neural Network Models

In the training group ($n = 168$), 133 (80%) patients were randomly selected to train the network, while 35(20%) for cross validation. Cross-validation was necessary for our neural networks to learn general predictive characteristics rather than memorizing the idiosyncrasies of the training data, which played a role in helping assisting model building, including stopping network training and to avoiding over-fitting.

In the training set of 168 patients, variables found to be significantly related to 8 month survival in the multivariate analysis and univariate analysis were entered into ANN models 1 and 2, respectively. All 32 variables were entered into ANN model 3. A total of 168 patients in the training group were selected to train the network, and the remaining 53 patients were used for testing. Our artificial neural network was built using the PyTorch framework. The search space of network configuration was based empirically on the number of features and the quantity of our available data. And then grid search was conducted to search the best network configuration based on the criteria of our cross-validation group (Bergstra and Bengio, 2012). We have tried three layers or five or more layers, all resulting dissatisfied or overfitting and the best performance was achieved with four layers based on computer experiments. So we built a four-layer feedforward neural network with 3 input nodes in the input layer, 5 and 3 nodes in the first and second hidden layers, respectively, and one output neuron in model 1; 7 input nodes, 8 and 3 neurons in two hidden layers, and one output neuron in model 2; and 32 input nodes, 10 and 8 neurons in two hidden layers, and one output neuron in model 3. **Figure 1** shows the diagrams of ANN models 1–3. The selection strategy was stratified sampling, which guaranteed that the ratios of positive and negative samples in both groups were equal. An early-stop strategy, which stops the training process when the performance of cross-validation no longer improves, was applied in the training of our neural networks.

Assessment of the Diagnostic Accuracy of the Models

The accuracy of the ANN and logistic regression models in predicting 8 month OS were compared using receiver operating characteristic (ROC) curve analysis, positive predictive values (PPV), and positive likelihood ratios (PLR). The performance parameters were calculated by the following formulas: sensitivity: TP/(TP+FN), specificity: TN/(FP+TN), accuracy: (TP+TN)/(P+N), positive predictive value: TP/(TP+FP), negative predictive value: TN/(TN+FN), and positive likelihood ratio = sensitivity/(1-specificity), where TP is true positive, FN is false negative, FP is false positive, TN is true negative, P is positive, and N is negative. The Hanley–McNeil method was used to compare ROC curves. The predictions of both the ANN and logistic regression models in the testing group of 53 patients were reported using Cohen's k coefficient using the formula: [Pr(a)–Pr(e)]/[1–Pr(e)]; Pr(a) is the relative observed agreement, and Pr(e) is the proportion of agreement expected to occur by chance alone (Landis and Koch, 1977). Statistical and ROC analyses were performed by MedCalc 7.2.1.0 (MedCalc software, Mariakerke, Belgium).

RESULTS

Patient Demographics

Of the 211 enrolled patients, 168 were enrolled in the training group, and 53 were enrolled in the testing group. The median overall survival time of the training group was 8 months, which was consistent with previous studies reporting that the median overall survival in advanced pancreatic cancer is approximately 6–11 months (Conroy et al., 2011; Von Hoff et al., 2013). Thus, the 8 month survival was set as the main endpoint of this work. The characteristics of the training and testing groups are listed in **Table 1**. The mean age of the training group was 61.05 ± 8.55 years, and that of the testing group was 61.17 ± 8.42 years ($p > 0.05$). There were 2, 42, 53, and 71 patients with stages T1–T4 disease, respectively, in the training group and 1, 10, 12, and 30 patients with stages T1–T4 disease, respectively, in the testing group ($p > 0.05$). A total of 155 (92.26%) patients were defined as M1 in the training group, and 48 (90.57%) patients

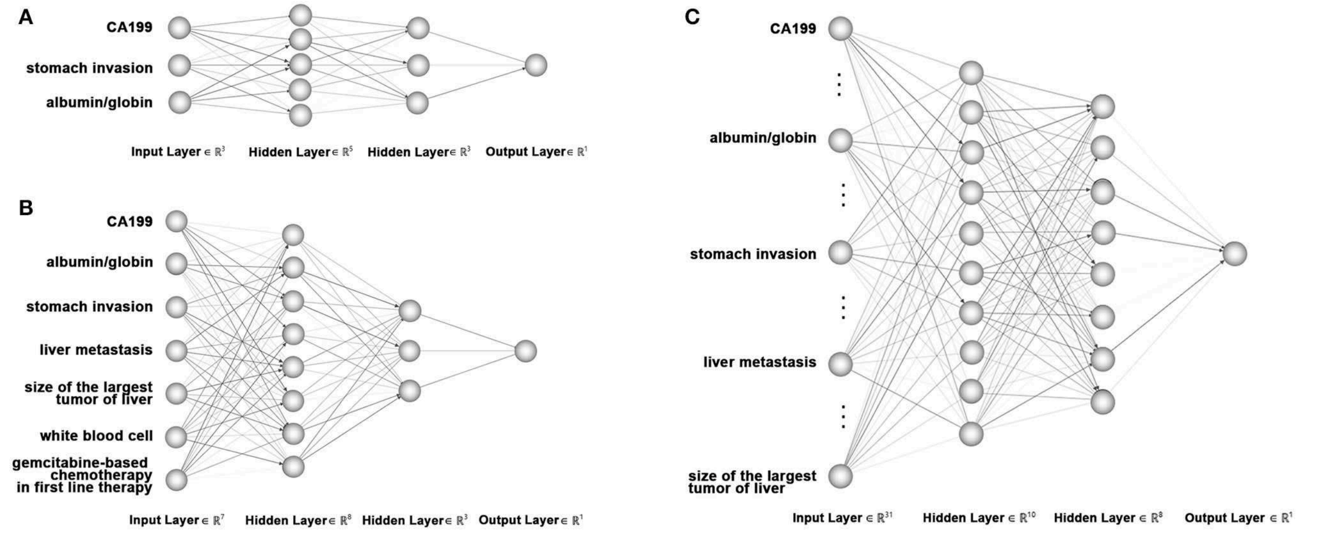


FIGURE 1 | Diagram of artificial neural network models used to predict 8 month survival of unresectable pancreatic cancer. **(A)** Artificial neural network model with 3 input nodes: stomach invasion, AGR and CA199. **(B)** Artificial neural network with 7 input nodes: liver metastasis, stomach invasion, size of the largest tumor of the liver, CA199, AGR, white blood cell count, and gemcitabine-based chemotherapy as the first-line therapy. **(C)** Artificial neural network with 32 input nodes. The output nodes of the three ANN models were 8 month survival.

were defined as M1 in the testing group ($p > 0.05$). There was no statistically significant difference in 8 month survival between these two groups ($p = 0.581$). All patients were treated with at least one dose of chemotherapy. Gemcitabine-based chemotherapy was the most common 1st-line chemotherapy regimen. There were 85.12% and 83.02% of patients who received less than third-line chemotherapy in the training group and testing group, respectively. There were no significant differences in any basic characteristics, including clinical parameters and biological parameters, between the two groups ($p > 0.05$). All continuous variables in the training and testing groups were depicted using violin plots (**Figure 2**).

Prognostic Factors for 8 Month Survival

In the training group of 168 patients, liver metastasis (HR 0.51, $p = 0.041$), stomach invasion (HR 0.408, $p = 0.007$), size of the largest tumor of the liver (HR 0.778, $p = 0.008$), CA199 (HR 0.685, $p = 0.002$), AGR (HR 2.885, $p = 0.002$), WBC (HR 0.092, $p = 0.016$), and gemcitabine-based chemotherapy as the first-line therapy (HR 7.401, $p = 0.009$) were related to 8 month survival in the univariate analysis (**Table 2**). ROC curve analysis was applied to categorize the optimal cutoff value of the AGR for 8 month survival, which was set as 1.48. We classified the patients into groups of 'high AGR ($\geq$1.48)' and 'low AGR ($<$1.48)'. These seven variables were selected as potential independent risk factors in the multivariate analysis. The multivariate logistic regression confirmed stomach invasion (HR 0.473, $p = 0.04$), CA199 (HR 0.754, $p = 0.046$), and AGR (HR 2.360, $p = 0.026$) as independent predictors of 8 month survival (**Table 2**). In the training group of 168 patients, the Kaplan–Meier curve indicated that the OS of patients with abnormal CA199 (median survival, 7.80 vs. 13.73 months, $p < 0.05$), stomach invasion (median survival, 6.83 vs. 9.10 months, $p < 0.05$) and low AGR (median survival, 6.10 vs. 9.10 months, $p < 0.05$) decreased significantly (**Figures 3A–C**).

Artificial Neural Network Models and Logistic Regression Models

Three independent predictors of 8 month survival, stomach invasion, AGR and CA199, were used to build the artificial neural network and logistic regression models labeled ANN model 1 and LR model 1, respectively. The area under the ROC curve (AUC) for ANN model 1 was 0.811 (95% C.I. = 0.743–0.867), higher than that of LR model 1 with 0.680 (95% C.I. = 0.603–0.749, $p < 0.05$) (**Figure 4A**). We applied a cutoff of 0.559 for ANN prediction, and ANN model 1 had a sensitivity of 64.83% and a specificity of 76.62%. ANN model 1 had a higher PPV for 8 month survival prediction than that of LR model 1, reflecting the good predictive power of ANN. The PLR of the ANN model for 8 month survival prediction also remained higher than that of the LR model.

Seven predictors for 8 month survival in the univariate analysis were used to build the ANN and logistic regression models labeled ANN model 2 and LR model 2. The performance of ANN model 2 was high, with an area under the ROC curve (AUC) of 0.844(95% C.I. = 0.780–0.895), compared to that of LR model 2, with an AUC of 0.722 (95% C.I. = 0.648–0.788, $p < 0.05$) (**Figure 4B**). A cutoff of 0.6292 was applied for ANN prediction. ANN model 2 had a sensitivity of 69.23% and a specificity of 87.01%. The PPV and PLR for 8 month survival prediction of ANN model 2 were higher than those from LR model 2.

All 32 clinical and biological parameters were used to build ANN model 3 and LR model 3 to predict 8 month survival. The area under the ROC curve (AUC) of ANN model 3 was 0.921 (95% C.I. = 0.869–0.957), which was higher than that

TABLE 1 | Basic characteristics of the study population.

Variables		**Training (*n* = 168)**	**Testing (*n* = 53)**	***p***
Age, years		61.05 ± 8.55	61.17 ± 8.42	0.928
Gender	Male	106 (63.10%)	38 (71.70%)	0.252
Main vascular invasion		71 (42.26%)	30 (56.60%)	0.068
T	T1	2 (1.19%)	1 (1.89%)	
	T2	42 (25%)	10 (18.87%)	
	T3	53 (31.55%)	12 (22.64%)	
	T4	71 (42.26%)	30 (56.60%)	0.297
N	N0	29 (17.26%)	7 (13.21%)	
	N1	139 (82.74%)	46 (86.79%)	0.486
M	M0	13 (7.74%)	5 (9.43%)	
	M1	155 (92.26%)	48 (90.57%)	0.694
Retroperitoneal lymph node metastasis		95 (56.55%)	31 (58.49%)	0.803
Liver metastasis		106 (63.10%)	36 (67.92%)	0.522
Lung metastasis		19 (11.31%)	5 (9.43%)	0.702
Peritoneal metastasis		21 (12.5%)	7 (13.21%)	0.893
Ascites		21 (12.5%)	4 (7.55%)	0.456
Size of the largest tumor of pancreas, cm		4.61 ± 1.67	4.94 ± 2.02	0.237
Tumor position of pancreas	Head and/or neck	66 (39.29%)	23 (43.40%)	
	Body and/or tail	102 (60.71%)	30 (56.60%)	0.595
Stomach invasion		60 (35.71%)	15 (28.30%)	0.406
Duodenum invasion		22 (13.10%)	9 (16.98%)	0.478
Liver metastasis number	<6	88 (52.38%)	27 (50.94%)	
	≥6	80 (47.62%)	26 (49.06%)	0.855
Size of the largest tumor of liver, cm		1.51 ± 1.85	1.89 ± 2.25	0.217
CEA, ng/mL (log-value)		0.96 ± 0.71	1.15 ± 0.77	0.103
CA199, U/mL (log-value)		2.96 ± 1.35	2.99 ± 1.10	0.865
Albumin/globin		1.63 ± 0.36	1.64 ± 0.37	0.855
ALT, U/L (log-value)		1.33 ± 0.32	1.32 ± 0.33	0.813
AST, U/L (log-value)		1.38 ± 0.24	1.34 ± 0.21	0.283
Creatinine, umol/L		63.81 ± 14.65	67.34 ± 16.48	0.141
Total bilirubin, umol/L (log-value)		1.12 ± 0.26	1.11 ± 0.27	0.849
Direct bilirubin, umol/L (log-value)		0.71 ± 0.34	0.73 ± 0.34	0.599
Indirect bilirubin, umol/L (log-value)		0.88 ± 0.24	0.84 ± 0.27	0.335
Hemoglobin, g/L (log-value)		2.09 ± 0.06	2.11 ± 0.06	0.066
Neutrophil/lymphocyte (log-value)		0.52 ± 0.26	0.57 ± 0.26	0.227
Platelet/lymphocyte (log-value)		2.14 ± 0.22	2.15 ± 0.19	0.622
WBC, 10^9/L (log-value)		0.78 ± 0.17	0.83 ± 0.16	0.079
HBV		12 (7.14%)	3 (5.67%)	0.724
Palliative 1st line protocol	FOLFIRINOX	13 (7.74%)	4 (7.55%)	
	Gemcitabine-based chemotherapy	151 (89.88%)	49 (92.45%)	
	Others	4 (2.38%)	0 (0%)	0.524
Chemotherapy beyond 1st line protocol	<3rd line palliative chemotherapy	143 (85.12%)	44 (83.02%)	
	≥3rd line palliative chemotherapy	25 (14.88%)	9 (16.98%)	0.712
Overall survival >8 months		91 (54.17%)	31 (58.49%)	0.581

CEA, carcinoembryonic antigen; CA199, carbohydrate antigen 199; ALT, alanine transaminase; AST, aspartate transaminase; WBC, white blood cell; HBV, hepatitis B virus.

of LR model 3 with 0.849 (95% C.I. = 0.785–0.899, $p < 0.05$) (**Figure 4C**). We built three ANN models, and all these models showed that the AUC of the ANN model was higher than that of the respective LR model, with ANN model 3 having the highest performance (**Table 3**).

All ANN and LR models were evaluated on the testing group of 53 patients. The accuracies of ANN model 1, ANN model 2 and ANN model 3 were 0.679, 0.698, and 0.774, respectively, which were all were higher than the accuracies of the respective LR models (0.623, 0.679, and 0.736). The k-statistics were 0.344,

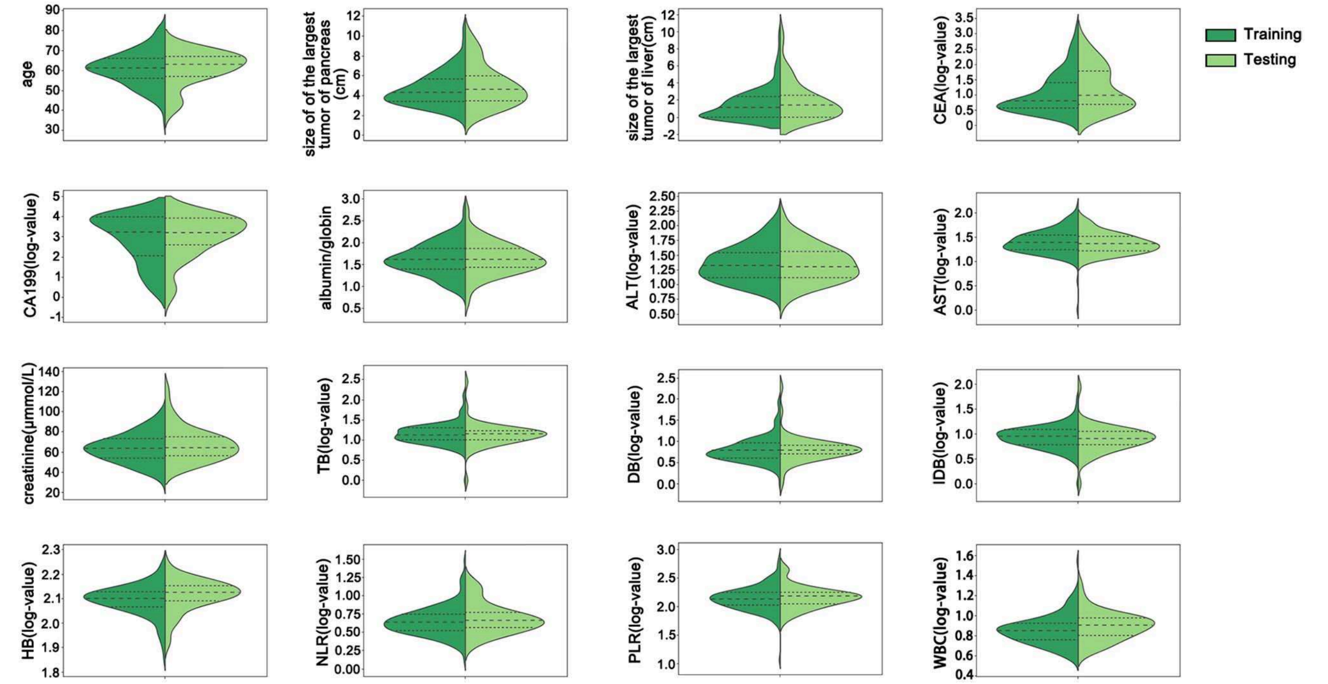


FIGURE 2 | The distribution of all continuous variables in the training and testing groups. There were no significant differences between the training and testing groups in any continuous variables. CEA, carcinoembryonic antigen; CA199, carbohydrate antigen 199; ALT, alanine transaminase; AST, aspartate transaminase; TB, total bilirubin; DB, direct bilirubin; IDB, indirect bilirubin; HB, hemoglobin; NLR, neutrophil/lymphocyte ratio; PLR, platelet/lymphocyte ratio; WBC, white blood cell count.

0.417, and 0.527 for ANN model 1, ANN model 2, and ANN model 3 and 0.233, 0.288, and 0.434 for LR model 1, LR model 2, and LR model 3, respectively. All LR models showed a lower accuracy **(Table 4)**.

DISCUSSION

Artificial neural networks have been developed as an effective statistical technique in the last 40 years (Dayhoff and DeLeo, 2001). They have been used in many fields and established as viable computational methodologies in computer science, biochemical and medical fields (Baxt and Skora, 1996; Milik et al., 1998; Gao et al., 2019; Yin et al., 2019; Deng et al., 2020; Yu et al., 2020). The network itself consists of an input layer, one or more hidden layers, and an output layer. Compared to logistic regression, ANN applies non-linear statistics and consists of a highly interconnected set of processing units (neurons) and weighted connections; the data used to build ANN can be applied to individual cases (Naguib et al., 1998).

For the ANN model, the usual ratio of training to testing group is 7:3 or 6:2:2 (when there is a validation dataset), but the radio is not strictly controlled, as previous studies have listed 5:2:3 or 6.4: 1.6: 0.2 (Cucchetti et al., 2010; Wu et al., 2017). In our study, the data before January 2018 were used as training group, and the data after January 2018 were used to simulate external validation. In the training group (n = 168), 133 (80%) patients were randomly selected to train the network, while 35 (20%) for cross validation. Thus, the total ratio is 6: 1.6: 2.4 (133:35:53), which was close to 6:2:2.

Many studies have demonstrated that ANN outperformed logistic regression in predicting survival, morbidity and mortality post-surgery and cancer diagnosis accuracy (Hanai et al., 2003; Pergialiotis et al., 2018; Wise et al., 2019). However, in the field of prostate cancer, the predictive accuracy of logistic regression is better than that of ANN (Chun et al., 2007; Kawakami et al., 2008). There are few applications of ANN in pancreatic cancer, and, the applications to date have been mainly in diagnosis and differential diagnosis (Ikeda et al., 1997; Norton et al., 2001; Honda et al., 2005). Very few studies have compared the abilities of ANN and logistic regression to predict the survival of advanced pancreatic cancer patients. Except for the significant clinical variables, some researchers showed non-significant variables still play important roles in prediction (Kawakami et al., 2008; Wu et al., 2017). So, we built three ANN models with different numbers of input to compare the AUC, PPV, PLR, sensitivity, and specificity, to help with patient stratification and clinical decision making in the absence of standardized prognostic risk scores for pancreatic cancer. ANN model 1 was built based on the three independent predictive factors for 8 month survival in the multivariate analysis, ANN model 2 was built based on the seven predictive factors for 8 month survival in the univariate analysis, and ANN model 3 was built based on all thirty-two variables. This is the first study comparing ANN and logistic regression in predicting unresectable pancreatic cancer patient survival. The median OS for metastatic pancreatic

TABLE 2 | Univariate and multivariate analyses of clinical characteristics associated with 8 month survival of the training group of 168 patients.

		Univariate analysis			**Multivariate analysis**		
Variables		HR	95% C.I.	*p*	HR	95% C.I.	*p*
Age, years		0.986	0.951–1.022	0.432			
Gender	Female	1					
	Male	0.702	0.372–1.323	0.274			
Main vascular invasion		1.419	0.764–2.633	0.268			
T	T1–T2	1					
	T3–T4	1.058	0.523–2.14	0.876			
N	N0	1					
	N1	0.566	0.245–1.303	0.181			
M	M0	1					
	M1	0.328	0.087–1.239	0.1			
Retroperitoneal lymph node metastasis		0.713	0.385–1.319	0.281			
Liver metastasis		0.51	0.268–0.972	**0.041**	0.854	0.35–2.08	0.727
Lung metastasis		1.186	0.451–3.116	0.729			
Peritoneal metastasis		0.741	0.296–1.851	0.521			
Ascites		0.921	0.369–2.302	0.861			
Size of of the largest tumor of pancreas, cm		0.892	0.741–1.073	0.224			
Tumor position of pancreas	Head and/or neck	1					
	Body and/or tail	1.078	0.579–2.007	0.812			
Stomach invasion		0.408	0.214–0.779	**0.007**	0.473	0.231–0.965	**0.04**
Duodenum invasion		0.669	0.272–1.646	0.381			
Liver metastasis number	<6	1					
	≥ 6	0.542	0.293–1.001	0.05			
Size of of the largest tumor of liver, cm		0.778	0.645–0.938	**0.008**	0.903	0.71–1.147	0.402
CEA, ng/mL (log-value)		1.132	0.733–1.748	0.575			
CA199, U/mL (log-value)		0.685	0.536–0.875	**0.002**	0.754	0.572–0.995	**0.046**
Albumin/globin	<1.48	1					
	≥ 1.48	2.885	1.487–5.596	**0.002**	2.36	1.106–5.038	**0.026**
ALT, U/L (log-value)		0.76	0.293–1.968	0.572			
AST, U/L (log-value)		0.62	0.171–2.248	0.467			
Creatinine, umol/L		1.009	0.987–1.03	0.427			
Total bilirubin, umol/L (log-value)		3.133	0.888–11.063	0.076			
Direct bilirubin, umol/L (log-value)		1.624	0.643–4.104	0.305			
Indirect bilirubin, umol/L (log-value)		3.81	0.996–14.578	0.051			
Hemoglobin, g/L (log-value)		0.099	0.001–13.453	0.356			
Neutrophil/lymphocyte (log-value)		0.409	0.121–1.378	0.149			
Platelet/lymphocyte (log-value)		0.818	0.2–3.346	0.78			
WBC, 10^9/L (log-value)		0.092	0.013–0.644	**0.016**	0.369	0.043–3.168	0.363
HBV		0.845	0.261–2.737	0.779			
Gemcitabine-based chemotherapy in 1st line		7.401	1.636–33.487	**0.009**	3.768	0.753–18.865	0.107

CEA, carcinoembryonic antigen; CA199, carbohydrate antigen 199; ALT, alanine transaminase; AST, aspartate transaminase; WBC, white blood cell; HBV, hepatitis B virus.
p-value $<$ 0.05 are indicated in bold.

cancer is approximately 6 months without systemic therapy. FOLFIRINOX offered enhanced median OS as compared to gemcitabine monotherapy (11.1 vs. 6.8 months) (Conroy et al., 2011). Gemcitabine plus nab-paclitaxel demonstrated superiority than gemcitabine with OS of 8.5 vs 6.7 months (Von Hoff et al., 2013). In our study, the median OS of the training group was 8 months, which is consistent with previous studies, so we chose 8 month survival as study's primary endpoint. The ANN models were found to be superior to linear discriminant analysis in predicting 8 month survival in the training group, and these results were further validated in the testing group. In addition, as the feature numbers increased, the prediction accuracy improved. Although ANN model 3 had the best performance, it was impractical, as 32 characters needed to be collected. Of the two rest models, ANN model 2 achieved higher accuracy than ANN model 1, and the number of characters needed to be collected were acceptable, so we recommend ANN model 2 for clinicians.

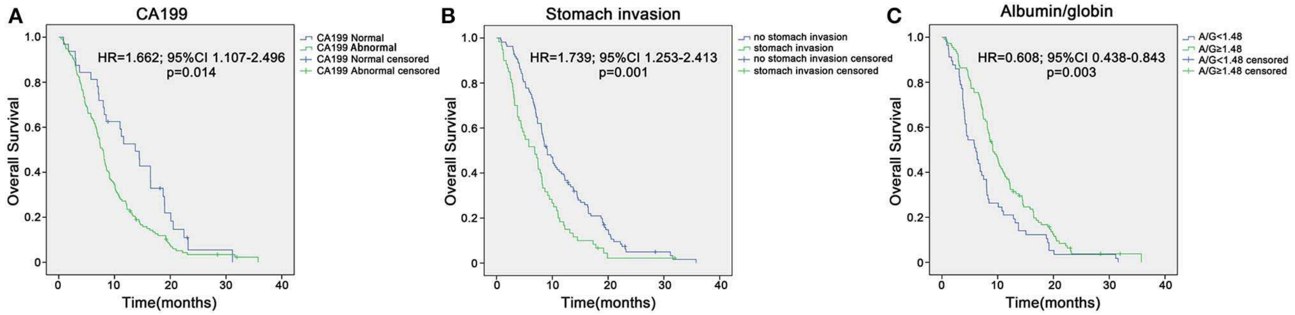


FIGURE 3 | Kaplan–Meier overall survival curves for the patients with unresectable pancreatic cancer in the training sample of 168 patients. **(A)** Overall survival of patients with abnormal CA199 decreased significantly compared with that of patients with normal CA199 (median survival, 7.80 vs. 13.73 months, $p < 0.05$). **(B)** Overall survival of patients with stomach invasion decreased significantly compared with that of patients with no stomach invasion (median survival, 6.83 vs. 9.10 months, $p < 0.05$). **(C)** Overall survival of patients with low AGR decreased significantly compared with that of patients with high AGR (median survival, 6.10 vs. 9.10 months, $p < 0.05$).

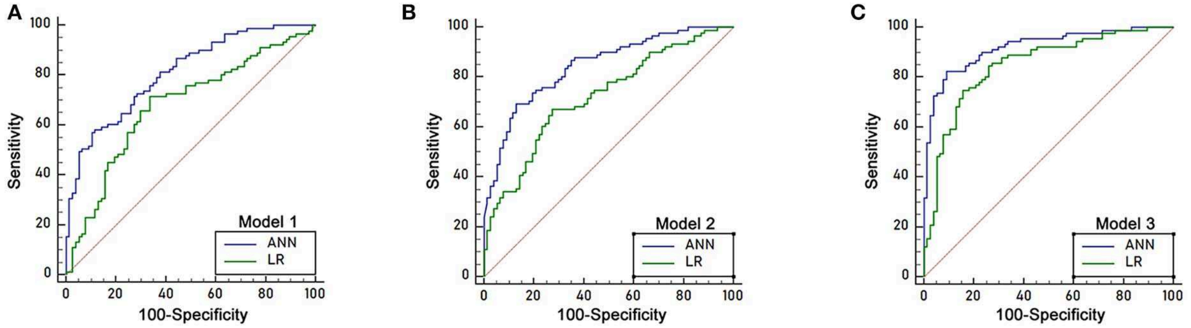


FIGURE 4 | ROC curve of the logistic regression models and ANN models in the training sample of 168 patients. **(A)** The area under the ROC curve (AUC) of ANN model 1 was 0.811 (95% C.I. = 0.743–0.867), which was higher than that of LR model 1 (AUC 0.680, 95% C.I. = 0.603–0.749, $p < 0.05$). **(B)** The area under the ROC curve (AUC) of ANN model 2 was 0.844 (95% C.I. = 0.780–0.895), which was higher than that of LR model 2 (AUC 0.722, 95% C.I. = 0.648–0.788, $p < 0.05$). **(C)** The area under the ROC curve (AUC) of ANN model 3 was 0.921 (95% C.I. = 0.869–0.957), which was higher than that of LR model 3 (AUC 0.849, 95% C.I. = 0.785–0.899, $p < 0.05$).

TABLE 3 | Accuracy of artificial neural network and logistic regression models in the training sample of 168 patients.

	AUC	95% C.I.	Cut-off	PPV		PLR		Sensitivity	Specificity
				OS ≤ 8 months	OS > 8 months	OS ≤ 8 months	OS > 8 months		
ANN model 1	0.811	0.743–0.867	0.559	0.6483	0.7662	0.4589	2.7735	0.6483	0.7662
LR model 1	0.680	0.603–0.749	0.5274	0.6578	0.7065	0.44	2.037	0.6493	0.7142
p-value	0.0008								
ANN model 2	0.844	0.780–0.895	0.6292	0.7052	0.863	0.3536	5.3307	0.6923	0.8701
LR model 2	0.722	0.648–0.788	0.5457	0.6511	0.7439	0.4532	2.4578	0.6703	0.7272
p-value	0.0006								
ANN model 3	0.921	0.869–0.957	0.4122	0.8117	0.9036	0.1962	7.9326	0.8241	0.8961
LR model 3	0.849	0.785–0.899	0.5601	0.7386	0.85	0.2994	4.7948	0.7472	0.8441
p-value	0.03								

ANN, artificial neural network; LR, logistic regression; PPV, positive predictive values; PLR, positive likelihood ratio. Stomach invasion, CA199, albumin/globin were used to build ANN model 1 and logistic model 1. Liver metastasis, stomach invasion, size of the largest tumor of liver, CA199, albumin/globin, white blood cell and gemcitabine-based chemotherapy in first line therapy were used to build ANN model 2 and LR model 2. All 32 characters were used to build ANN model 3 and LR model 3.

All patients included had unresectable pancreatic cancer. We collected as many clinical markers related to tumor prognosis as possible. Finally, we addressed the prognostic significance of AGR, CA199 and stomach invasion in univariate and multivariate analyses. Albumin and globulin are human serum proteins. Albumin reflects nutritional status and systemic

TABLE 4 | Prediction accuracy of ANN and logistic regression models in the testing group of 53 patients.

	Model 1		Model 2		Model 3	
	Accuracy	k	Accuracy	k	Accuracy	k
ANN	0.679	0.344	0.698	0.417	0.774	0.527
LR	0.623	0.233	0.679	0.288	0.736	0.434

ANN, artificial neural network; LR, logistic regression.

inflammatory response in cancer patients (McMillan et al., 2001). Poor nutrition status (hypoalbuminemia) has been proven to be a negative factor of survival in multiple cancers, including hepatobiliary, lung, gastrointestinal, CNS, reproductive, and breast cancers (Onate-Ocana et al., 2007; Gupta and Lis, 2010). On the other hand, haemoglobulin plays an important role in immunity and inflammation. Chronic inflammation is considered a contributor to tumor proliferation, immune evasion and metastasis. Therefore, low albumin and high haemoglobulin may decrease the survival of cancer patients. In previous studies, the AGR has been used as a prognostic indicator in diverse human cancers (Azab et al., 2013; Lv et al., 2018). However, AGR cutoff values are diverse in different studies (Lv et al., 2018), and more accurate AGR cutoff values are expected to be found.

Tumor invasion of adjacent structures is not captured in the TNM classification of pancreatic cancer from the 8th American Joint Committee on Cancer. However, a multidisciplinary consensus group recently created a standardized language for the reporting of imaging results, and reporting the presence of extrapancreatic tumor extension was recommended (Al-Hawary et al., 2014). Stomach, as one of the adjacent structures to pancreas, were recommended to be reported present or absent of tumor involved. Stomach invasion carries the risk of haematemesis. Although the incidence of haematemesis is low, it can be life-threatening if it occurs. Additionally, according to NCCN guidelines, SBRT should not be used if invasion of the stomach is observed on imaging. These results prove that stomach invasion is a problem worthy of clinical concern. In our study, Kaplan–Meier analysis showed that overall survival decreased significantly in the stomach invasion group. To the best of our knowledge, this is the first report indicating that stomach invasion is an independent prognostic factor for the 8 month survival of advanced pancreatic cancer patients. These features deserve the doctors' attention.

Treatment option is another important factor that impacts patients' prognosis. In our study, gemcitabine-based chemotherapy as the first-line therapy (HR 7.401, $p = 0.009$) were related to 8 month survival in the univariate analysis in the training group. However, it was not confirmed in the multivariate analysis. Different from randomized clinical trial, patients' status varied in retrospective study. As there was a preference among doctors and patients to select treatment based on performance status and fitness to withstand toxicities, bias is hard to be avoided. The relative small sample size may be another reason that failed to meet the statistical significance in multivariate analysis.

In addition to selecting predictive factors for 8 month survival, we also tried to identify predictive factors for 4 month progression-free survival. Even though nine factors (liver metastasis, stomach invasion, liver metastasis number, size of the largest tumor of the liver, CA199, AGR, neutrophil/lymphocyte ratio, platelet/lymphocyte ratio, and WBC count) showed statistical significance in univariate analysis, none of them were confirmed in the multivariate analysis based on the training group data (**Supplementary Table 1**).

Our study had several strengths. Our study made full use of clinical data that is very convenient and easy to obtain to build models to predict the survival of patients. Our models help make more accurate predictions of OS, thus optimizing patient selection for appropriate treatment and achieving more personalized management. In addition, more accurate prediction of OS will facilitate well-balanced arms in clinical trials (Vernerey et al., 2016) and allow cross-study comparisons for research purposes. Moreover, the clinical and biological parameters in the training and testing groups were comparable ($p > 0.05$), and the testing group displayed convincing performance. However, as our models were built and tested on data that originated from one center, a multicentre study should be performed in the future to verify our findings.

CONCLUSIONS

AGR, CA199, and stomach invasion were independent predictive factors for 8 month survival in unresectable pancreatic cancer patients. We developed convenient and reliable ANN models predicting the 8 month survival of patients with unresectable pancreatic cancer, and the validation showed superior predictive accuracy of ANN over logistic regression models. Our models may help clinicians evaluate the 8 month survival time and make appropriate recommendations for the most suitable treatment options for their patients.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the First Affiliated Hospital, Zhejiang University School of Medicine. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZT wrote the manuscript. HM, JZ, BL, and XB analyzed the data. YL, XX, and CG collected the clinical and pathological information from the cancer patients. YZ and LL designed the study. WF, SD, and PZ revised the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81472346), the Natural Science Foundation of Zhejiang Province (LY20H160033). Educational Science Plan of Zhejiang Province (2020SCG200).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00196/full#supplementary-material

REFERENCES

Al-Hawary, M. M., Francis, I. R., Chari, S. T., Fishman, E. K., Hough, D. M., Lu, D. S., et al. (2014). Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the society of abdominal radiology and the american pancreatic association. *Gastroenterology* 146, 291–304.e1. doi: 10.1053/j.gastro.2013.11.004

Are, C., Afuh, C., Ravipati, L., Sasson, A., Ullrich, F., and Smith, L. (2009). Preoperative nomogram to predict risk of perioperative mortality following pancreatic resections for malignancy. *J. Gastrointest. Surg.* 13, 2152–2162. doi: 10.1007/s11605-009-1051-z

Azab, B., Kedia, S., Shah, N., Vonfrolio, S., Lu, W., Naboush, A., et al. (2013). The value of the pretreatment albumin/globulin ratio in predicting the long-term survival in colorectal cancer. *Int. J. Colorectal Dis.* 28, 1629–1636. doi: 10.1007/s00384-013-1748-z

Baxt, W. G., and Skora, J. (1996). Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet* 347, 12–15. doi: 10.1016/S0140-6736(96)91555-X

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13, 281–305.

Bradley, E. L. III. (2008). Long-term survival after pancreatoduodenectomy for ductal adenocarcinoma: the emperor has no clothes? *Pancreas* 37, 349–351. doi: 10.1097/MPA.0b013e31818e9100

Braga, M., Capretti, G., Pecorelli, N., Balzano, G., Doglioni, C., Ariotti, R., et al. (2011). A prognostic score to predict major complications after pancreaticoduodenectomy. *Ann. Surg.* 254, 702–707; discussion 707–8. doi: 10.1097/SLA.0b013e31823598fb

Chun, F. K., Karakiewicz, P. I., Briganti, A., Walz, J., Kattan, M. W., Huland, H., et al. (2007). A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. *BJU Int.* 99, 794–800. doi: 10.1111/j.1464-410X.2006.06694.x

Conroy, T., Desseigne, F., Ychou, M., Bouche, O., Guimbaud, R., Becouarn, Y., et al. (2011). FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N. Engl. J. Med.* 364, 1817–1825. doi: 10.1056/NEJMoa1011923

Cress, R. D., Yin, D., Clarke, L., Bold, R., and Holly, E. A. (2006). Survival among patients with adenocarcinoma of the pancreas: a population-based study (United States). *Cancer Causes Control* 17, 403–409. doi: 10.1007/s10552-005-0539-4

Cucchetti, A., Piscaglia, F., Grigioni, A. D., Ravaioli, M., Cescon, M., Zanello, M., et al. (2010). Preoperative prediction of hepatocellular carcinoma tumour grade and micro-vascular invasion by means of artificial neural network: a pilot study. *J. Hepatol.* 52, 880–888. doi: 10.1016/j.jhep.2009.12.037

Dasari, B. V., Roberts, K. J., Hodson, J., Stevens, L., Smith, A. M., Hubscher, S. G., et al. (2016). A model to predict survival following pancreaticoduodenectomy for malignancy based on tumour site, stage and lymph node ratio. *HPB* 18, 332–338. doi: 10.1016/j.hpb.2015.11.008

Dayhoff, J. E., and DeLeo, J. M. (2001). Artificial neural networks: opening the black box. *Cancer* 91(Suppl. 8), 1615–1635. doi: 10.1002/1097-0142(20010415)91:8+<1615::aid-cncr1175>3.0.co;2-l

Deng, S., Xiang, Z., Taheri, J., Mohammad, K. A., Yin, J., Zomaya, A., et al. (2020). Optimal application deployment in resource constrained distributed edges. *IEEE Trans. Mobile Comput.* 99, 1–1. doi: 10.1109/TMC.2020.2970698

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

Gao, W., Zhu, Y., Zhang, W., Zhang, K., and Gao, H. (2019). A hierarchical recurrent approach to predict scene graphs from a visual-attention-oriented perspective. *Comput. Intell.* 35, 496–516. doi: 10.1111/coin.12202

Ghoshal, U. C., and Das, A. (2008). Models for prediction of mortality from cirrhosis with special reference to artificial neural network: a critical review. *Hepatol. Int.* 2, 31–38. doi: 10.1007/s12072-007-9026-1

Gupta, D., and Lis, C. G. (2010). Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature. *Nutr. J.* 9:69. doi: 10.1186/1475-2891-9-69

Hanai, T., Yatabe, Y., Nakayama, Y., Takahashi, T., Honda, H., Mitsudomi, T., et al. (2003). Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer Sci.* 94, 473–477. doi: 10.1111/j.1349-7006.2003.tb01467.x

Hidalgo, M., Cascinu, S., Kleeff, J., Labianca, R., Lohr, J. M., Neoptolemos, J., et al. (2015). Addressing the challenges of pancreatic cancer: future directions for improving outcomes. *Pancreatology* 15, 8–18. doi: 10.1016/j.pan.2014.10.001

Honda, K., Hayashida, Y., Umaki, T., Okusaka, T., Kosuge, T., Kikuchi, S., et al. (2005). Possible detection of pancreatic cancer by plasma protein profiling. *Cancer Res.* 65, 10613–10622. doi: 10.1158/0008-5472.CAN-05-1851

Ikeda, M., Ito, S., Ishigaki, T., and Yamauchi, K. (1997). Evaluation of a neural network classifier for pancreatic masses based on CT findings. *Comput. Med. Imaging Graph.* 21, 175–183. doi: 10.1016/S0895-6111(97)00006-2

Kawakami, S., Numao, N., Okubo, Y., Koga, F., Yamamoto, S., Saito, K., et al. (2008). Development, validation, and head-to-head comparison of logistic regression-based nomograms and artificial neural network models predicting prostate cancer on initial extended biopsy. *Eur. Urol.* 54, 601–611. doi: 10.1016/j.eururo.2008.01.017

Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V., et al. (2016). Pancreatic cancer. *Nat. Rev. Dis. Primers* 2:16022. doi: 10.1038/nrdp.2016.22

Kuhlmann, K. F., de Castro, S. M., Wesseling, J. G., ten Kate, F. J., Offerhaus, G. J., Busch, O. R., et al. (2004). Surgical treatment of pancreatic adenocarcinoma; actual survival and prognostic factors in 343 patients. *Eur. J. Cancer* 40, 549–558. doi: 10.1016/j.ejca.2003.10.026

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lisboa, P. J., and Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw.* 19, 408–415. doi: 10.1016/j.neunet.2005.10.007

Lv, G. Y., An, L., Sun, X. D., Hu, Y. L., and Sun, D. W. (2018). Pretreatment albumin to globulin ratio can serve as a prognostic marker in human cancers: a meta-analysis. *Clin. Chim. Acta* 476, 81–91. doi: 10.1016/j.cca.2017.11.019

McMillan, D. C., Watson, W. S., O'Gorman, P., Preston, T., Scott, H. R., and McArdle, C. S. (2001). Albumin concentrations are primarily determined by the body cell mass and the systemic inflammatory response in cancer patients with weight loss. *Nutr. Cancer.* 39, 210–213. doi: 10.1207/S15327914nc392_8

Milik, M., Sauer, D., Brunmark, A. P., Yuan, L., Vitiello, A., Jackson, M. R., et al. (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat. Biotechnol.* 16, 753–756. doi: 10.1038/nbt0898-753

Miura, T., Hirano, S., Nakamura, T., Tanaka, E., Shichinohe, T., Tsuchikawa, T., et al. (2014). A new preoperative prognostic scoring system to predict prognosis in patients with locally advanced pancreatic body cancer who undergo distal pancreatectomy with en bloc celiac axis resection: a retrospective cohort study. *Surgery* 155, 457–467. doi: 10.1016/j.surg.2013.10.024

Naguib, R. N., Robinson, M. C., Neal, D. E., and Hamdy, F. C. (1998). Neural network analysis of combined conventional and experimental prognostic markers in prostate cancer: a pilot study. *Br. J. Cancer*. 78, 246–250. doi: 10.1038/bjc.1998.472

Naito, Y., Ishikawa, H., Sadashima, E., Okabe, Y., Takahashi, K., Kawahara, R., et al. (2019). Significance of neoadjuvant chemoradiotherapy for borderline resectable pancreatic head cancer: pathological local invasion and microvessel invasion analysis. *Mol. Clin. Oncol*. 11, 225–233. doi: 10.3892/mco.2019.1885

Neoptolemos, J. P., Kleeff, J., Michl, P., Costello, E., Greenhalf, W., and Palmer, D. H. (2018). Therapeutic developments in pancreatic cancer: current and future perspectives. *Nat. Rev. Gastroenterol. Hepatol*. 15, 333–348. doi: 10.1038/s41575-018-0005-x

Norton, I. D., Zheng, Y., Wiersema, M. S., Greenleaf, J., Clain, J. E., and Dimagno, E. P. (2001). Neural network analysis of EUS images to differentiate between pancreatic malignancy and pancreatitis. *Gastrointest. Endosc*. 54, 625–629. doi: 10.1067/mge.2001.118644

Onate-Ocana, L. F., Aiello-Crocifoglio, V., Gallardo-Rincon, D., Herrera-Goepfert, R., Brom-Valladares, R., Carrillo, J. F., et al. (2007). Serum albumin as a significant prognostic factor for patients with gastric carcinoma. *Ann. Surg. Oncol*. 14, 381–389. doi: 10.1245/s10434-006-9093-x

Penny, W., and Frost, D. (1996). Neural networks in clinical medicine. *Med. Decis. Making* 16, 386–398. doi: 10.1177/0272989X9601600409

Pergialiotis, V., Pouliakis, A., Parthenis, C., Damaskou, V., Chrelias, C., Papantoniou, N., et al. (2018). The utility of artificial neural networks and classification and regression trees for the prediction of endometrial cancer in postmenopausal women. *Public Health* 164, 1–6. doi: 10.1016/j.puhe.2018.07.012

Vernerey, D., Huguet, F., Vienot, A., Goldstein, D., Paget-Bailly, S., Van Laethem, J. L., et al. (2016). Prognostic nomogram and score to predict overall survival in locally advanced untreated pancreatic cancer (PROLAP). *Br. J. Cancer* 115, 281–289. doi: 10.1038/bjc.2016.212

Von Hoff, D. D., Ervin, T., Arena, F. P., Chiorean, E. G., Infante, J., Moore, M., et al. (2013). Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N. Engl. J. Med*. 369, 1691–1703. doi: 10.1056/NEJMoa1304369

Wise, E. S., Amateau, S. K., Ikramuddin, S., and Leslie, D. B. (2019). Prediction of thirty-day morbidity and mortality after laparoscopic sleeve gastrectomy: data from an artificial neural network. *Surg. Endosc.* doi: 10.1007/s00464-019-07130-0. [Epub ahead of print].

Wu, C. F., Wu, Y. J., Liang, P. C., Wu, C. H., Peng, S. F., and Chiu, H. W. (2017). Disease-free survival assessment by artificial neural networks for hepatocellular carcinoma patients after radiofrequency ablation. *J. Formosan Med. Assoc*. 116, 765–773. doi: 10.1016/j.jfma.2016.12.006

Xu, J., Shi, K. Q., Chen, B. C., Huang, Z. P., Lu, F. Y., and Zhou, M. T. (2017). A nomogram based on preoperative inflammatory markers predicting the overall survival of pancreatic ductal adenocarcinoma. *J. Gastroenterol. Hepatol*. 32, 1394–1402. doi: 10.1111/jgh.13676

Yin, Y., Chen, L., Xu, Y., Wan, J., Zhang, H., and Mai, Z. (2019). QoS prediction for service recommendation with deep feature learning in edge computing environment. *Mobile Netw. Appl.* doi: 10.1007/s11036-019-01241-7

Yu, J., Zhu, C., Zhang, J., Huang, Q., and Tao, D. (2020). Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE Trans. Neural Netw. Learn. Syst*. 31, 661–674. doi: 10.1109/TNNLS.2019.2908982

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

frontiers
in Bioengineering and Biotechnology

BRIEF RESEARCH REPORT
published: 15 May 2020
doi: 10.3389/fbioe.2020.00479

P-Wave Area Predicts New Onset Atrial Fibrillation in Mitral Stenosis: A Machine Learning Approach

Gary Tse[1,2], *Ishan Lakhani*[3], *Jiandong Zhou*[4*], *Ka Hou Christien Li*[5], *Sharen Lee*[3], *Yingzhi Liu*[3], *Keith Sai Kit Leung*[6], *Tong Liu*[1], *Adrian Baranchuk*[7] and *Qingpeng Zhang*[4*]

[1] Tianjin Key Laboratory of Ionic-Molecular Function of Cardiovascular Disease, Department of Cardiology, Tianjin Institute of Cardiology, Second Hospital of Tianjin Medical University, Tianjin, China, [2] Xiamen Cardiovascular Hospital, Xiamen University, Xiamen, China, [3] Laboratory of Cardiovascular Physiology, Li Ka Shing Institute of Health Sciences, Shatin, China, [4] School of Data Science, City University of Hong Kong, Kowloon, China, [5] Faculty of Medicine, Newcastle University, Newcastle, United Kingdom, [6] Aston Medical School, Aston University, Birmingham, United Kingdom, [7] Heart Rhythm Service, Kingston General Hospital, Queen's University, Kingston, ON, Canada

OPEN ACCESS

Edited by:
Wenbing Zhao,
Cleveland State University,
United States

Reviewed by:
Olaf Doessel,
Karlsruhe Institute of Technology
(KIT), Germany
Andrew Teh,
Eastern Health, Australia

***Correspondence:**
Jiandong Zhou
jiandzhou3-c@my.cityu.edu.hk
Qingpeng Zhang
qingpeng.zhang@cityu.edu.hk

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

Received: *21 January 2020*
Accepted: *24 April 2020*
Published: *15 May 2020*

Citation:
Tse G, Lakhani I, Zhou J, Li KHC, Lee S, Liu Y, Leung KSK, Liu T, Baranchuk A and Zhang Q (2020) P-Wave Area Predicts New Onset Atrial Fibrillation in Mitral Stenosis: A Machine Learning Approach. Front. Bioeng. Biotechnol. 8:479. doi: 10.3389/fbioe.2020.00479

Introduction: Mitral stenosis is associated with an atrial cardiomyopathic process, leading to abnormal atrial electrophysiology, manifesting as prolonged *P*-wave duration (PWD), larger *P*-wave area, increased *P*-wave dispersion ($\text{PWD}_{\text{max}} - \text{PWD}_{\text{min}}$), and/or higher *P*-wave terminal force on lead V1 (PTFV1) on the electrocardiogram.

Methods: This was a single-center retrospective study of Chinese patients, diagnosed with mitral stenosis in sinus rhythm at baseline, between November 2009 and October 2016. Automated ECG measurements from raw data were determined. The primary outcome was incident atrial fibrillation (AF).

Results: A total 59 mitral stenosis patients were included (age 59 [54–65] years, 13 (22%) males). New onset AF was observed in 27 patients. Age (odds ratio [OR]: 1.08 [1.01–1.16], $P = 0.017$), systolic blood pressure (OR: 1.03 [1.00–1.07]; $P = 0.046$), mean *P*-wave area in V3 (odds ratio: 3.97 [1.32–11.96], $P = 0.014$) were significant predictors of incident AF. On multivariate analysis, age (OR: 1.08 [1.00–1.16], $P = 0.037$) and *P*-wave area in V3 (OR: 3.64 [1.10–12.00], $P = 0.034$) remained significant predictors of AF. Receiver-operating characteristic (ROC) analysis showed that the optimum cut-off for *P*-wave area in V3 was 1.45 Ashman units (area under the curve: 0.65) for classification of new onset AF. A decision tree learning model with individual and non-linear interaction variables with age achieved the best performance for outcome prediction (accuracy = 0.84, precision = 0.84, recall = 0.83, *F*-measure = 0.84).

Conclusion: Atrial electrophysiological alterations in mitral stenosis can detected on the electrocardiogram. Age, systolic blood pressure, and *P*-wave area in V3 predicted new onset AF. A decision tree learning model significantly improved outcome prediction.

Keywords: mitral stenosis, mitral valve, *P*-wave area, decision tree, machine learning

INTRODUCTION

Inter-atrial block (IAB) results from impaired conduction of action potentials along Bachmann's bundle that connects the right and left atria (Tse et al., 2016). It is characterized electrocardiographically by a prolonged *P*-wave duration of >120 ms. This condition results in delayed and asynchronous activation of the left atrium (Agarwal et al., 2003; Budeus et al., 2005; Caldwell et al., 2014). IAB has been associated with higher incidence of stroke as well as cardiovascular and all-cause mortality (Ariyarajah et al., 2007; Magnani et al., 2011). However, it is unclear any benefit derived from early initiation of anti-coagulation in IAB before the development of atrial fibrillation (AF), and the risk may differ depending on the severity of IAB and the presence of other cardio-metabolic co-morbidities. Two other measures have been used to assess atrial electrophysiological remodeling. Firstly, *P*-wave dispersion, defined as the difference between maximum and minimum *P*-wave duration (PWD), is a measure of heterogeneous and discontinuous atrial activation. Secondly, *P*-wave terminal force in V1 (PTFV1) is a marker of left atrial disease independently of structural or pressure changes in the left atrium (Morris and Thompson, 1964) and has been shown to be a predictor of future incident AF (Martin Garcia et al., 2012). Prolonged PWDs, measured from amplified and digitized ECG signals obtained in sinus rhythm, predicted AF recurrence after pulmonary isolation procedures (Jadidi et al., 2018). Moreover, the area of the *P*-wave initial portion was independently associated with the development of AF in patients with left atrial overload (Ishida et al., 2010). AF complexity parameters derived from the ECG also predicted long-term outcomes following catheter ablation (Lankveld et al., 2016).

Mitral stenosis is a valvular disease frequently seen in parts of Asia, causing significant morbidity and mortality. In this condition, the most common arrhythmia encountered is AF, but there are limited data on electrocardiographic changes that reflect ongoing atrial cardiomyopathic process that precedes the development of fibrillation. Mitral stenosis patients have longer PWDs and higher *P*-wave dispersion than control subjects without mitral stenosis (Guntekin et al., 2008). Another study confirmed this observation and further demonstrated a significant correlation between maximum PWDs and left atrial size, transmitral valve gradient, and a negative correlation with mitral valve area (Rezaian et al., 2007). PTFV1 is higher in mitral stenosis and is a predictor of disease severity (Yuce et al., 2011). However, there are limited published data regarding the incidence of IAB, the relative contributions of partial and advanced IAB, and whether these indices predict incident AF in mitral stenosis.

METHODS

This study received approval from The Joint Chinese University of Hong Kong—New Territories East Cluster Clinical Research Ethics Committee. Clinical and electrocardiographic details of a cohort of Chinese patients referred to our center, which is a tertiary referral center and teaching hospital, between November 2009 and October 2016, for echocardiography, were analyzed retrospectively. Inclusion criteria were mitral stenosis patients with raw ECG data files available for analysis.

Definitions, Data Extraction, Electrocardiographic Measurements, and Primary Outcome

The following clinical details were obtained from the patients: age, gender, blood pressure, smoking status, diabetes mellitus, hypertension, hypercholesterolemia, and ischemic heart disease. For electrocardiographic parameters, data were extracted from patients who had ECGs that did not show atrial fibrillation (AF). The following parameters were manually measured by two investigators from the ECGs showing sinus rhythm. The following *P*-wave variables were determined from ECGs of patients in sinus rhythm. Automated measurements from raw ECG data were extracted from the Philips ECGVue program (Standard Edition). The ECG waveform data is captured at a sample rate of 4 MHz and reduced to 500 samples per second with 5 μV resolution. The mean, minimum, maximum, and standard deviation of different *P*-wave variables were calculated from values from all 12 leads (**Figure 1**). *P*-wave dispersion was defined as the maximum difference in PWD. *P*-wave terminal force in V1 (PTFV1) was defined as the area subtended by the terminal negative component of a biphasic *P*-wave in lead V1, with the area calculated by multiplication of the duration and depth of the waveform (He et al., 2017). The primary endpoint of this study was new onset persistent or permanent atrial fibrillation (AF). Paroxysmal AF at baseline or detected follow-up was excluded. The endpoint was met if AF was detected in at least two ECGs on follow-up 1 year apart in an absence of sinus rhythm detection in the intervening period.

Statistical Analysis, Non-linear Variables, and Decision Tree Learning

Data were expressed as median [lower quartile to upper quartile]. Categorical data were analyzed by Fisher's exact test. Differences between study groups were tested using Kruskal-Wallis ANOVA. $P < 0.05$ was considered statistically significant. Non-linear interactions (e.g., interactions formed by some important individual variable) play an important role in predicting the outcome. The consideration of non-linear interactions overcomes linear model's assumption that the dependent and independent variables are linearly related. In this study, the logarithmic form of the multiplication non-linear items formed by age (important individual variable) and other continuous variables were considered, i.e., $\log(\text{age}^* x_i)$, where x_i denotes the *i*th continuous variable. The adoption of logarithmic transformation is to obtain equivalent inference on variable-outcome associations while avoid the bias due to exponentiation on some squared and cubed variables. The non-linear variables considered were log(age*systolic blood pressure), log(age*diastolic blood pressure), log(age*diabetes mellitus), log(age*hypercholesterolaemia, log(age*ischaemic heart disease), log(age*left atrial diameter), log(age*mitral valve area), log(age*mitral valve gradient), log(age*mitral

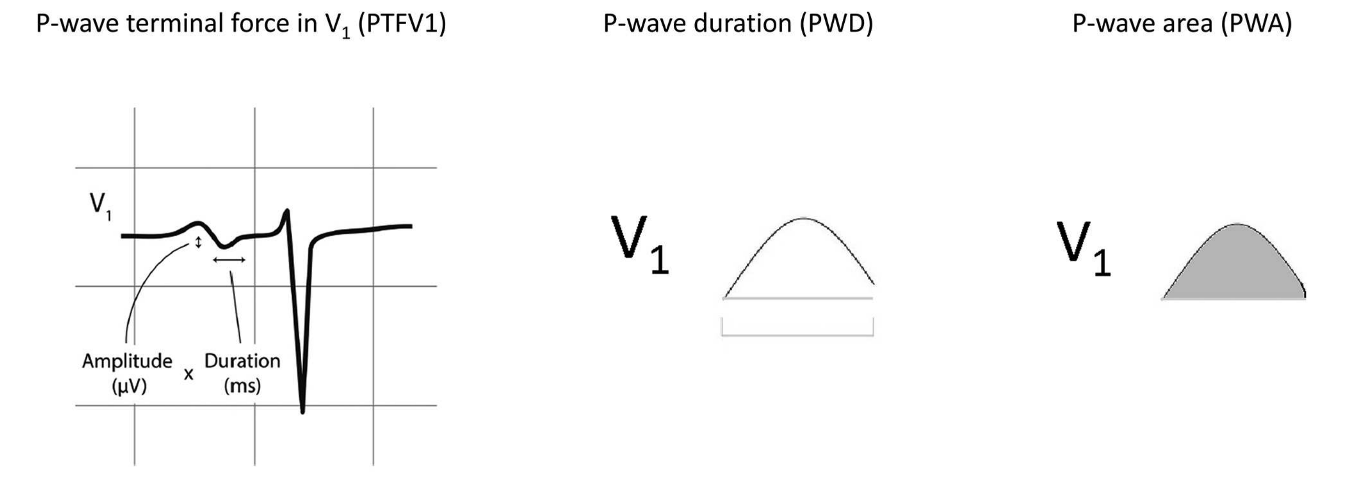


FIGURE 1 | Normal inter-atrial conduction, partial and advanced inter-atrial block and left atrial enlargement. *P*-wave terminal force in V1 (PTFV1), *P*-wave duration (PWD), and *P*-wave area. Adapted from (He et al., 2017) with permission.

stenosis severity), log(age* *P*-wave area in v3). For instance, for a patient whose age is 58 years old and has systolic blood pressure 110 mmHg, we generate the value of non-linear variable log(age*systolic blood pressure) by calculating log(58*110) = 3.8048. The values for the other non-linear variables are obtained in a similar way. Then these non-linear variables and the individual variables are together used as input in the risk prediction model. The reproduction of the non-linear variables can be obtained since there exists one-one mapping between the non-linear variables and the individual variable pairs that form them.

Decision tree learning uses a decision tree module (as a predictive model) to determine the outcome (or target value, represented by leaves) of a sample based on the associated observations (represented by branches) for model classification and prediction. The principles of decision tree are illustrated in **Figure 2**. In this study, a decision tree learning approach (classification and regression tree, CART Rutkowski et al., 2014 was used to predict new onset AF. Specifically, the non-linear variables (including log(age*systolic blood pressure), log(age*diastolic blood pressure), log(age*diabetes mellitus), log(age*hypercholesterolaemia), log(age*ischaemic heart disease), log(age*left atrial diameter), log(age*mitral valve area), log(age*mitral valve gradient), log(age*mitral stenosis severity), log(age* *P*-wave area in v3) together with individual variables (including sex, age, systolic blood pressure, diastolic blood pressure, diabetes mellitus, hypercholesterolaemia, ischaemic heart disease, left atrial diameter, mitral valve area, mitral valve gradient, mitral stenosis severity, *P*-wave area in v3) were used as input to the DTL model, in order to predict the new onset of AF outcome in mitral stenosis. In the DTL model, leaves represent class label of new onset AF and branches represent feature conjunctions (both of non-linear variables and individual variables) that lead to new onset AF. DTL uses Gini index to construct a decision tree, which is calculated by the formula $Gini = 1 - \sum_{i=1}^{C} \left(p_i\right)^2$ representing the probability of a particular patient being wrongly classified when it is randomly chosen and c denotes the number of class ($c = 2$ for new onset AF classification in this study). The Gini index is used to create split points by considering a binary split for each variable in DTL. It varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

Here we present an example to show the construction process of DTL with Gini index. Firstly, the frequency table of new onset AF outcome (27 with "yes" and 32 with "no") was considered: there are 19 of 46 females with new onset AF, while eight of 13 males with new onset AF. The Gini index for sex as a decision node to split the tree was calculated: Gini index (new onset AF, male) = 2*(8/13)*(1–8/13) =0.4734, Gini index (new onset AF, female) = 2*(19/46)*(1–19/46) = 0.4849. In addition, the frequency table of ischaemic heart disease was considered: eight of 12 with ischaemic heart disease had new onset AF, and 19 of 47 without ischaemic heart disease had new onset AF. The corresponding Gini index was calculated by Gini(new onset AF, ischaemic heart disease as "yes") = 2*(8/12)*(1–8/12)=0.4444, Gini index(new onset AF, ischaemic heart disease as "no")=2*(19/47)*(1–19/47) = 0.4817. In the same manner, the Gini indices of all the variables (the Information Gain of a continuous variable, such as age, was discretized first) was calculated. DTL splits the dataset on different variables by referring to the Gini indices, and the variable with the lowest Gini index value was selected as the decision node. DTL divided the dataset by its branches and repeat the same process on every branch. A branch with zero Gini index becomes a leaf node, while a branch with Gini index larger than zero needs further splitting. DTL is run recursively in a similar way on the non-leaf branches, until all data were classified.

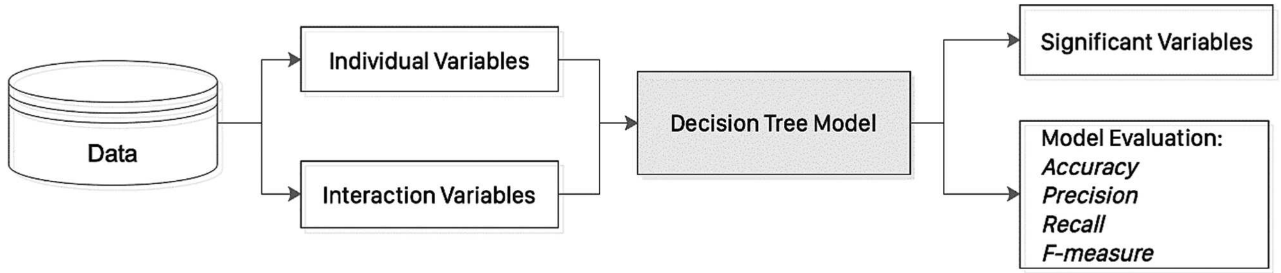


FIGURE 2 | An illustration of the decision tree learning approach.

RESULTS

Electrocardiograms of patients with mitral stenosis ($n = 155$) were screened. In this cohort, 96 had atrial fibrillation on admission to our hospital without prior ECGs in sinus rhythm available for analysis. The remaining 59 patients were in sinus rhythm and were analyzed further. A graphical representation of the different *P*-wave indices is shown in **Figure 1**. The baseline characteristics of this cohort are shown in **Supplementary Table 1**. The median age was 59 [54–65] years old, and 13 patients (22%) were male.

A total of 27 patients developed new onset AF over a median follow-up of 58 [48–76] months. Patients with new onset AF had similar mean PWD (102 [95–118] vs. 101 [89–115] ms), minimum PWD (56 [40–68] vs. 52 [38–65]), maximum PWD (152 [132–164] vs. 136 [123–160] ms), *P*-wave dispersion (84 [62–116] vs. 82 [57–110] ms), standard deviation of PWD (28 [17–33] vs. 24 [16–33] ms), mean *P*-wave amplitude (0.11 [0.09–0.15] vs. 0.11 [0.09–0.13] mV), minimum *P*-wave amplitude (0.05 [0.03–0.07] vs. 0.05 [0.03–0.06] mV), maximum *P*-wave amplitude (0.18 [0.15–0.25] vs. 0.19 [0.16–0.23] mV), dispersion of *P*-wave amplitude (0.14 [0.10–0.17] vs. 0.14 [0.11–0.19] mV), standard deviation of *P*-wave amplitude (0.04 [0.03–0.06] vs. 0.04 [0.03–0.06] mV), mean *P*-wave area (0.12 [0.09–0.14] vs. 0.10 [0.09–0.13] Ashman units [40 ms × 0.1 mV]), minimum *P*-wave area (0.05 [0.04–0.07] vs. 0.05 [0.03–0.06] Ashman units), maximum *P*-wave area (0.22 [0.17–0.26] vs. 0.18 [0.14–0.26] Ashman units), dispersion of *P*-wave area (0.16 [0.11–0.20] vs. 0.12 [0.11–0.16] Ashman units), standard deviation of *P*-wave area (0.05 [0.04–0.06] vs. 0.04 [0.03–0.04] Ashman units), Neither *P*-wave initial force in V1 (PIFV1: 7.6 [3.7–11.7] vs. 3.6 [1.7–11.0] ms.mV) nor *P*-wave terminal force in V1 (PTFV1: 2.7 [0–6.9] vs. 3.8 [0–8.3] ms.mV) differed between the groups. Similarly, no difference in left atrial diameter was detected (4.9 [4.3–5.0] vs. 4.4 [3.9–4.9] cm, $P = 0.1179$). By contrast, *P*-wave area in V3 was significantly higher in the new onset AF group (1.0 [0.7–1.9] vs. 0.8[0.5–1.1]; $P = 0.045$).

The results of univariate logistic regression are shown in **Supplementary Table 2**. Age (odds ratio [OR]: 1.08 [1.01–1.16], $P = 0.017$), systolic blood pressure (OR: 1.03 [1.00–1.07]; $P = 0.046$), mean *P*-wave area in V3 (odds ratio: 3.97 [1.32–11.96], $P = 0.014$) were significant predictors of incident AF. Variables that achieved $P < 0.10$ in univariate logistic regression were included in the multivariable model (**Supplementary Table 3**). On multivariate analysis, age (OR: 1.08 [1.00–1.16], $P = 0.037$) and *P*-wave area in V3 (OR: 3.64 [1.10–12.00], $P = 0.034$) remained significant predictors of AF. Receiver-operating characteristic (ROC) analysis showed that the optimum cut-off for *P*-wave area in V3 was 1.45 Ashman units with an area under the curve of 0.65 for classifying new onset AF.

A decision tree learning (DTL) model was then employed to generate the decision rules based on only individual variables as shown in **Figure 3A**, while the decision rules generated by DTL model based on both individual and non-linear interaction items are shown in **Figure 3B**. In each model, 80% of the sample ($n = 47$) were randomly selected and the remaining 20% ($n = 12$) were used for validation. For the decision tree without interacting variables, *P*-wave area in V3 is the first predictor generated with a Gini index of 0.496 (**Figure 3A**). In addition, the interaction between age and left atrial diameter is the first variable in the generated decision rule with a Gini index of 0.495, while *P*-wave area in V3 as the second most important variable with Gini index 0.278 (**Figure 3B**). Both decision trees generated by machine learning can be used as an efficient tool for accurate risk stratification in mitral stenosis. The decision rule incorporating interactions between variables is more accurate.

To determine the out-of-sample prediction ability of the model, five-fold cross-validation was adopted. The evaluation metrics evaluated are accuracy, precision, recall, and F-measure. Comparisons of the prediction performance of logistic regression and DTL are shown in **Supplementary Table 4**. DTL with individual and non-linear variables outperforms the logistic regression for predicting incident AF in mitral stenosis. The non-linear variable formed by *P*-wave area in V3 and age was significantly predictive for classifying new onset AF outcome in mitral stenosis. We can also observe that several non-linear variables are more predictive than individual variables, indicating the importance of considering the non-linear patterns in clinical characteristics to improve the performance of predicting new onset AF.

DISCUSSION

The major findings of this study are that (i) a high proportion of patients with mitral stenosis had IAB, (ii) age and *P*-wave

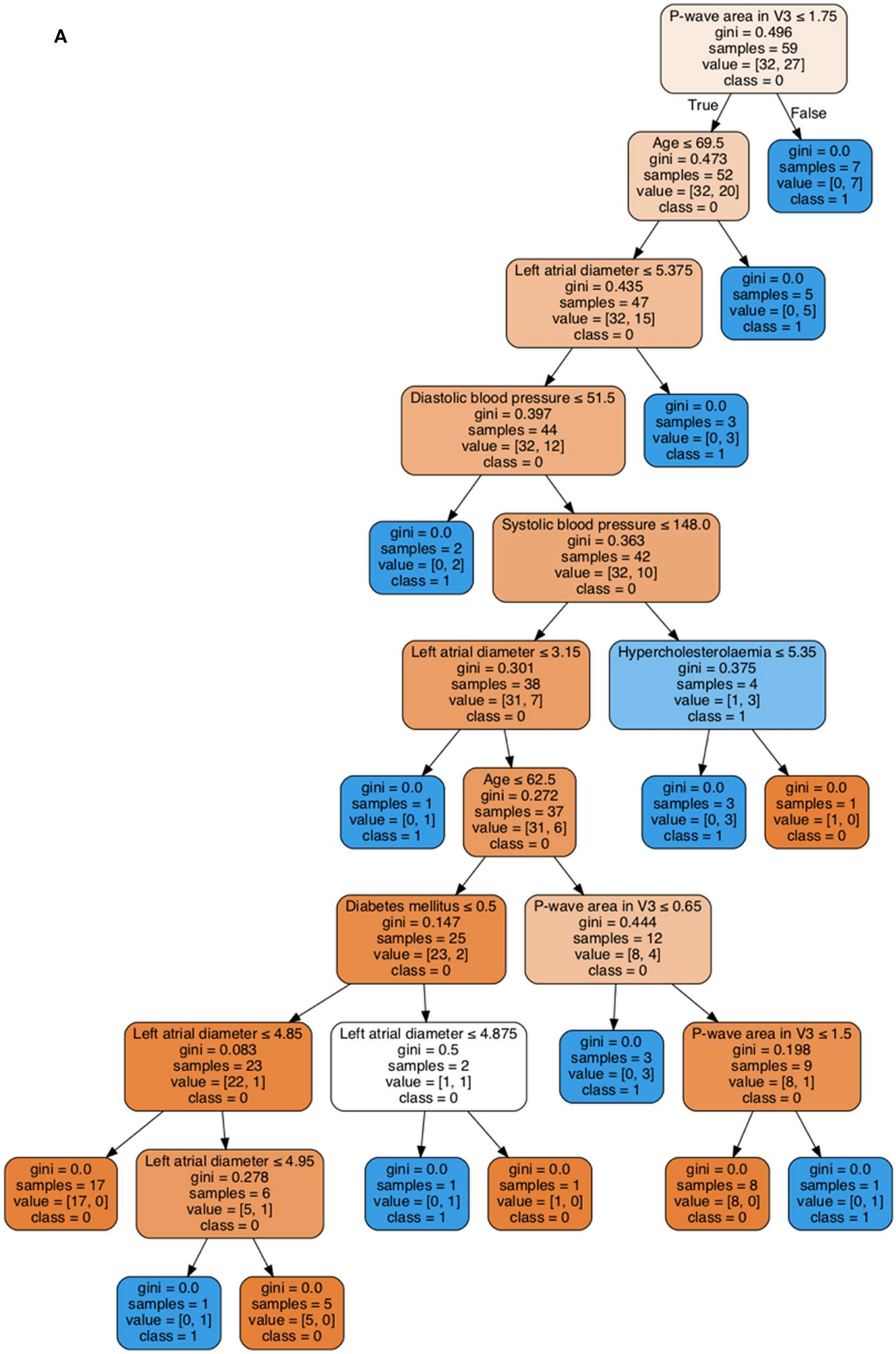


FIGURE 3 | Continued

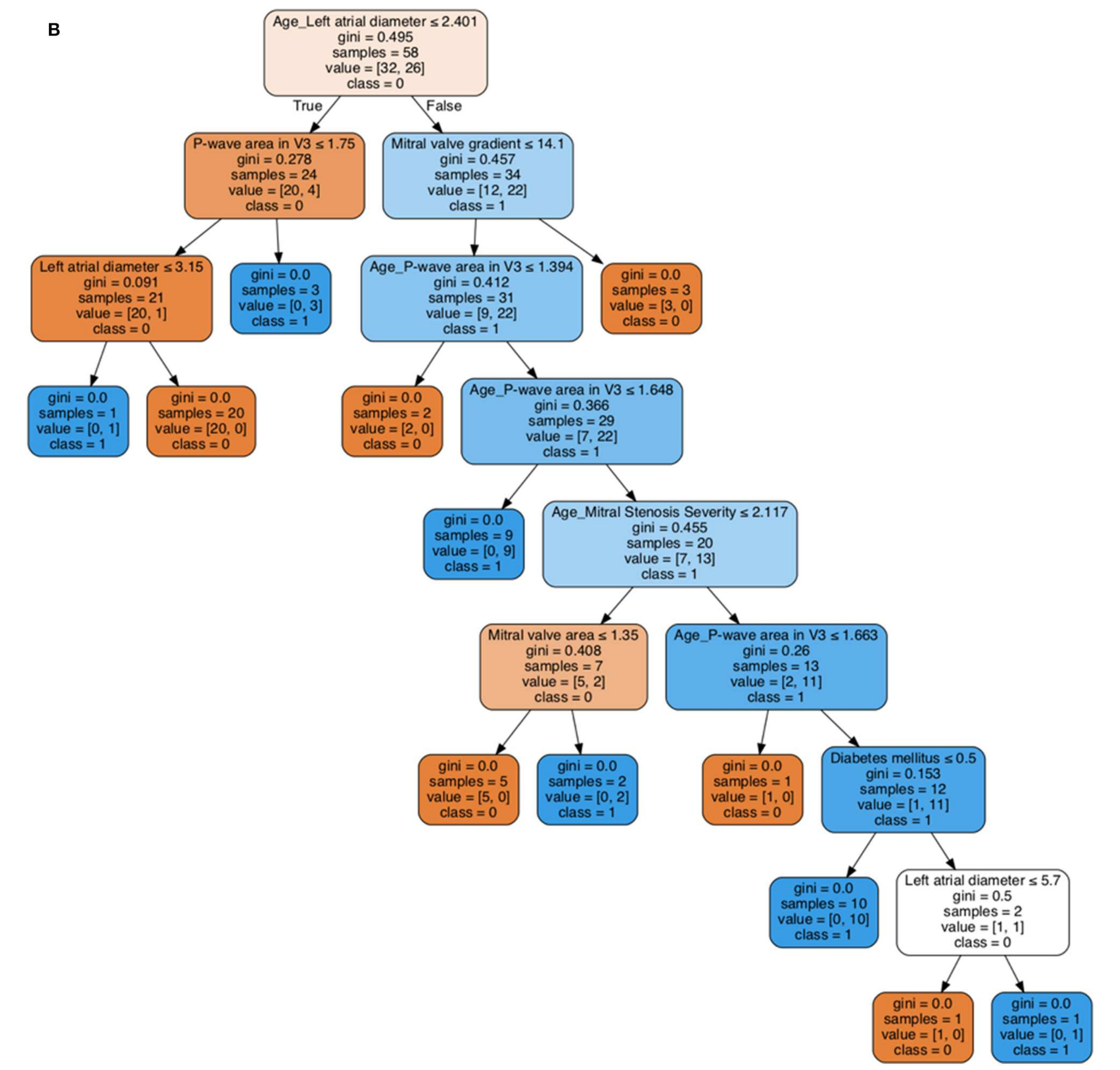


FIGURE 3 | Visualization of decision tree learning with individual variables **(A)**. Visualization of decision tree learning with both individual and non-linear interaction variables **(B)**.

area in V3 predicted new onset AF, and (iii) a stepwise improvement in the predictive performance after incorporation of interaction variables and machine learning using a decision tree approach.

Prior studies have investigated alterations in *P*-wave morphologies and indices in non-valvular atrial fibrillation. Whilst some reports have dealt with the relationship between mitral stenosis and *P*-wave indices, few studies have examined whether these indices can predict new onset AF. Atrial electrophysiology in mitral stenosis is abnormal due to a complex process of electrophysiological remodeling. IAB is the conduction delay along the Bachmann's bundle between left and right atria, diagnosed by its characteristic ECG pattern (Agarwal et al., 2003; Bayes de Luna et al., 2012). Partial IAB and advanced IAB are defined as PWD ≥ 120 ms in the presence and absence of biphasic *P*-waves in the inferior leads. The association between IAB and supraventricular tachyarrhythmias, especially AF, is known as Bayés syndrome. Mitral stenosis is a major risk factor for AF through atrial dilatation with progressive structural remodeling and interstitial fibrosis, predisposing to re-entrant activity within the atrium (Markides and Schilling, 2003; O'neal et al., 2016). In addition to IAB, abnormal atrial electrophysiology can be detected by alterations in *P*-wave morphology on the electrocardiogram (ECG), including *P*-wave dispersion and abnormal *P*-wave terminal force in V1 (PTFV1) (Yamada et al., 1999; Dogan et al., 2004; Wong et al., 2004; Koide et al., 2008; Tsioufis et al., 2010; Yoshizawa et al., 2014). A previous study involving 30 mitral stenosis patients found that maximum *P*-wave duration and *P*-wave dispersion were significantly higher than patients without mitral stenosis (Guntekin et al., 2008). In the same previous study, baseline maximum and minimum PWDs and *P*-wave dispersion all correlated with mitral valve area and mean mitral gradient (Guntekin et al., 2008). Another study involving a prospective follow-up of 116 mitral stenosis patients similarly reported these associations, and additionally correlated these ECG parameters with increased pulmonary artery pressure, and a poor NYHA class (Yuce et al., 2011). Moreover, the extent to which atrial electrical abnormalities can predict incident AF remains less explored in these previous studies. Whilst previous studies have demonstrated the predictive value of various *P*-wave indices for incident AF or progression from paroxysmal to persistent AF (Koide et al., 2008), to date there are no studies specifically on their values in mitral stenosis. In our cohort, we utilized automated ECG measurements and found that *P*-wave area significantly predicted new onset AF.

In clinical practice, it is often useful to use cut-off values to identify categorize whether a patient is at high or low risk of adverse events. For example, a previous study identified that PWDs longer than 150 ms predicted AF recurrence after pulmonary isolation procedures (Jadidi et al., 2018). It should be noted that in the prior study, the ECG signals were amplified to 0.2–0.25 mV/cm before manual measurements were made, and the methodology therefore differs from that used here. In a cohort of patients with left atrial overload, area of the initial portion of the *P*-wave $\geq$ 65 μV.ms was associated with a four-fold increased risk of developing future AF (Ishida et al., 2010). In our study, the optimum cut-off for *P*-wave area in V3 was 1.45 Ashman units, with an area under the ROC curve of 0.65. Our novelty lies with the demonstrations that *P*-wave area can be used to predict incident AF in mitral stenosis and the use of machine learning approaches significantly improve outcome classification.

LIMITATIONS

Some limitations should be recognized. Firstly, this was a single center study with a small sample size and retrospective analyses. Some inherent bias could affect the results. However, electronic health records in Hong Kong are comprehensive with accessible information across different hospitals within the public system and multiple follow-ups per year in the outpatient and inpatient settings. Secondly, the effects of drugs were not explored in the current study. Thirdly, left atrial diameter was the only available metric on atrial dimensions, as this was the only variable described in the echocardiography reports. Future work could explore (i) whether left atrial area, volume or volume index can predict incident AF and (ii) the potential effects on atrial reverse remodeling by drugs. Our main conclusion that *P*-wave area predicts incident AF needs to be validated by larger prospective studies.

CONCLUSION

Atrial electrophysiological alterations in mitral stenosis can detected on the electrocardiogram. Age, systolic blood pressure, and mean PWD predicted new onset AF. A decision tree learning model significantly improve outcome prediction.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Joint Chinese University of Hong Kong—New Territories East Cluster Clinical Research Ethics Committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

GT: conception of study and literature search, figures, study design, data collection, data analysis, data contribution, manuscript drafting, and critical revision of manuscript. IL: literature search, data analysis, data contribution, manuscript drafting, and critical revision of manuscript. JZ and KLi: data analysis, manuscript drafting, and critical revision of manuscript. SL, YL, KLe, and TL: data interpretation, and critical revision of manuscript. AB: literature search, data analysis, critical revision of manuscript, and study supervision. QZ: literature search, figures, study design, data analysis, manuscript drafting, critical revision of manuscript, and study supervision.

FUNDING

This study was partially supported by National Natural Science Foundation of China (NSFC: 71972164 and 71972164) and the Health and Medical Research Fund (HMRF: 16171991).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00479/full#supplementary-material

REFERENCES

Agarwal, Y. K., Aronow, W. S., Levy, J. A., and Spodick, D. H. (2003). Association of interatrial block with development of atrial fibrillation. *Am. J. Cardiol.* 91:882. doi: 10.1016/S0002-9149(03)00027-4

Ariyarajah, V., Puri, P., Apiyasawat, S., and Spodick, D. H. (2007). Interatrial block: a novel risk factor for embolic stroke? *Ann. Noninvasive Electrocardiol.* 12, 15–20. doi: 10.1111/j.1542-474X.2007.00133.x

Bayes de Luna, A., Platonov, P., Cosio, F. G., Cygankiewicz, I., Pastore, C., Baranowski, R., et al. (2012). Interatrial blocks. A separate entity from left atrial enlargement: a consensus report. *J. Electrocardiol.* 45, 445–451. doi: 10.1016/j.jelectrocard.2012.06.029

Budeus, M., Hennersdorf, M., Perings, C., Wieneke, H., Erbel, R., and Sack, S. (2005). Prediction of the recurrence of atrial fibrillation after successful cardioversion with *P* wave signal-averaged ECG. *Ann. Noninvasive Electrocardiol.* 10, 414–419. doi: 10.1111/j.1542-474X.2005.00059.x

Caldwell, J., Koppikar, S., Barake, W., Redfearn, D., Michael, K., Simpson, C., et al. (2014). Prolonged *P*-wave duration is associated with atrial fibrillation recurrence after successful pulmonary vein isolation for paroxysmal atrial fibrillation. *J. Interv. Card. Electrophysiol.* 39, 131–138. doi: 10.1007/s10840-013-9851-1

Dogan, A., Avsar, A., and Ozturk, M. (2004). *P*-wave dispersion for predicting maintenance of sinus rhythm after cardioversion of atrial fibrillation. *Am. J. Cardiol.* 93, 368–371. doi: 10.1016/j.amjcard.2003.09.064

Guntekin, U., Gunes, Y., Tuncer, M., Gunes, A., Sahin, M., and Simsek, H. (2008). Long-term follow-up of P-wave duration and dispersion in patients with mitral stenosis. *Pacing Clin. Electrophysiol.* 31, 1620–1624. doi: 10.1111/j.1540-8159.2008.01235.x

He, J., Tse, G., Korantzopoulos, P., Letsas, K. P., Ali-Hasan-Al-Saegh, S., Kamel, H., et al. (2017). *P*-wave indices and risk of ischemic stroke: a systematic review and meta-analysis. *Stroke* 48, 2066–2072. doi: 10.1161/STROKEAHA.117.017293

Ishida, K., Hayashi, H., Miyamoto, A., Sugimoto, Y., Ito, M., Murakami, Y., et al. (2010). *P* wave and the development of atrial fibrillation. *Heart Rhythm* 7, 289–294. doi: 10.1016/j.hrthm.2009.11.012

Jadidi, A., Muller-Edenborn, B., Chen, J., Keyl, C., Weber, R., Allgeier, J., et al. (2018). The duration of the amplified sinus-*P*-wave identifies presence of left atrial low voltage substrate and predicts outcome after pulmonary vein isolation in patients with persistent atrial fibrillation. *JACC Clin. Electrophysiol.* 4, 531–543. doi: 10.1016/j.jacep.2017.12.001

Koide, Y., Yotsukura, M., Ando, H., Aoki, S., Suzuki, T., Sakata, K., et al. (2008). Usefulness of P-wave dispersion in standard twelve-lead electrocardiography to predict transition from paroxysmal to persistent atrial fibrillation. *Am. J. Cardiol.* 102, 573–577. doi: 10.1016/j.amjcard.2008.04.065

Lankveld, T., Zeemering, S., Scherr, D., Kuklik, P., Hoffmann, B. A., Willems, S., et al. (2016). Atrial fibrillation complexity parameters derived from surface ECGs predict procedural outcome and long-term follow-up of stepwise catheter ablation for atrial fibrillation. *Circ. Arrhythm. Electrophysiol.* 9:e003354. doi: 10.1161/CIRCEP.115.003354

Magnani, J. W., Gorodeski, E. Z., Johnson, V. M., Sullivan, L. M., Hamburg, N. M., Benjamin, E. J., et al. (2011). P wave duration is associated with cardiovascular and all-cause mortality outcomes: the National health and nutrition examination survey. *Heart Rhythm* 8, 93–100. doi: 10.1016/j.hrthm.2010.09.020

Markides, V., and Schilling, R. J. (2003). Atrial fibrillation: classification, pathophysiology, mechanisms and drug treatment. *Heart* 89, 939–943. doi: 10.1136/heart.89.8.939

Martin Garcia, A., Jimenez-Candil, J., Hernandez, J., Martin Garcia, A., Martin Herrero, F., and Martin Luengo, C. (2012). P wave morphology and recurrence after cardioversion of lone atrial fibrillation. *Rev. Esp. Cardiol.* 65, 289–290. doi: 10.1016/j.rec.2011.04.020

Morris, J. J. Jr, Estes, E. H. Jr, Whalen, R. E., Thompson, H. K. Jr, and Mcintosh, H. D. (1964). P-wave analysis in valvular heart disease. *Circulation* 29, 242–252. doi: 10.1161/01.CIR.29.2.242

O'neal, W. T., Venkatesh, S., Broughton, S. T., Griffin, W. F., and Soliman, E. Z. (2016). Biomarkers and the prediction of atrial fibrillation: state of the art. *Vasc. Health Risk Manag.* 12, 297–303. doi: 10.2147/VHRM.S75537

Rezaian, G. R., Rezaian, S., Liaghat, L., and Zare, N. (2007). P-wave dispersion in patients with rheumatic mitral stenosis. *Int. J. Angiol.* 16, 20–23. doi: 10.1055/s-0031-1278239

Rutkowski, L., Jaworski, M., Pietruczuk, L., and Duda, P. (2014). The CART decision tree for mining data streams. *Inf. Sci.* 266, 1–15. doi: 10.1016/j.ins.2013.12.060

Tse, G., Lai, E. T., Yeo, J. M., and Yan, B. P. (2016). Electrophysiological mechanisms of bayes syndrome: insights from clinical and mouse studies. *Front. Physiol.* 7:188. doi: 10.3389/fphys.2016.00188

Tsioufis, C., Syrseloudis, D., Hatziyianni, A., Tzamou, V., Andrikou, I., Tolis, P., et al. (2010). Relationships of CRP and P wave dispersion with atrial fibrillation in hypertensive subjects. *Am. J. Hypertens.* 23, 202–207. doi: 10.1038/ajh.2009.231

Wong, T., Davlouros Periklis, A., Li, W., Millington-Sanders, C., Francis Darrel, P., and Gatzoulis Michael, A. (2004). Mechano-electrical interaction late after fontan operation. *Circulation* 109, 2319–2325. doi: 10.1161/01.CIR.0000129766.18065.DC

Yamada, T., Fukunami, M., Shimonagata, T., Kumagai, K., Sanada, S., Ogita, H., et al. (1999). Dispersion of signal-averaged P wave duration on precordial body surface in patients with paroxysmal atrial fibrillation. *Eur. Heart J.* 20, 211–220. doi: 10.1053/euhj.1998.1281

Yoshizawa, T., Niwano, S., Niwano, H., Igarashi, T., Fujiishi, T., Ishizue, N., et al. (2014). Prediction of new onset atrial fibrillation through P wave analysis in 12 lead ECG. *Int. Heart J.* 55, 422–427. doi: 10.1536/ihj.14-052

Yuce, M., Davutoglu, V., Akkoyun, C., Kizilkan, N., Ercan, S., Akcay, M., et al. (2011). Interatrial block and P-terminal force: a reflection of mitral stenosis severity on electrocardiography. *J. Heart Valve Dis.* 20, 619–623.

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fbioe.2020.00720

Detection and Severity Assessment of Peripheral Occlusive Artery Disease via Deep Learning Analysis of Arterial Pulse Waveforms: Proof-of-Concept and Potential Challenges

Sooho Kim[1], *Jin-Oh Hahn*[2]* *and Byeng Dong Youn*[1,3]*

[1] *Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul, South Korea,* [2] *Department of Mechanical Engineering, University of Maryland, College Park, MD, United States,* [3] *OnePredict, Inc., Seoul, South Korea*

OPEN ACCESS

Edited by:
Wenbing Zhao,
Cleveland State University,
United States

Reviewed by:
Xishuang Dong,
Prairie View A&M University,
United States
Chenxi Huang,
Xiamen University, China

***Correspondence:**
Jin-Oh Hahn
jhahn12@umd.edu
Byeng Dong Youn
bdyoun@snu.ac.kr

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

Received: *10 February 2020*
Accepted: *08 June 2020*
Published: *30 June 2020*

Citation:
Kim S, Hahn J-O and Youn BD (2020) Detection and Severity Assessment of Peripheral Occlusive Artery Disease via Deep Learning Analysis of Arterial Pulse Waveforms: Proof-of-Concept and Potential Challenges. Front. Bioeng. Biotechnol. 8:720. doi: 10.3389/fbioe.2020.00720

Toward the ultimate goal of affordable and non-invasive screening of peripheral occlusive artery disease (PAD), the objective of this work is to investigate the potential of deep learning-based arterial pulse waveform analysis in detecting and assessing the severity of PAD. Using an established transmission line model of arterial hemodynamics, a large number of virtual patients associated with PAD of a wide range of severity and the corresponding arterial pulse waveform data were created. A deep convolutional neural network capable of detecting and assessing the severity of PAD based on the analysis of brachial and ankle arterial pulse waveforms was constructed, evaluated for efficacy, and compared with the state-of-the-art ankle-brachial index (ABI) using the virtual patients. The results suggested that deep learning may diagnose PAD more accurately and robustly than ABI. In sum, this work demonstrates the initial proof-of-concept of deep learning-based arterial pulse waveform analysis for affordable and convenient PAD screening as well as presents challenges that must be addressed for real-world clinical applications.

Keywords: peripheral artery disease, cardiovascular disease, deep learning, machine learning, pulse wave analysis, arterial hemodynamics, ankle-brachial index, convolutional neural network

INTRODUCTION

Peripheral artery occlusive disease (PAD) is a highly prevalent vascular disease associated with high morbidity and mortality risks. It was estimated that >8 million and >200 million people were suffering from PAD in the United States (in 2000) (Allison et al., 2007) and globally (in 2010) (Fowkes et al., 2013), and the number of PAD patients is projected to sharply increase with societal aging. It makes a significant adverse impact on morbidity and quality of life, and also carries significant mortality implications as a powerful predictor of coronary artery disease and cerebrovascular disease (Golomb et al., 2006). Nonetheless, PAD is underdiagnosed with low primary care awareness (Hirsch et al., 2001).

In clinical practice today, PAD diagnosis necessitates angiography techniques (Guthaner et al., 1983; Romano et al., 2004; Cavallo et al., 2019). These techniques are not ideally suited to affordable and convenient PAD detection and severity assessment. The current gold standard is the digital subtraction angiography, which is an invasive technique. Other non-invasive imaging-based angiography techniques including the computed tomography angiography and magnetic resonance angiography require X-ray radiation and expensive equipment not appropriate for affordable settings. The ankle-brachial index (ABI) is a relatively low-cost technique and is widely used for PAD screening. However, it is often criticized for its limited accuracy and robustness in diagnosing PAD (Nelson et al., 2012).

Machine learning (ML) is increasingly exploited in cardiovascular disease (CVD) detection and prognosis. In particular, ML has exhibited promising efficacy in heart disease detection and prediction (Dogan et al., 2018; Abdar et al., 2019; Vallée et al., 2019) as well as CVD risk and CV death prognosis (Ambale-Venkatesh et al., 2017; Steele et al., 2018; Alaa et al., 2019). Recent reports increasingly exploit deep learning (DL) to capitalize on its ability to automatically select characteristic features, especially in conjunction with medical imaging techniques (Abdolmanafi et al., 2018; Poplin et al., 2018; Zhang et al., 2019). In contrast to the large body of existing work on ML-based CVD detection and CV mortality prediction, relatively small number of work on ML applications to PAD is available, including detection and mortality prognosis using electronic health record as well as genomic and imaging data (Ross et al., 2016; Arruda-Olson et al., 2018).

The analysis of arterial pulse waveforms [called hereafter the pulse waveform analysis (PWA)] may play a complementary role to ML in PAD diagnosis. In fact, our prior work shows that model-based PWA has the potential to estimate CV risk predictors (Ghasemi et al., 2018) and diagnose CVD (Ebrahimi Nejad et al., 2017) using diametric arterial pulses. A recent work illustrated the theoretical feasibility of PAD diagnosis (including detection, localization, and severity assessment) using a hybrid model- and ML-based analysis of central aortic and peripheral arterial pulses (Xiao et al., 2016a). A practical advantage of PWA is that it may be relevant to affordable PAD screening and diagnosis with convenient arterial pulse measurements at the extremity locations (e.g., arm and ankle).

Despite the complementary value of DL and PWA in advancing the diagnosis of PAD (and even other CVDs), the fusion of DL and PWA for PAD diagnosis has never been pursued to the best of our knowledge. In fact, the state-of-the-art of DL-based PWA appears to be limited to rudimentary classification of CV health state (e.g., hypertension, atherosclerosis, and diabetes mellitus) (Li et al., 2019). Hence, DL-PWA fusion is a novel conceptual idea worthy of pursuit in the context of CVD diagnosis (including PAD).

Toward the long-term goal of affordable and non-invasive PAD screening and diagnosis, the objective of this work is to investigate the potential of DL-based arterial PWA in detecting and assessing the severity of PAD. Using an established transmission line (TL) model of arterial hemodynamics, a large number of virtual patients associated with PAD of a wide range of severity and the corresponding arterial pulse waveform data were created. A deep convolutional neural network (CNN) capable of detecting and assessing the severity of PAD based on the analysis of brachial and ankle arterial pulse waveforms was constructed, evaluated for efficacy, and compared with the state-of-the-art ABI using the virtual patients.

This paper is organized as follows. Section "Materials and Methods" presents a multi-branch TL model of arterial hemodynamics used in this work, creation of virtual PAD patients together with the corresponding arterial pulse waveforms to investigate DL-based PWA for PAD diagnosis, a DL-based PWA approach based on the CNN for PAD detection and severity assessment, and data analysis methods to evaluate the efficacy of the DL-based PWA approach. Section "Results" presents results, which are discussed in section "Discussion." Section "Conclusion" concludes this work with future directions.

MATERIALS AND METHODS

Transmission Line Model of Arterial Hemodynamics

We used a multi-branch TL model of arterial hemodynamics developed in a prior work (**Figure 1**; He et al., 2012). In brief, the model is composed of 55 TLs, each of which represents an arterial segment characterized by segment-specific viscous, elastic, and inertial properties. In each TL, the propagation of arterial blood pressure (BP) and flow (BF) waves is dictated by the propagation and reflection constants as well as the arterial length:

$$\begin{aligned} p_O &= p_I(1+\Gamma)\Big/\left(e^{\gamma l}+\Gamma e^{-\gamma l}\right) \\ q_O &= q_I(1-\Gamma)\Big/\left(e^{\gamma l}-\Gamma e^{-\gamma l}\right) \end{aligned} \tag{1}$$

where p_I and p_O are BP waves at the inlet and outlet of the artery, q_I and q_O are BF waves at the inlet and outlet of the artery, γ is the propagation constant, Γ is the reflection constant, and l is the arterial length. BP and BF waves at the inlet of the artery are related by the input impedance of the arterial segment:

$$p_I = q_I Z_I = q_I Z_C\left(e^{\gamma l}+\Gamma e^{-\gamma l}\right)\Big/\left(e^{\gamma l}-\Gamma e^{-\gamma l}\right) \tag{2}$$

where Z_I and Z_C are the input impedance and characteristic impedance of the artery, respectively. If an arterial segment is terminated by a bifurcation, its load impedance is given by the parallel connection of the input impedances associated with the two descendent arteries. If an arterial segment is connected to a single descendent artery, its load impedance is given simply by the input impedance associated with the descendent artery. If an arterial segment itself is a terminal artery connected to a peripheral load, its load impedance is given by the impedance associated with the load. Full details of the TL model is provided in He et al. (2012). This model was validated with physiological data and the results of other studies, and was

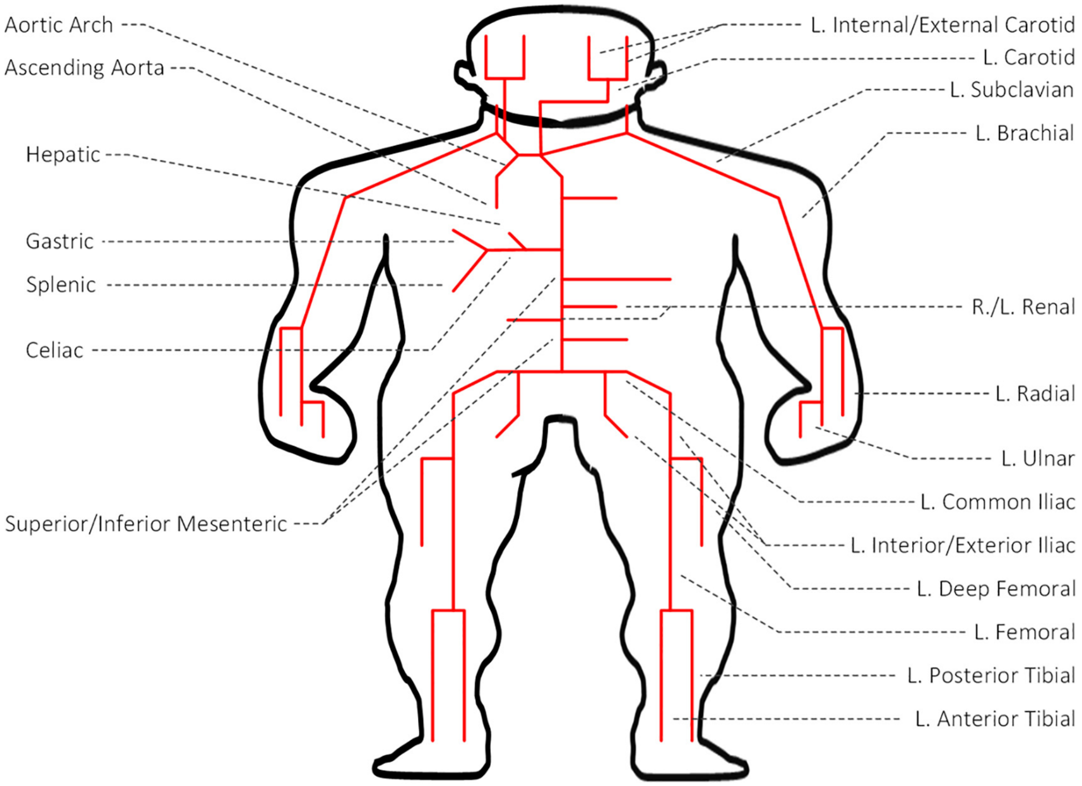


FIGURE 1 | Transmission line (TL) model of arterial hemodynamics consisting of 55 TLs, each of which represents an arterial segment characterized by segment-specific viscous, elastic, and inertial properties.

used in the study of arterial stenosis and arterial viscoelasticity (Xiao et al., 2016a,b, 2017).

Creation of Virtual PAD Patients

We created a large number of virtual patients to investigate the potential and challenges in DL-based PWA for PAD diagnosis using the aforementioned multi-branch TL model. To create realistic virtual patients, we considered three layers of variabilities: inter-individual, intra-individual, and PAD severity. First, we considered the inter-individual variability in the arterial hemodynamics associated with the virtual patients by widely varying five anatomical and physiological parameters in the multi-branch TL model: arterial length, diameter, and thickness, arterial elasticity, and peripheral load resistance. These parameters were varied up to ±20% around the nominal values reported in He et al. (2012) in an increment of 10%, which resulted in a total of 5^5 = 3125 virtual patients associated with 5^5 distinct arterial hemodynamic properties. Second, we considered the PAD severity variability in each virtual patent by widely varying the degree of the artery occlusion in the multi-branch TL model. In this exploratory work, we limited our focus to PAD occurring in the abdominal aorta, which is one of the most common PAD sites. In each of the 3125 virtual patients, we included PAD by varying diameter associated with the abdominal aorta. We considered PAD severity of 0–80% in an increment of 10% for training and validation datasets and in an increment of 1% for test dataset, where severity is measured as the degree of artery area occlusion (0% implies no occlusion while 100% implies complete occlusion). This resulted in a total of 3125 × 9 = 28,125 virtual patients, associated with distinct arterial hemodynamics and PAD, as the basis to construct training and validation datasets and 3125 × 81 = 253,125 virtual patients, associated with distinct arterial hemodynamics and PAD, as the basis to construct test dataset. Third, we considered the intra-individual variability in the arterial hemodynamics in each virtual patient to account for the uncertainty due to model imperfection as well as random anatomical and physiological variations. We assumed that the five anatomical and physiological parameters in the multi-branch TL model used to account for the inter-individual arterial hemodynamic variability have log-normal distributions around the individual-specific values as mean values with coefficient of variation of 0.01 in each virtual patient. Finally, we constructed training and validation datasets by sampling 100 and 10 times from each of the 28,125 virtual patients equipped with random anatomical and physiological variations, and likewise constructed test dataset by sampling 10 times from each of the 253,125 virtual patients equipped with random anatomical and physiological variations. Then, we created arterial BP and BF waveforms associated with each of these samples by inputting a representative heart blood flow waveform used in He et al. (2012; **Figure 2**) to the multi-branch TL model characterized by the sample-specific anatomical and physiological parameters (including PAD severity). In this way, training and validation datasets were composed of 2,812,500 and 281,250 arterial BP and BF waveform data samples corresponding to 28,125 virtual patients, while test dataset was composed of 2,531,250 arterial BP and BF waveform data samples corresponding to 253,125 virtual patients.

PAD Diagnosis via Deep Learning-Based Pulse Waveform Analysis

We developed our DL-based PWA approach to PAD diagnosis using the training and validation datasets constructed in section "Creation of Virtual PAD Patients." Specifically, we constructed a deep CNN that can predict PAD severity by the analysis of arterial pulse waveforms. We in particular selected brachial and ankle BP waveforms as inputs to our deep CNN in order to make our approach compatible to the state-of-the-art ABI technique, so that (i) our approach and ABI can be directly compared and (ii) the potential for real-world application of our approach is maximized. Details follow.

Our deep CNN was built upon the AlexNet (Krizhevsky et al., 2012; Han et al., 2017; Wang et al., 2019), which was regarded as appropriate in dealing with 1-D arterial pulse waveforms associated with less complexity than 2-D images relative to other deeper CNN architectures such as ResNet (He et al., 2016) and DenseNet (Huang et al., 2017). To obviate extensive tuning of hyper-parameters, we adopted the original AlexNet architecture (five convolution layers and three fully connected layers), but with modest modifications (**Figure 3**). First, we employed the LeakyReLU as the activation function for the entire network to promote stable convergence in the training phase (Goodfellow et al., 2016). Second, we employed batch normalization in all the convolution layers to promote stable back propagation of gradient as well as regularization (Goodfellow et al., 2016). Third, we reduced the size of the fully connected layer to 64 to match it to the number of latent features outputted by the last convolution layer in our CNN. Using the network architecture thus specified, we constructed the deep CNN in such a way that brachial and ankle arterial pulses are convoluted independently (**Figure 3**). For this purpose, brachial and ankle arterial pulses undergo channel-wise concatenation so that these arterial pulses can be convoluted separately from each other by a shared kernel in the convolution layer. In this way, discriminative features of PAD severity embedded in the brachial and ankle arterial pulses can be extracted independently while computational efficiency can be gained with the use of shared kernels. In addition, mutual interactions between the discriminative features associated with the two arterial pulses can be exploited in the fully connected layer of the network.

To train the deep CNN, we used NVIDIA Titan Xp GPU and PyTorch libraries. We used the mean squared error loss between the true vs. model-predicted PAD severity as the cost function. We used the ADAM optimization ($\alpha = 0.9$, $\beta = 0.999$) with initial learning rate of 0.0002. To assess the robustness of the deep CNN, we examined the sensitivity of the cost function with respect to the local perturbations in the hyper-parameters including the number (increased by 1.5 and 2 times) and size (increased by 1 and 2) of kernels in the convolution layer. Note that the deep CNN thus trained with the above regression cost can be used to both detect and assess the severity of PAD. In particular, it can be used to detect PAD simply by labeling PAD in terms of PAD severity (i.e., classifying a subject as PAD patient if the subject's PAD severity exceeds a pre-specified PAD severity threshold).

Evaluation

We evaluated our DL-based PWA approach to PAD diagnosis and compared its efficacy with the state-of-the-art ABI technique, in terms of PAD detection and severity assessment efficacy, using the test dataset constructed in section "Creation of Virtual PAD Patients." Details follow.

First, we evaluated our approach for its PAD detection performance. We considered a range of PAD severity threshold levels in labeling healthy subjects and PAD patients (10–70%, in an increment of 10%). For each PAD labeling threshold level, we randomly selected 2000 virtual patients from test dataset (consisting of 253,125 virtual patients; see section "Creation of Virtual PAD Patients") so that the selected patients include equal number of healthy subjects and PAD patients (i.e., 1000 healthy subjects and 1000 PAD patients; for example, in case of 40% PAD severity threshold for labeling, 1000 virtual patients with <40% PAD severity were randomly chosen to form healthy subjects while 1000 virtual patients with ≥40% PAD severity were randomly chosen to form PAD patients). Then, we evaluated our approach and ABI technique using the 20,000 arterial BP and BF waveform data of these 2000 virtual patients (see section "Creation of Virtual PAD Patients") by (i) classifying each arterial BP and BF waveform data sample into healthy or PAD category based on the PAD severity predicted by the deep CNN when the brachial and ankle BP waveforms in the sample were inputted and the ABI value computed from the waveforms, (ii) aggregating the classification results across all the 20,000 data samples associated with all the 2000 virtual patients, and (iii) computing the sensitivity and specificity as well as the accuracy of PAD detection. In the context of PAD detection, sensitivity was defined as the proportion of the 10,000 PAD patient samples which were actually detected as such (with the PAD severity predicted to be higher than the PAD labeling threshold), while specificity was defined as the proportion of the 10,000 healthy subject samples which were actually detected as such (with the PAD severity predicted to be lower than the PAD labeling threshold). Accuracy was defined as the proportion of the 20,000 test samples whose labels were classified correctly.

Second, we evaluated our approach for its PAD severity assessment performance. We randomly selected 2,000 virtual patients from test dataset (consisting of 253,125 virtual patients; see section "Creation of Virtual PAD Patients") so that the selected patients are distributed uniformly across all the PAD severity levels (1–80% in an increment of 1%, which amounts to 25 virtual patients per PAD severity level). Then, we evaluated our approach and ABI technique using the 20,000 arterial BP and BF waveform data samples of these 2000 virtual patients (see section "Creation of Virtual PAD Patients"), in terms of the Bland-Altman statistics between the true PAD severity vs. the PAD severity predicted by our deep CNN and ABI. To map

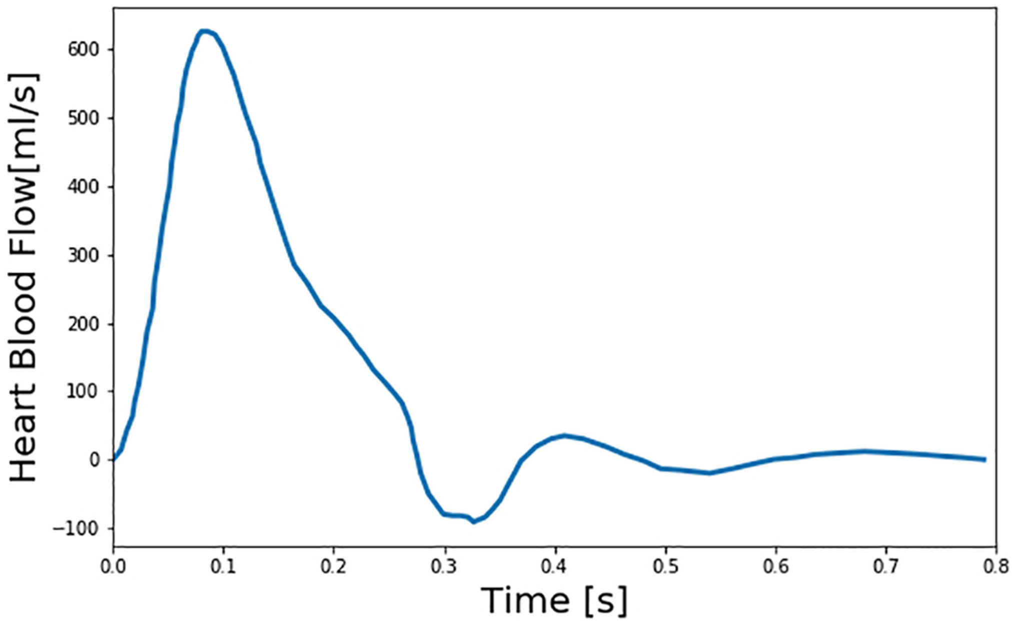


FIGURE 2 | Representative heart blood flow waveform used as input to the multi-branch transmission line (TL) model of arterial hemodynamics associated with virtual patients.

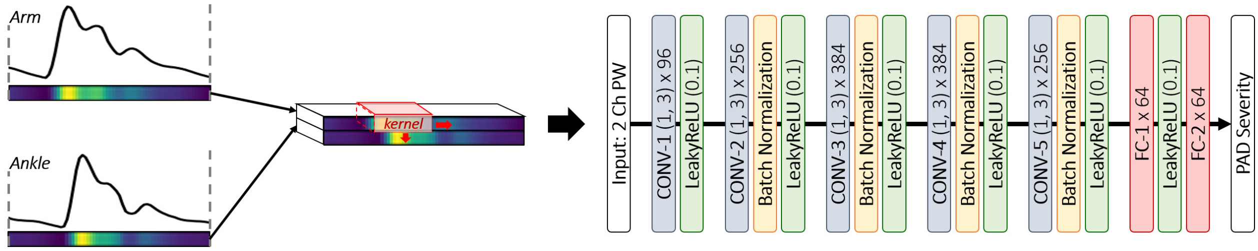


FIGURE 3 | Deep convolutional neural network (CNN) architecture for PAD diagnosis via deep learning-based arterial pulse waveform analysis. CONV-*n* (*h*, *l*) × *k*: *n*th convolution layer with height *h*, length *l* and the number of kernel *k*. LeakyReLU (*a*): LeakyReLU activation with slope *a* on negative inputs. FC-*n* × *m*: *n*th fully connected layer with the number of node *m*.

ABI value to PAD severity, we pre-calibrated the ABI values to the corresponding PAD severity level based on a polynomial regression model relating ABI to PAD severity (which was obtained from the nominal virtual patient characterized by the nominal anatomical and physiological parameter values). Third, we analyzed the latent feature space associated with our deep CNN using the t-distributed stochastic neighbor embedding (t-SNE) algorithm. This analysis was conducted to examine the presence of a smooth manifold relating the latent features to PAD severity. We applied t-SNE to visualize the input space and the space of latent features at the last convolution layer into 2-dimensional space. Then, we investigated the distributions of the input and latent features in the 2-dimensional space for a connected manifold in the direction of PAD severity. Fourth, we analyzed our deep CNN using the gradient-weighted class activation mapping (GradCAM) algorithm (Selvaraju et al., 2017) to interpret the discriminative input features exploited by our deep CNN in predicting PAD severity. We applied GradCAM to visualize the discriminative features (i.e., regions) in the brachial and ankle arterial BP waveforms which largely contributed in predicting PAD severity. Then, we assessed the physiological relevance of the input features exploited by the deep CNN in diagnosing PAD by comparing these discriminative features and the available clinical knowledge on the relationship between PAD severity and arterial pulse waveforms.

To derive a robust estimate of detection and diagnosis performance, we repeated the above evaluation 10 times and reported the average values of the sensitivity, specificity, and accuracy as well as the Bland-Altman statistics.

RESULTS

Figure 4 presents brachial and ankle BP waveforms corresponding to (a) nominal virtual patient, (b) nominal virtual patient with intra-individual variability, and (c) all the

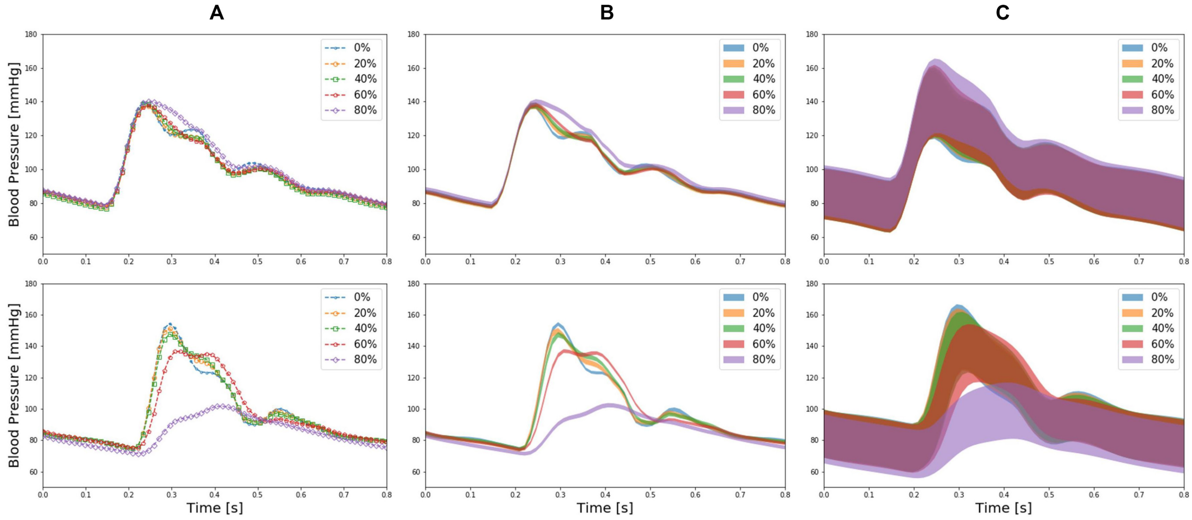


FIGURE 4 | Brachial (upper panel) and ankle (lower panel) blood pressure (BP) waveforms corresponding to **(A)** nominal virtual patient (i.e., virtual patient with nominal anatomical and physiological parameter values), **(B)** nominal virtual patient with intra-individual variability, and **(C)** all the virtual patients with inter- and intra-individual variability in the test dataset, all associated with varying PAD severity levels.

virtual patients with inter- and intra-individual variability in the test dataset, all associated with varying PAD severity levels. **Table 1** summarizes the PAD detection performance of our approach and ABI (measured in terms of detection sensitivity, specificity, and accuracy), both corresponding to varying PAD severity threshold levels for labeling of healthy subjects and PAD patients. **Figure 5** shows the receiver operating characteristic (ROC) curves associated with our approach and ABI, both corresponding to varying PAD severity threshold levels for labeling of healthy subjects and PAD patients. **Figure 6** shows the Bland-Altman plots between true PAD severity vs. PAD severity predicted by our approach and ABI. **Figure 7** presents the 2-dimensional t-SNE visualization of the input and latent feature spaces associated with the fully trained and validated deep CNN, while **Figure 8** presents discriminative input features of our deep CNN localized by GradCAM associated with low and high PAD severity levels.

TABLE 1 | PAD detection performance of the deep learning-based pulse waveform analysis approach and ankle-brachial index, both corresponding to varying PAD severity threshold levels for labeling of healthy subjects and PAD patients.

Labeling threshold		**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
DL	Sensitivity	0.97	0.96	0.94	0.95	0.93	0.92	0.85
	Specificity	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Accuracy	0.99	0.98	0.97	0.97	0.96	0.95	0.91
	AUC	0.99	0.99	0.99	0.99	0.99	0.99	0.99
ABI	Sensitivity	0.96	0.94	0.73	0.64	0.60	0.58	0.59
	Specificity	0.50	0.50	0.64	0.75	0.91	0.99	0.99
	Accuracy	0.50	0.51	0.68	0.68	0.66	0.64	0.65
	AUC	0.73	0.74	0.76	0.79	0.83	0.88	0.92

DL, deep learning-based pulse waveform analysis approach; ABI, ankle-brachial index.

DISCUSSION

PAD is a highly prevalent CVD with profound morbidity and mortality implications, but it is frequently undiagnosed due to the limitations associated with the cost, comfort, and accuracy of existing angiography and ABI techniques. In this work, we investigated an affordable, convenient, and accurate PAD screening and diagnosis approach via DL-based PWA. Using a large number of virtual patients created with a validated multi-branch TL model of arterial hemodynamics, we illustrated its potential and challenges to overcome.

Validity of Virtual Patients

The virtual patients created with the multi-branch TL model of arterial hemodynamics could reproduce the clinically observed trends in the shape of the arterial pulse waveforms in response to varying degree of PAD severity. In particular, the multi-branch TL model predicted that ankle BP pulse undergoes the following morphological changes with an increase in the PAD severity level: (i) systolic peak flattens; (ii) secondary diastolic peak disappears; (iii) pulse amplitude decreases; (iv) crest time (time interval between diastolic trough and systolic peak) increases; and (v) pulse width at half amplitude increases (**Figure 4**). It also predicted that brachial pulse amplitude increases, which contributes to a decrease in ABI with an increase in the PAD severity level. These predictions are consistent with a number of existing clinical observations (Carter, 1968; Davies et al., 2014; Sumpio and Benitez, 2015; Dhanoa et al., 2016; Mao et al., 2017; Sibley et al., 2017) at least from qualitative standpoint. In sum, it was concluded that the virtual patients used in our work can produce realistically plausible arterial pulse waveforms with respect to varying degree of PAD severity, which provided a solid

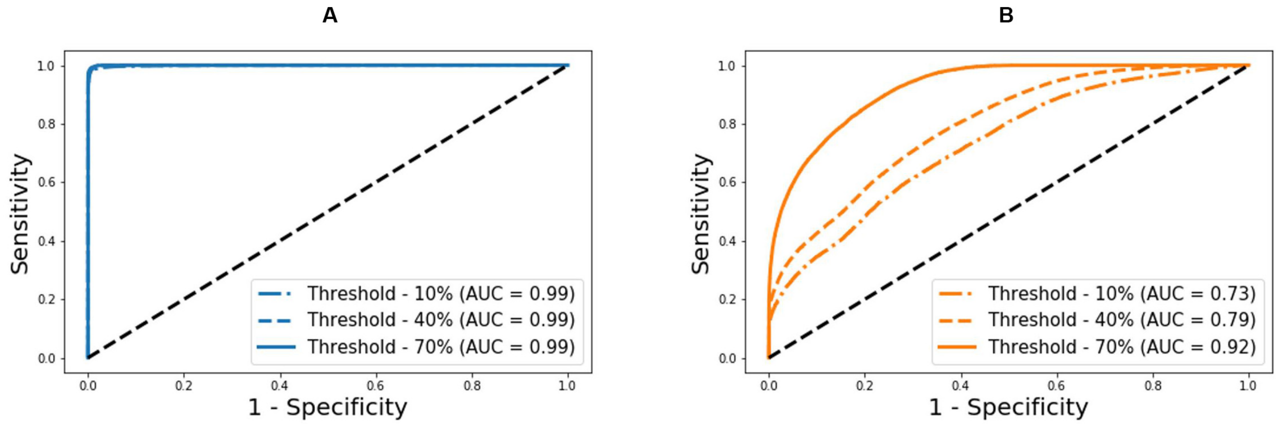


FIGURE 5 | Receiver operating characteristic curves associated with the deep learning-based pulse waveform analysis approach and ankle-brachial index (ABI), both corresponding to varying PAD severity threshold levels for labeling of healthy subjects and PAD patients. **(A)** DL-based pulse waveform analysis approach. **(B)** Ankle-brachial index.

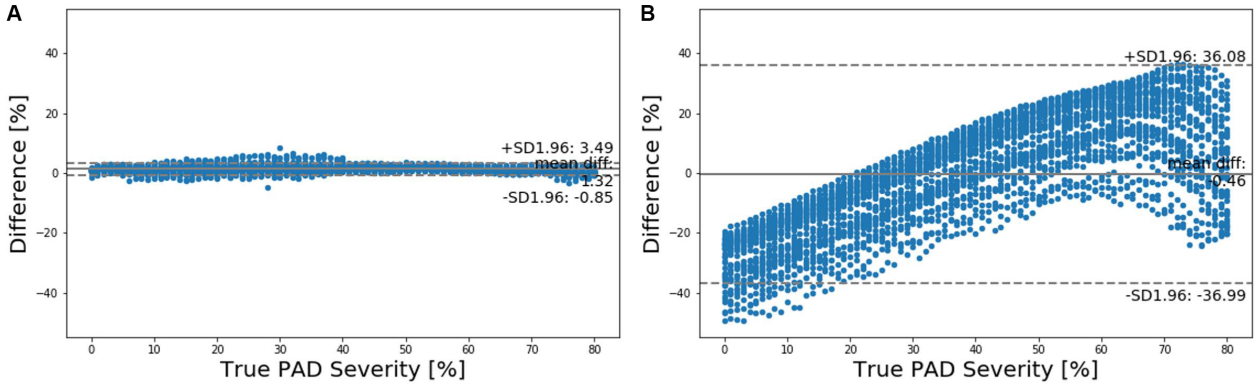


FIGURE 6 | Bland-Altman plots between true PAD severity vs. PAD severity predicted by **(A)** deep learning-based pulse waveform analysis approach and **(B)** ankle-brachial index (ABI).

basis to investigate the strengths and weaknesses of our DL-based PWA approach to PAD screening and diagnosis especially in comparison with the widely used ABI technique.

PAD Detection and Severity Assessment Efficacy

Our approach boasted robust PAD detection performance superior to the ABI technique against a wide range of PAD severity threshold levels for labeling of healthy subjects and PAD patients (**Table 1** and **Figure 5**). The sensitivity, specificity, and accuracy values computed at the PAD classification threshold levels identical to the labeling threshold values [note that (i) the deep CNN was calibrated to the true PAD severity as part of training, and (ii) a PAD severity level can be mapped to its corresponding ABI by using the polynomial regression model relating ABI to PAD severity in section “Evaluation”] were consistently higher in our approach than the ABI technique (**Table 1**). Our approach also boasted PAD severity assessment performance largely superior to the ABI technique, as indicated by its much smaller limits of agreement between the true vs. predicted PAD severity levels in comparison to its ABI counterparts (**Figure 6**). Overall, it appears that ABI is susceptible to the inter-individual variability in anatomical and physiological parameters which affect the systolic peak values associated with brachial and ankle arterial pulses, whereas our approach can cope with those confounding factors via highly sophisticated analysis of the two arterial pulse waveforms to exploit morphological characteristics beyond systolic peak values. The PAD detection and severity assessment performance remained consistent against repeated tests: the sensitivity, specificity, and accuracy values exhibited small coefficients of variation of the order of 10^{-3} across the 10 repeated tests outlined in section “Evaluation.” Lastly, the deep CNN appeared to be robust against modest perturbations in its hyper-parameters in that the alteration in the cost function with respect to the hyper-parameter perturbations considered in this work was small (<2.3%). This suggests that the AlexNet architecture used in this work was adequate, if not ideal.

Our approach exhibited a tendency for slight underestimation of PAD severity, especially at high PAD severity levels

(**Figure 6A**). This may explain its imperfect sensitivity relative to specificity at high PAD labeling threshold (**Table 1**), because underestimation of PAD severity in general makes the deep CNN conservative in detecting PAD. In contrast, the ABI technique suffered from a tendency for severe overestimation of PAD severity in low-severity PAD and also severe underestimation of PAD severity in high-severity PAD (**Figure 6B**). This may explain its deteriorating sensitivity and improving specificity (and the suboptimal accuracy as a whole) with respect to the increase in the PAD labeling threshold (**Table 1**). In our virtual patients, ABI tended to remain at a normal constant level up to ∼50% PAD severity level, beyond which it started to sharply decrease (not shown). Hence, the sensitivity of ABI is high in low PAD labeling thresholds (since it overestimates the severity in low PAD severity regime) but is low in high PAD labeling thresholds (since it underestimates the severity in high PAD severity regime). For the same reason, the specificity of ABI is low in low PAD labeling thresholds but is high in high PAD labeling thresholds. It is worth noting that this trend is in accordance with prior clinical observations on the low sensitivity and high specificity of ABI in detecting symptomatic PAD patients (Stein et al., 2006; Wikström et al., 2008).

Latent Feature and Interpretability Analysis

Two inherent challenges associated with DL is its susceptibility to overfitting and lack of transparency. We employed (i) t-SNE to examine if our deep CNN was properly trained and (ii) GradCAM to examine if our deep CNN exploits appropriate input features in diagnosing PAD.

The t-SNE visualization of the input and latent feature spaces clearly illustrates that the deep CNN was properly trained to capture the relationship between the latent features extracted from the brachial and ankle pulse waveforms and PAD severity (**Figure 7**). In particular, the input feature space contains a number of small and scattered clusters associated with varying PAD severity levels (**Figure 7A**), which presumably represent the inter-individual variability associated with the virtual patients. In contrast, the latent feature space clearly shows a manifold smoothly connecting low (upper left) to high (lower right) PAD severity levels (**Figure 7B**). Hence, it may be claimed that the notable performance of the DL-based PWA approach originates from its appropriate learning of the latent features indicative of PAD severity rather than from overfitting to the data.

The discriminative input features localized by GradCAM provide support for the transparency of the deep CNN constructed in this work. Indeed, main discriminative input features included (i) the systolic up-stroke and (ii) diastolic down-stroke (including secondary peaks when exists) (**Figure 8**), which are the regions in the brachial and ankle arterial pulses in which salient morphological changes occur as PAD develops according to the existing clinical literature (Carter, 1968; Davies et al., 2014; Sumpio and Benitez, 2015; Dhanoa et al., 2016; Mao et al., 2017; Sibley et al., 2017). Hence, it can be claimed that the DL-based PWA approach may detect and assess the severity of PAD by analyzing brachial and ankle arterial pulse waveforms in a way similar to how experienced clinicians analyze them, although the exact mechanisms underlying how the deep CNN compiles and interprets the observed morphological changes into PAD severity are unknown.

Limitations and Opportunities

All in all, this work demonstrated the proof-of-concept of integrating DL and PWA for affordable and non-invasive PAD screening and diagnosis. However, this work has a number of limitations to be addressed. In addition, this work also sheds light on outstanding opportunities toward its real clinical application.

First and foremost, this work was conducted using data collected from virtual rather than real patients. We employed a validated multi-branch TL model to create virtual patients. We also showed that arterial pulse waveforms produced by the virtual patients exhibit the morphological characteristics observed in real PAD patients. Yet, discrepancy between virtual vs. real patients may be inevitable at least to some extent, and there are a few potential sources that can obscure the initial success of this work when applied to real clinical data. In particular, the inter- and intra-individual variability considered in this work is somewhat ad-hoc. Furthermore, we accounted for variability associated only with arterial anatomical and physiological parameters but not cardiac parameters (such as stroke volume and ejection duration). In the near term, the efficacy of our approach against variabilities not considered in this work may be investigated using the same virtual patients. But ultimately, future work must confirm the proof-of-concept obtained in this work using clinical data collected from real patients. Regardless of this limitation, this work may still have unique value as an exploratory study of DL-based arterial pulse waveform analysis for PAD diagnosis in a reasonably realistic yet resource-effective and controlled setting. Indeed, our work may provide a strong justification for conducting a (potentially large-scale and resource-intensive) clinical data collection study for experimental investigation of DL-based PWA approaches to PAD diagnosis (and perhaps other CVDs as well).

Second, this work was limited to the detection and severity assessment of PAD in a single arterial site. In contrast, an ideal PAD screening and diagnosis tool is required to also localize PAD. Hence, our approach must be extended to a technique capable of simultaneously detecting, localizing, and assessing the severity of PAD. This requirement may present additional challenge when PAD at multiple sites with different levels of severity must be diagnosed. Future work must investigate how to extend our approach to also include PAD localization capability. A possible initial strategy may be to leverage the deep CNN trained in this work in conjunction with the multi-task learning, pre-training, and continuation methods established in the DL domain so as to extend the current deep CNN to also embed the ability to localize PAD.

Third, this work assumed the availability of a large amount of data associated with a wide range of variability in anatomical and physiological characteristics as well as PAD severity levels, which may not be practically realistic. For example, the majority of PAD data may be associated with aged patients, and our

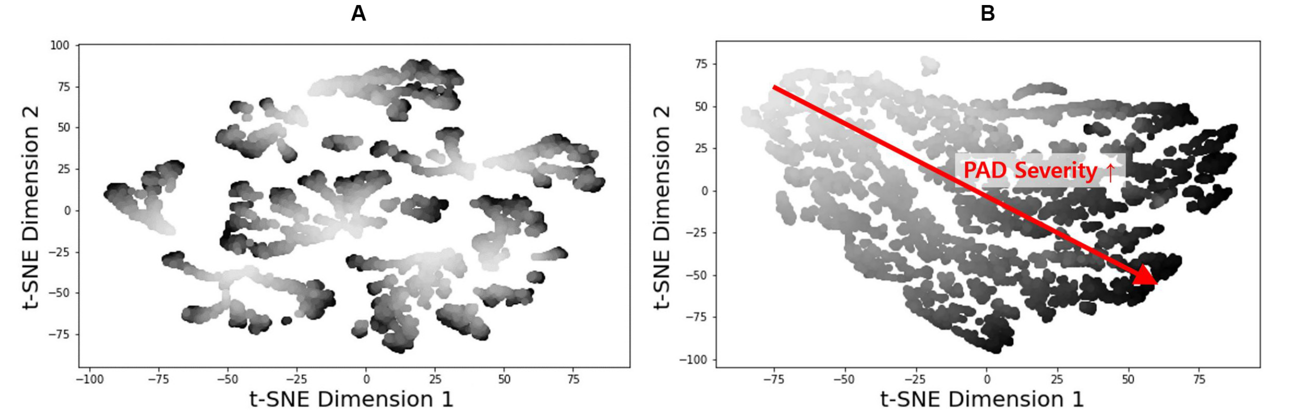


FIGURE 7 | 2-dimensional t-distributed stochastic neighbor embedding (t-NSE) visualization of **(A)** input and **(B)** latent feature spaces associated with the fully trained and validated deep convolutional neural network.

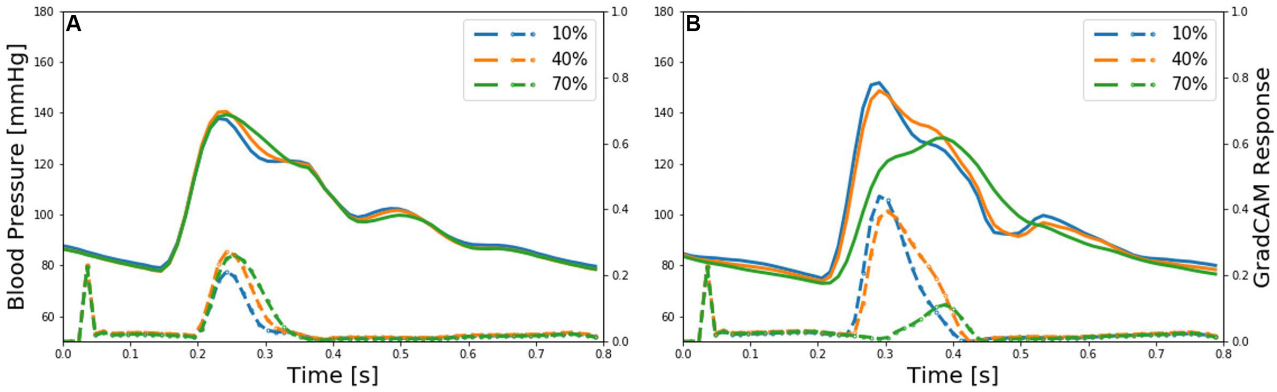


FIGURE 8 | Representative brachial and ankle pulse waveforms (solid lines) and discriminative features (dotted lines) of deep convolutional neural network (CNN) localized by the gradient-weighted class activation mapping (GradCAM) associated with low (10%), medium (40%), and high (70%) PAD severity levels. **(A)** Brachial arterial pulse. **(B)** Ankle arterial pulse.

approach when trained with such data may not generalize well to young patients (who are associated with low PAD incidence but screening/diagnosing whom is still crucial for CV risk management). Likewise, our approach when trained with data associated with one ethnic population may not generalize well to another subject to a large inter-ethnic anatomical and physiological discrepancies. Future work on coping with limited data and enormous inter-individual variability must be conducted. A possible initial strategy may be to exploit the domain adaptation and transfer techniques as well as adversarial training to guide the deep CNN work with latent features invariant to ethnic, anatomical, and physiological characteristics.

Lastly, this work used arterial BP waveforms, which may not be easy to measure non-invasively. Practically affordable non-invasive arterial pulse waveforms (e.g., pulse volume recording waveforms; Davies et al., 2014; Sumpio and Benitez, 2015; Ghasemi et al., 2018) are typically measured at the skin level and thus exhibit subtle morphological differences relative to arterial BP waveforms (Lee et al., 2018). Hence, future work must be conducted to investigate adverse effect of using non-invasive arterial pulse waveform measurements on our approach as well as innovative strategies to realize our approach using affordable and non-invasive arterial pulse measurements.

CONCLUSION

This work demonstrated the proof-of-concept of a novel DL-based PWA approach to PAD diagnosis. The results suggest that PAD detection and severity assessment may be feasible with data-driven analysis of arterial pulse waveforms. This work also outlined outstanding opportunities and challenges toward real-world deployment of our approach, including (i) validation with data collected from real patients, (ii) PAD localization, (iii) generalizable implementation with limited data and robustness against confounding factors, and (iv) practical embodiment with affordable and non-invasive arterial pulse waveforms. Future work to explore and address these opportunities and challenges, including the development of innovative DL-based PWA algorithms capable of addressing the outstanding obstacles, may serve as key cornerstones to realize affordable and convenient PAD screening and diagnosis.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SK, J-OH, and BY conceived the study and analyzed and interpreted the results. SK and J-OH created the virtual PAD patient data, developed the DL-based pulse wave analysis, and wrote and revised the manuscript. BY reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Research Foundation of Korea under Grant 2019H1D3A2A01061350 and the American Heart Association under Grant 18IPA34110466.

ACKNOWLEDGMENTS

We thank Dr. Hanguang Xiao at the Chongqing University of Technology for providing the access to and guidance on the use of his transmission line model of arterial hemodynamics.

REFERENCES

Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., and Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* 179:104992. doi: 10.1016/j.cmpb.2019.104992

Abdolmanafi, A., Duong, L., Dahdah, N., Adib, I. R., and Cheriet, F. (2018). Characterization of coronary artery pathological formations from OCT imaging using deep learning. *Biomed. Opt. Express* 9:4936. doi: 10.1364/boe.9.004936

Alaa, A. M., Bolton, T., Angelantonio, E., Di Rudd, J. H. F., and van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PLoS One* 14:e0213653. doi: 10.1371/journal.pone.0213653

Allison, M. A., Ho, E., Denenberg, J. O., Langer, R. D., Newman, A. B., Fabsitz, R. R., et al. (2007). Ethnic-specific prevalence of peripheral arterial disease in the United States. *Am. J. Prev. Med.* 32, 328–333. doi: 10.1016/j.amepre.2006.12.010

Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Gregory Hundley, W., McClelland, R., et al. (2017). Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ. Res.* 121, 1092–1101. doi: 10.1161/CIRCRESAHA.117.311312

Arruda-Olson, A. M., Afzal, N., Mallipeddi, V. P., Said, A., Pacha, H. M., Moon, S., et al. (2018). Leveraging the electronic health record to create an automated real-time prognostic tool for peripheral arterial disease. *J. Am. Heart Assoc.* 7:e009680. doi: 10.1161/JAHA.118.009680

Carter, S. A. (1968). Indirect systolic pressures and pulse waves in arterial occlusive diseases of the lower extremities. *Circulation* 37, 624–637. doi: 10.1161/01.CIR.37.4.624

Cavallo, A. U., Koktzoglou, I., Edelman, R. R., Gilkeson, R., Mihai, G., Shin, T., et al. (2019). Noncontrast magnetic resonance angiography for the diagnosis of peripheral vascular disease. *Circ. Cardiovasc. Imaging* 12:e008844. doi: 10.1161/CIRCIMAGING.118.008844

Davies, J., Lewis, J., and Williams, E. (2014). The utility of pulse volume waveforms in the identification of lower limb arterial insufficiency. *EWMA J.* 14, 21–25.

Dhanoa, D., Baerlocher, M. O., Benko, A. J., Benenati, J. F., Kuo, M. D., Dariushnia, S. R., et al. (2016). Position statement on noninvasive imaging of peripheral arterial disease by the society of interventional radiology and the Canadian Interventional Radiology Association. *J. Vasc. Interv. Radiol.* 27, 947–951. doi: 10.1016/j.jvir.2016.03.049

Dogan, M. V., Grumbach, I. M., Michaelson, J. J., and Philibert, R. A. (2018). Integrated genetic and epigenetic prediction of coronary heart disease in the framingham heart study. *PLoS One* 13:e0190549. doi: 10.1371/journal.pone.0190549

Ebrahimi Nejad, S., Carey, J. P., McMurtry, M. S., and Hahn, J. O. (2017). Model-based cardiovascular disease diagnosis: a preliminary in-silico study. *Biomech. Model. Mechanobiol.* 16, 549–560. doi: 10.1007/s10237-016-0836-8

Fowkes, F. G. R., Rudan, D., Rudan, I., Aboyans, V., Denenberg, J. O., McDermott, M. M., et al. (2013). Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis. *Lancet* 382, 1329–1340. doi: 10.1016/S0140-6736(13)61249-0

Ghasemi, Z., Lee, J. C., Kim, C. S., Cheng, H. M., Sung, S. H., Chen, C. H., et al. (2018). Estimation of cardiovascular risk predictors from non-invasively measured diametric pulse volume waveforms via multiple measurement information fusion. *Sci. Rep.* 8:10433. doi: 10.1038/s41598-018-28604-6

Golomb, B. A., Dang, T. T., and Criqui, M. H. (2006). Peripheral arterial disease: morbidity and mortality implications. *Circulation* 114, 688–699. doi: 10.1161/CIRCULATIONAHA.105.593442

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Guthaner, D. F., Wexler, L., Enzmann, D. R., Riederer, S. J., Keyes, G. S., Collins, W. F., et al. (1983). Evaluation of peripheral vascular disease using digital subtraction angiography. *Radiology* 147, 393–398. doi: 10.1148/radiology.147.2.6340157

Han, X., Zhong, Y., Cao, L., and Zhang, L. (2017). Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sens* 9:848. doi: 10.3390/rs9080848

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

He, W., Xiao, H., and Liu, X. (2012). Numerical simulation of human systemic arterial hemodynamics based on a transmission line model and recursive algorithm. *J. Mech. Med. Biol.* 12, 1–19. doi: 10.1142/S0219519411004587

Hirsch, A. T., Criqui, M. H., Treat-Jacobson, D., Regensteiner, J. G., Creager, M. A., Olin, J. W., et al. (2001). Peripheral arterial disease detection, awareness, and treatment in primary care. *J. Am. Med. Assoc.* 286, 1317–1324. doi: 10.1001/jama.286.11.1317

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI: IEEE), 4700–4708.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Red Hook, NY, 1097–1105. doi: 10.1201/9781420010749

Lee, J., Ghasemi, Z., Kim, C., Cheng, H., Chen, C., Sung, S., et al. (2018). Investigation of viscoelasticity in the relationship between carotid artery blood pressure and distal pulse volume waveforms. *IEEE J. Biomed. Heal. Informatics* 22, 460–470. doi: 10.1109/JBHI.2017.2672899

Li, G., Watanabe, K., Anzai, H., Song, X., Qiao, A., and Ohta, M. (2019). Pulse-wave-pattern classification with a convolutional neural network. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-51334-2

Mao, Y., Huang, Y., Yu, H., Xu, P., Yu, G., Yu, J., et al. (2017). Incidence of peripheral arterial disease and its association with pulse pressure: a prospective cohort study. *Front. Endocrinol.* 8:333. doi: 10.3389/fendo.2017.00333

Nelson, M. R., Quinn, S., Winzenberg, T. M., Howes, F., Shiel, L., and Reid, C. M. (2012). Ankle-brachial index determination and peripheral arterial disease diagnosis by an oscillometric blood pressure device in primary care: validation and diagnostic accuracy study. *BMJ Open* 2, 1–6. doi: 10.1136/bmjopen-2012-001689

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164. doi: 10.1038/s41551-018-0195-0

Romano, M., Mainenti, P. P., Imbriaco, M., Amato, B., Markabaoui, K., Tamburrini, O., et al. (2004). Multidetector row CT angiography of the abdominal aorta and lower extremities in patients with peripheral arterial occlusive disease: diagnostic accuracy and interobserver agreement. *Eur. J. Radiol.* 50, 303–308. doi: 10.1016/S0720-048X(03)00118-9

Ross, E. G., Shah, N. H., Dalman, R. L., Nead, K. T., Cooke, J. P., and Leeper, N. J. (2016). The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J. Vasc. Surg.* 64, 1515–1522.e3. doi: 10.1016/j.jvs.2016.04.026

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, Venice, 618–626.

Sibley, R. C., Reis, S. P., MacFarlane, J. J., Reddick, M. A., Kalva, S. P., and Sutphin, P. D. (2017). Noninvasive physiologic vascular studies: a guide to diagnosing peripheral arterial disease. *Radiographics* 37, 346–357. doi: 10.1148/rg.2017160044

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., and Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* 13:e0202344. doi: 10.1371/journal.pone.0202344

Stein, R., Hriljac, I., Halperin, J. L., Gustavson, S. M., Teodorescu, V., and Olin, J. W. (2006). Limitation of the resting Ankle-Brachial index in symptomatic patients with peripheral arterial disease. *Vasc. Med.* 11, 29–33. doi: 10.1191/1358863x06vm663oa

Sumpio, B. E., and Benitez, E. (2015). Pulse volume recording for peripheral vascular disease diagnosis in diabetes patients. *J. Vasc. Diagnostics* 3, 33–39. doi: 10.2147/jvd.s68048

Vallée, A., Cinaud, A., Blachier, V., Lelong, H., Safar, M. E., and Blacher, J. (2019). Coronary heart disease diagnosis by artificial neural networks including Aortic Pulse wave velocity index and clinical parameters. *J. Hypertens.* 37, 1682–1688. doi: 10.1097/hjh.0000000000002075

Wang, S. H., Xie, S., Chen, X., Guttery, D. S., Tang, C., Sun, J., et al. (2019). Alcoholism identification based on an AlexNet transfer learning model. *Front. Psychiatry* 10:205. doi: 10.3389/fpsyt.2019.00205

Wikström, J., Hansen, T., Johansson, L., Lind, L., Ahlström, H., Hansen, T., et al. (2008). Ankle brachial Index <0.9 underestimates the prevalence of peripheral artery occlusive disease assessed with whole-body magnetic resonance angiography in the elderly ankle brachial index v 0 . 9 underestimates the prevalence of peripheral artery occlus. *Acta Radiol.* 49, 143–149. doi: 10.1080/02841850701732957

Xiao, H., Avolio, A., and Huang, D. (2016a). A novel method of artery stenosis diagnosis using transfer function and support vector machine based on transmission line model: a numerical simulation and validation study. *Comput. Methods Programs Biomed.* 129, 71–81. doi: 10.1016/j.cmpb.2016.03.005

Xiao, H., Avolio, A., and Zhao, M. (2016b). Modeling and hemodynamic simulation of human arterial stenosis via transmission line model. *J. Mech. Med. Biol.* 16:1650067. doi: 10.1142/S0219519416500676

Xiao, H., Tan, I., Butlin, M., Li, D., and Avolio, A. P. (2017). Arterial viscoelasticity: role in the dependency of pulse wave velocity on heart rate in conduit arteries. *Am. J. Physiol. Heart Circ. Physiol.* 312, H1185–H1194. doi: 10.1152/ajpheart.00849.2016

Zhang, N., Yang, G., Gao, Z., Xu, C., Zhang, Y., Shi, R., et al. (2019). Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. *Radiology* 291, 606–617. doi: 10.1148/radiol.2019182304

Conflict of Interest: BY was employed by company OnePredict, Inc.

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 28 August 2020
doi: 10.3389/fbioe.2020.01005

Automated Detection of Acute Lymphoblastic Leukemia From Microscopic Images Based on Human Visual Perception

Alexandra Bodzas, Pavel Kodytek and Jan Zidek*

Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava, Czechia

Edited by:
Ying Li,
Zhejiang University, China

Reviewed by:
Itaru Kato,
Kyoto University, Japan
Neetika Nath,
Universitätsmedizin Greifswald, Germany
Xiaoxian Yang,
Shanghai Second Polytechnic University, China

****Correspondence:***
Alexandra Bodzas
alexandra.bodzas@vsb.cz

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

Received: *29 February 2020*
Accepted: *31 July 2020*
Published: *28 August 2020*

Citation:
Bodzas A, Kodytek P and Zidek J (2020) Automated Detection of Acute Lymphoblastic Leukemia From Microscopic Images Based on Human Visual Perception. Front. Bioeng. Biotechnol. 8:1005. doi: 10.3389/fbioe.2020.01005

Microscopic image analysis plays a significant role in initial leukemia screening and its efficient diagnostics. Since the present conventional methodologies partly rely on manual examination, which is time consuming and depends greatly on the experience of domain experts, automated leukemia detection opens up new possibilities to minimize human intervention and provide more accurate clinical information. This paper proposes a novel approach based on conventional digital image processing techniques and machine learning algorithms to automatically identify acute lymphoblastic leukemia from peripheral blood smear images. To overcome the greatest challenges in the segmentation phase, we implemented extensive pre-processing and introduced a three-phase filtration algorithm to achieve the best segmentation results. Moreover, sixteen robust features were extracted from the images in the way that hematological experts do, which significantly increased the capability of the classifiers to recognize leukemic cells in microscopic images. To perform the classification, we applied two traditional machine learning classifiers, the artificial neural network and the support vector machine. Both methods reached a specificity of 95.31%, and the sensitivity of the support vector machine and artificial neural network reached 98.25 and 100%, respectively.

Keywords: automated leukemia detection, blood smear image analysis, cell segmentation, leukemic cell identification, acute leukemia, image processing, machine learning

INTRODUCTION

Leukemia is a term describing a group of hematological malignancies that are manifested by the tumourous proliferation or increased life span of immature white blood cells (WBCs) in the bone marrow (American Dental Association [ADA], 2012). Leukocytes are highly differentiated for their specialized functions, and they play an essential role in the immune system (Rogers, 2011). The malignancy of this disease varies from non-malignant to highly aggressive forms, and the immature cells are not able to fulfill their normal function (Serfontein, 2011). The excessive production of these type of cells, denoted as blasts or leukemic cells crowds out healthy leukocytes in the bone marrow and suppresses normal hematopoiesis, causing difficulties in fighting infections, transporting oxygen and controlling bleeding (Daniels and Nicoll, 2012). Clinically, leukemia is categorized on the basis of the rapidity of the disease progression to acute and chronic forms. Whereas the acute form of leukemia develops quickly and the number of leukemic cells increases rapidly, chronic leukemia progresses slowly over time, and the more mature leukocytes can carry out some of their normal functions (Serfontein, 2011). According to the type of affected cell from which the malignancy develops, leukemia is further divided into myelogenous and lymphoid forms (Manisha, 2012). Acute lymphoblastic leukemia (ALL), which is the only form we consider in this paper, is the second most common type of leukemia in adults and the most common type of childhood malignancy, accounting for approximately one-third of all pediatric cancers (Rose, 2013). Heterogeneous malignancy is caused by genetic alterations and chromosomal mutations of lymphocyte progenitor cells at an early phase of cell differentiation (Rose, 2013). The excessive production of these cells, called lymphoblasts, which do not develop into mature B and T lymphocytes, gradually displaces normal cells in the bone marrow and may spread to essential organs such as the liver, lymph nodes, spleen, and central nervous system (Katz et al., 2015).

The diagnosis of ALL requires a broad spectrum of information derived from several modalities, including morphology, cell phenotyping, cytochemistry, cytogenetics, and molecular genetics (Inaba et al., 2013). Despite technological advances in medicine, morphology remains the frontline hematological diagnostic technique. The observation of excessive leukemic cell buildup and morphological anomalies in cellular structures during the visual examination of peripheral blood smears arouses the first suspicion of leukemia. Because manual microscopic examination is a time-consuming process that requires a considerable amount of experience and is prone to humane error (Inaba et al., 2013), such an automated inspection is needed, which would standardize the examination process and circumvent the drawbacks of this diagnostic technique.

To minimize human intervention and overcome the abovementioned limitations, several computerized methods have been explored. Most of these methods utilize conventional image processing and machine learning techniques, which involve mainly segmentation, feature extraction, and classification methods. Especially the segmentation and feature extraction phases are considered the most significant and challenging tasks (Neoh et al., 2015). The main reason lies in the large variety of blood smear images, taken under different conditions, and the potential morphological differences between blast cells. Although some of these proposed methods were found to be faster and more cost effective than manual examination, their impact and accuracy remain insufficient (Shafique and Thesin, 2018). Whereas, Wang et al. (2019) achieved a detection speed of 14 to 100 milliseconds by utilizing convolution neural networks and GPU, most proposed methods produce false-negative errors and achieve overall accuracy in the range of 93–98% (Bagasjvara et al., 2016).

In this study, we propose a novel combination of techniques to overcome the most challenging parts of the detection process and present detailed insights into the greatest shortcomings of the existing classification methodologies, such as the overfitting and the reliability of particular classifications. To improve our segmentation phase, we introduce extensive pre-processing based on the proposed color transformation and design a three-phase filtration that ensures the elimination of surrounding blood components and artifacts without disrupting particular regions of leukocytes. After the whole segmentation process, involving seven stages, a robust set of features is extracted from all segmented regions. Extracting morphological and texture features from specific cell regions in a similar way to the visual interpretation of a domain expert heightens the performance of the selected classifiers. The final recognition of ALL from peripheral blood smear images is accomplished by an artificial neural network (ANN) and optimized support vector machine (SVM).

LITERATURE REVIEW OF THE PREVIOUSLY PROPOSED METHODOLOGIES

Extensive research has recently been conducted to explore the possibilities for the automated detection of leukemia from microscopic blood smear images (Alsalem et al., 2018). Most previously proposed methods employ sequential image pre-processing, cell segmentation, feature extraction, and cell classification (Bodzas, 2019). The main aim of the pre-processing phase is to enhance the image quality for subsequent processing. Many authors have enhanced blood smear images by converting them to another color domain, which highlights the particular features of the objects and therefore increases the efficiency of region detection (Aljaboriy et al., 2019). For example, Putzu et al. (2014) and Hariprasath et al. (2019) stated that the identification of WBCs is possible with conversion to the CMYK color model. The reason is that leukocytes have a higher contrast in the Y component since the yellow color is present in all elements except WBCs.

On the other hand, Moradiamin et al. (2015) converted images from the RGB color space to HSV, which reduced the correlation between the color channels in comparison to RGB and enabled the three H, S, and V channels to be dealt with separately. They additionally complemented this with a pre-processing phase with histogram equalization, which reduced the effect of different lightening conditions. After nucleus segmentation by the fuzzy C-means clustering algorithm, the authors extracted five geometrical and 72 statistical features. The dimensionality of the feature set was reduced by principal component analysis to eight features, which were subsequently applied to the SVM classifier.

A different approach was introduced by Kazemi et al. (2016) by implementing selective median filtering in combination with conversion to the CIEL*a*b model, in which the perceptual difference between colors is proportional to the Cartesian distance. In simple terms, the formula CIEL*a*b takes the XYZ tristimulus values and the white reference to produce correlates to the luminence, chroma, and hue elements (Fairchild, 2005). To extract the nucleus of WBCs, color-based clustering segmentation with additional morphological filtering was implemented. The set of features, including irregularity, the Hausdorff dimension, shape, color, and texture, was extracted from a whole image containing multiple nuclei. By applying a two-class SVM, they were able to achieve an overall accuracy of 96%.

In addition to the clustering segmentation method, many authors have used thresholding-based techniques to segment WBCs. In particular, Joshi et al. (2013) reported the usage of Otsu's global thresholding on an enhanced greyscale image. To differentiate blasts in a microscopic blood smear image, they extracted the area, perimeter, and circularity from the equivalent binary image and employed the K-nearest neighbor decision algorithm for classification.

Due to the absence of spatial information, threshold techniques cannot always produce relevant and precise results. Hence, they are often combined with mathematical morphology or other image processing techniques. For instance, Wang et al. (2008) proposed a segmentation algorithm that combined adaptive thresholding with an edge-based technique and seeded watershed to recognize cell nuclei in different cycle phases. Moreover, unlike other studies using off-line learning algorithms, the authors in this study deployed an online SVM classifier, which removed the support vectors from the older model and assigned weights to the new samples according to their importance to accommodate changing conditions.

Concerning feature extraction and classification, recent research has shown that the most preferred methodologies use a combination of morphological and texture features with supervised learning algorithms. In particular, SVM and multilayer perceptron have provided higher accuracy than methods using other classifiers (Aljaboriy et al., 2019). For instance, research by Neoh et al. (2015) extracted a total of 80 feature descriptors containing color, shape, and texture information to compare the classification performance of the SVM and multilayer perceptron. Both classifier results reached a similar accuracy, over 95%, with slightly higher accuracy for the multilayer perceptron classifier.

MATERIALS AND METHODS

The main goal of this work is to develop a fully automated system for ALL detection that can be applied to complete blood smear images containing multiple WBCs. The solution presented in this paper is based on conventional image processing techniques and comprises four main stages, which are described in the following subchapters.

Blood Smear Image Dataset

The proposed system was trained as well as tested on a local dataset, which was provided by the Department of Haemato-oncology at the University Hospital Ostrava. The anonymized dataset consists of 18 microscopic blood smear images obtained from patients without pathological findings and 13 blood smear images from patients with diagnosed ALL. On average, six blood smear images with a resolution of 4,080 × 3,072 were captured per patient. Since WBCs are distributed unevenly, with a predominance of large cells on the border and smaller cells in the center of the blood smear, systematic data acquisition was required (Bodzas, 2019, p. 45). This was carried out by the meander inspection pattern, which allowed microscopic images to be captured from different consecutive locations, particularly from both edges and the center of the blood smear. All slides in the dataset were stained with Giemsa stain and were captured under the same lighting conditions by an Olympus CX43 microscope under a magnification of 50 times with an oil immersion objective lens and an effective magnification of 500 (Bodzas, 2019, p. 45).

The manual examination of blood smear images was conducted by local domain experts. During this visual examination, the hematology specialists used several morphological criteria to distinguish between lymphoblasts and normal cells. The most significant criteria included the nucleus position and shape, chromatin structure, presence of nucleoli, nucleocytoplasmic ratio, size of the cell, and color or structure of the cytoplasm. Following the WHO classification system, ALL is divided into B-lymphoblastic leukemia/lymphoma, T-lymphoblastic leukemia/lymphoma, and acute leukemias of ambiguous lineage. Because, from a morphological point of view, there are no reproducible criteria to distinguish between B and T lineage lymphoblastic leukemia, ALL subtype classification is not considered in this study (Chiaretti et al., 2014).

Pre-processing

During the acquisition process, numerous variable factors, such as different illumination conditions, staining time, blood film thickness and film defects, may introduce undesirable visual artifacts or cause different color distributions among the images (Díaz and Manzanera, 2009). To deal with potential microscopic image artifacts and enhance the contrast of the individual blood elements, we introduced a pre-processing method based on the standard arithmetic operations followed by gamma correction and contrast enhancement algorithms. The proposed color transformation is described by the following formula

$$g\left(x,y\right) = [(L-1) - B] - \{\,[(L-1) - \mathrm{G}]\,0.5\} \tag{1}$$

where $g(x,\ y)$ is the transformed image, L is the number of distinct gray levels in the image and B and G are the blue and green color spaces. Using arithmetic operations on the individual color spaces enhanced the blood smear images and allowed finer differentiation of the leukocytes, even for cells with scanty cytoplasm (Bodzas, 2019, p. 46).

Leukocyte Segmentation

After applying the pre-processing step, the segmentation phase was performed. The segmentation phase, which is concerned with extracting individual object components carrying pivotal information, is considered the most essential and challenging task. The aim of this task is to reduce the computational complexity of the subsequent steps, and to reduce the size of the high-resolution images, which heavily burden the storage capacity of the hospital's server (Chen et al., 2020). From a morphological point of view, leukemic cells can be distinguished from mature leukocytes by having a large nucleus with finely dispersed chromatin, moderate and non-granular cytoplasm, and one or more prominent nucleoli (Wiernik, 2001). The challenging process in this work comprises two main steps: leukocyte localization and region extraction, which separates the specific cell components (nucleus and cytoplasm). The entire segmentation process, divided into these two main parts, is shown in **Figure 1**.

The most precise segmentation results of the leukocyte localization phase were achieved by an algorithm involving four fundamental stages, which can be seen in the diagram above. The main aim of this phase was to remove the background and the surrounding blood components and to separate any touching cells. The first step of this challenge is the conversion of the image into a binary format, which was performed by the histogram-based thresholding segmentation method. Due to the sensitive pre-processing phase, thresholding reduced the background and part of the erythrocytes, while the full size of the WBCs was retained. Considering that erythrocytes usually have the shape of a biconcave disk with an inclination to overlap each other and that platelets lie in a different color spectrum, the process of thresholding often results in an image with additional noise. To eliminate the residual parts of the cell components and blood film defects from the image, we present a three-phase filtration (Bodzas, 2019, pp. 47–48).

The first phase of the three-phase filtration is focused on the removal of small objects, which is performed by the modified morphological opening operation using a disk-shaped small structuring element (Bodzas, 2019, p. 48). The modification of this operation lied in the uneven ratio between the number of iterations of the dilatation and the erosion parts of the closing operation (in particular, using the ratio 8:1). Using different iteration ratios allows the regions containing the WBCs to be preserved without a considerable reduction of the cell, and effectively removes smaller objects, such as the remaining parts of the erythrocytes and the platelets resistant to the thresholding operation. The first phase of the proposed algorithm is complemented with the second filtration step, which is based on connected component labeling followed by histogram-based filtration.

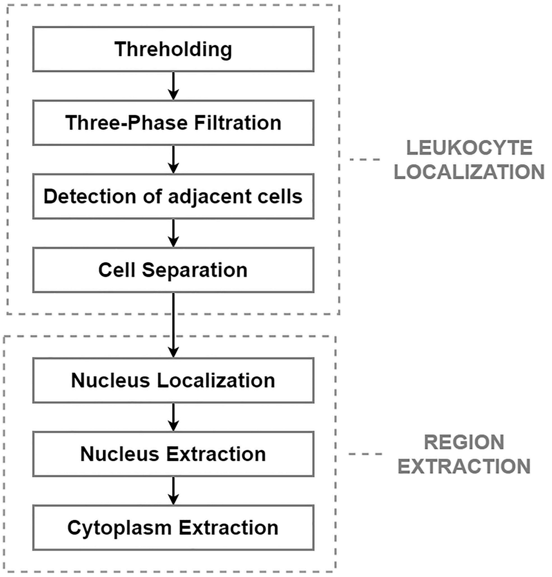


FIGURE 1 | The proposed segmentation algorithm.

This second phase of filtration is described by the following equations, where x and y are image coordinates that belong to the set of natural numbers, C_i denotes the cumulative sum of the same valued pixels in the image array and $I \in <0, n>$, where n is the number of distinct gray levels in the image.

$$f_{\mathrm{x,y}}i = \begin{cases} 1 & f_{\mathrm{x,y}}=i \\ 0 & \text{else} \end{cases} \tag{2}$$

$$\mathrm{C_i} = \sum_{x}\sum_{y} f_{\mathrm{x,y}}\ (i) \tag{3}$$

To remove all small objects in the image, we calculate the set S (see Eq. 4), where each value of i that satisfies the condition of "being small" is included. Based on the histogram evaluation, we select a threshold value T_s of 4,000. Values of i that do not satisfy the condition are excluded.

$$S_{\mathrm{i}} = \begin{cases} i & \mathrm{C_i} > T_{\mathrm{s}} \\ 0 & else \end{cases} \tag{4}$$

The output image $g(x,y)$ is constructed from the input image $f(x,y)$ in such a way that only the pixels with a nominal intensity belonging to a subpart of the set S are distributed to the output image, while the rest are set to 0. Thus, we ensure that the least commonly occurring intensity numbers are removed from the image.

$$g_{\mathrm{x,y}} = \begin{cases} f_{\mathrm{x,y}} & f_{\mathrm{x,y}} \notin S \\ 0 & else \end{cases} \tag{5}$$

Applying the second filtration step helps to smooth the image and remove all objects of small and medium size that are resistant to our opening operation. The last phase of the proposed three-phase filtration process is focused on the elimination of large blood film artifacts, which usually arise during the staining process. Since large artifacts such as precipitated stains and crushed cells have a very distinct texture and color spectrum, the mean particle color derived from the histogram is applied in combination with the particle area (Bodzas, 2019, p. 48). Using the histogram of a green color space, where the WBCs are more contrasted, prevents filtering of normal cells and cells with size abnormalities. The process of the localization of leukocytes, including the fundamental steps, is shown in **Figure 2**.

The blast cells tend to aggregate in clumps. The presence of such adjacent cells in an image often introduces high inaccuracy in the subsequent image processing stages. In particular, shape-based features such as the perimeter and area are highly dependent on the segmentation results. In clinical practice, to minimize the risk of miscounting, domain experts usually avoid adjacent cells or, in specific cases, solely examine clumps where the cytoplasm and nucleus are clearly identifiable. Each clearly detectable clump or adjacent cell in the image should therefore

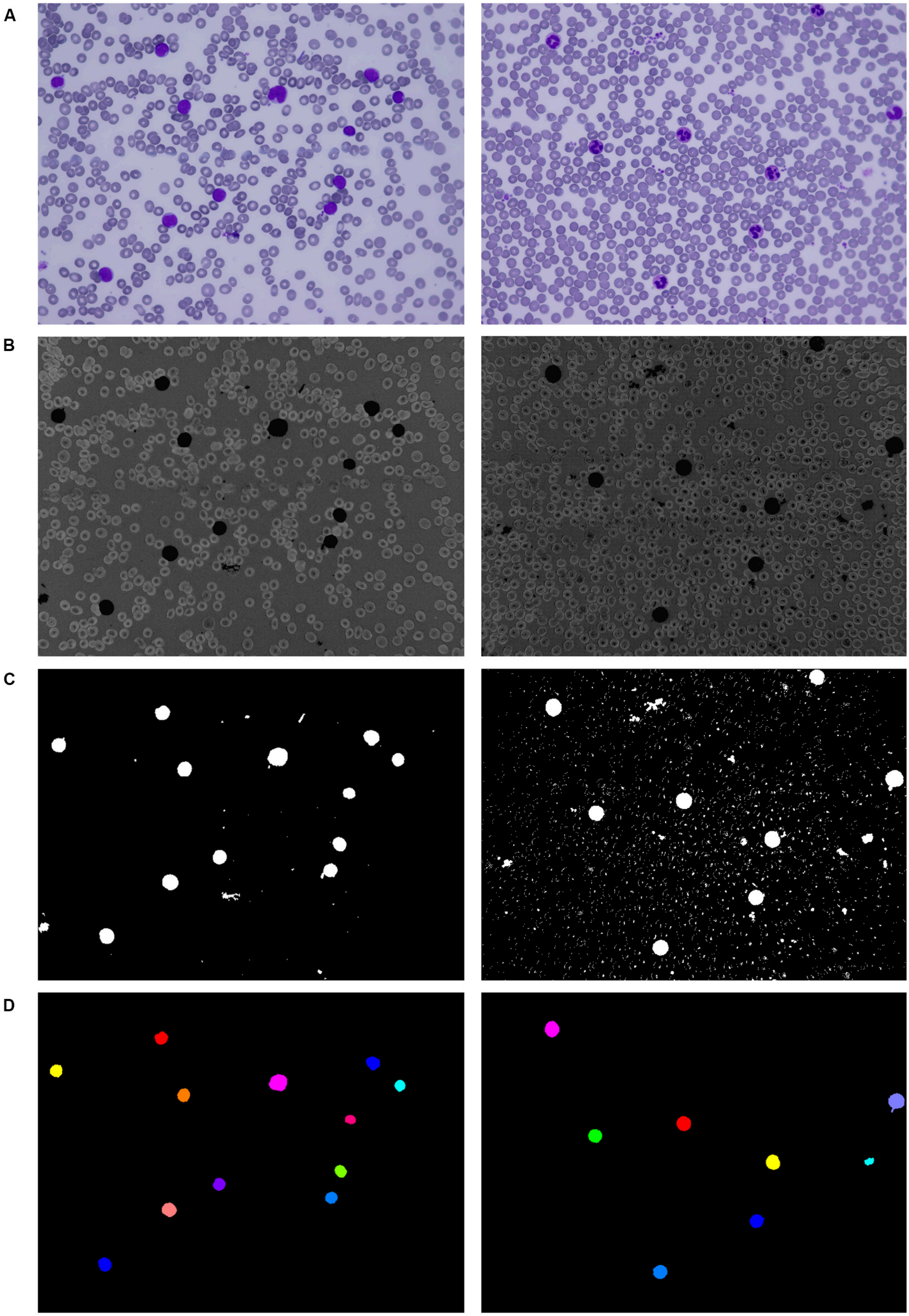


FIGURE 2 | Localization of white blood cells. **(A)** Original blood smear image. **(B)** Pre-processing results. **(C)** Thresholding segmentation results. **(D)** Application of the three-phase filtration with image labeling.

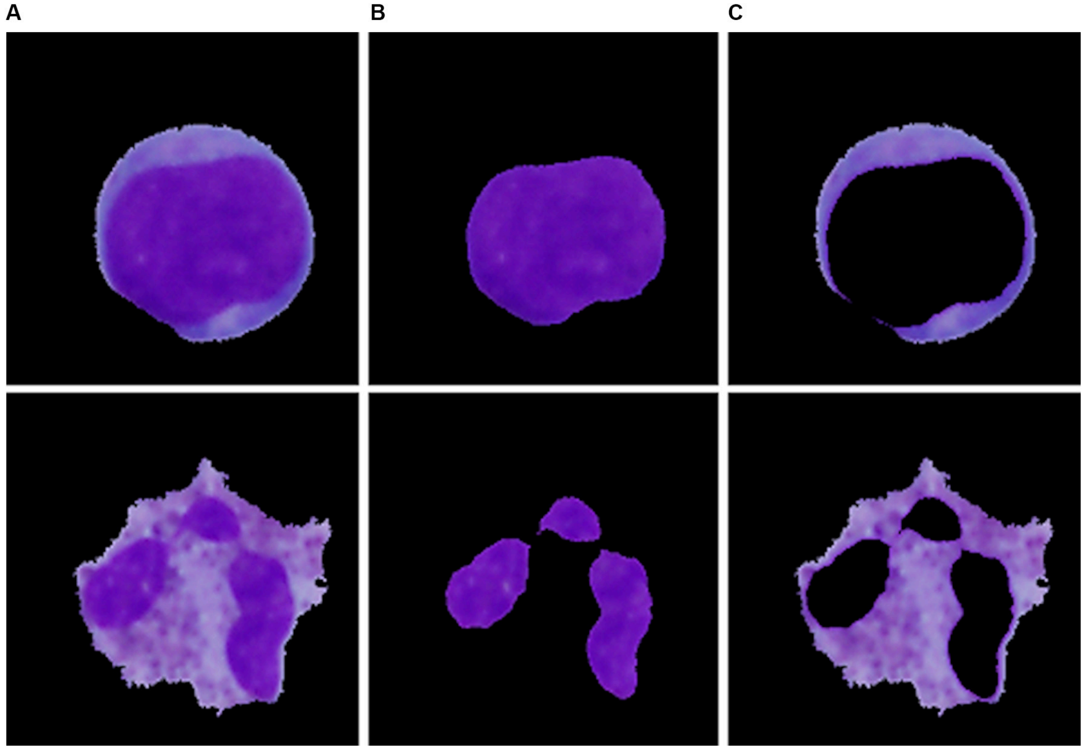


FIGURE 3 | The particular segmentation results of the blast cell (top) and normal leukocyte (bottom). **(A)** Segmented cell. **(B)** Segmented nucleus. **(C)** Segmented cytoplasm.

be identified and then separated into individual cells. For the identification of adjacent cells and cell clumps, the total particle area computation and morphological erosion, in combination with particle counting, are implemented. Morphological erosion is, in this case, used to separate touching objects that can be subsequently counted. Since the blast cells are nearly round and the touching edge length is smaller than the radius of either object, the touching cells can be separated well without concern that the objects will be eroded into nothing. After detecting the adjacent cells, the cells are separated by applying the Sobel edge detection technique, which specifies the approximate region of the splitting boundary (Bodzas, 2019, p. 49).

Single-cell sub-image extraction was performed in this work by an automatic image crop using the bounding rectangle size, which is the smallest rectangle containing a particular component. Once the single leukocytes had been identified and cropped into single-cell sub-images, we finally proceeded to the second segmentation stage (region extraction), which focuses on the extraction of the nucleus and the cytoplasm into individual parts. Thus involves the following steps: nucleus localization, nucleus extraction, and extraction of the cytoplasm. To localize the nucleus, we employed equalization in the luma plane and performed color thresholding to extract the saturation channel from the HSL space, where the border of the nucleus seemed to be the most prominent. The process of nucleus extraction was accomplished by multiplying the original sub-image with the obtained binary image. Finally, the separated nucleus was used to obtain the cytoplasm by subtracting the nucleus from the original image. The results of the region extraction algorithm are shown in **Figure 3**.

Features Extraction

In general, the extracted features describe the texture or shape information obtained from the segmented pattern and thereby help to reduce the dimensionality of the image to produce a result that is more informative and less redundant than the original image (Wan and Mak, 2015). In this phase, we aimed to extract the descriptive information from an image in the way that domain experts do. The proper selection of the features is considered the second most challenging step in the field of automated identification of leukemic cells. To construct an effective feature set, several published articles and their feature selection methods were studied. In this work, we implemented sixteen widely used features, of which nine had morphological characteristics and seven had statistical characteristics (Bodzas, 2019, p. 51). Another approach to extract features is the use of a convolution neural network model, which extracts a collection of feature vectors (Gao et al., 2019). In contrast to our approach, this feature space does not carry fully comprehensible information, and therefore cannot be interpreted in deep detail.

Morphological Features

According to hematology experts, the shape of the nucleus has proven to be a good measure for immature cell recognition. Apart from rudimentary measures such as the nucleus and cytoplasm area and nucleus perimeter, the following shape descriptors were considered.

Nuclear-cytoplasmic ratio

The ratio of the area of the cell nucleus to the cytoplasm area. This measure is a pivotal feature for the assessment of the maturity of the cell and, in turn, the prediction of cell malignancy. In general, the size of the nucleus decreases with increasing degree of leukocyte maturity.

Nucleus compactness

The extent to which the shape is compact. Depending on the maturity and the type of the WBC, the shape of the nucleus varies greatly. Mature cells usually have multi-lobed nuclei with lobes connected by thin strands or bands. Furthermore, in specific cases, the nucleus can have kidney bean or horseshoe-shaped contours. By contrast, leukemic cell nuclei are generally ovoid or round in shape and exhibit higher overall compactness than the nuclei of to mature cells. The compactness measure is given by the following formula (Chan et al., 2010).

$$\text{Compactness} = \frac{\text{Perimeter}^2}{\text{Area}} \tag{6}$$

Nucleus form factor

A measure of shape irregularities independent on the object's size. In general, a circular nucleus has the greatest area to perimeter ratio, and this measure is equal to 1 for a perfect circle. Consequently, for the nuclei of leukemic cells, this ratio converges to a value of 1, while the nuclei of normal cells which depart from roundness have a lower value. The form factor is defined as

$$\text{Form factor} = \frac{4{*}\pi{*}\text{Area}}{\text{Perimeter}^2} \tag{7}$$

Nucleus eccentricity

Nucleus eccentricity indicates the deviation from a circular shape. This measure is calculated as the ratio of the length and width of the minimal bounding rectangle of the region of interest. Unlike the form factor, this measure takes into account the elliptic shapes or circular lobes of the nucleus.

Nucleus elongation

Nucleus elongation indicates abnormal bulging. This measure is calculated as the ratio of the maximum and minimum distance from the center of gravity to the boundary. This feature highlights WBCs with a multi-lobed elongated nucleus.

Nucleus solidity

Nucleus solidity defines the degree to which the shape is convex or concave and is computed as the ratio of the area and the convex hull area (Ahmed et al., 2016).

Statistical Features

Other indispensable descriptors used for the identification of blast cells are based on changes in the nuclear chromatin pattern reflecting DNA formation and on cytoplasmic changes. To capture the crucial information of the structural arrangement of the nucleus and the entire cell, two types of statistical measures were used. The first-order statistical measures are based on the histogram of the greyscale image, e.g., the cytoplasm and the nucleus mean color, and the second-order statistical measures are derived from the gray level co-occurrence matrix (GLCM), which carries information about the spatial relationships of the image pixels. The second-order statistical features selected in this study are defined by the equations below, where $P(i,\ j)$ is the element of the normalized GLCM at the coordinates i and j, N_g denotes the number of distinct gray levels and μ_x,μ_y and σ_x,σ_y represent the means and standard deviations of the normalized gray level co-occurrence matrix, respectively (Bodzas, 2019, pp. 52–53).

Nucleus energy

A measure of the local textural uniformity of gray levels, defined as

$$Energy = \sum_{i,\ j=0}^{N_g-1} \left(P_{\mathrm{i,j}}\right)^2 \tag{8}$$

Cell contrast

Cell contrast measures the number of local variations in the GLCM. This measure is given by the relation.

$$Contrast = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g-1} \sum_{j=1}^{N_g-1} P(i,j) \right\},\ |i-j| = n \tag{9}$$

Nucleus correlation

Nucleus correlation represents the linear dependency of gray tone values in the GLCM. The correlation measure is given by the following formula.

$$Correlation = \frac{\sum_i \sum_j (i,j)P(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{10}$$

Cell dissimilarity

Cell dissimilarity calculates the mean of the gray level difference distribution of a region and is given by the relation.

$$Dissimilarity = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} |i-j|\, P(i,j) \tag{11}$$

Cell entropy

Cell entropy measures the randomness or complexity of texture. The entropy can be calculated using the following formula (Batchelor and Waltz, 2001; Nailon, 2010; Ahmed et al., 2016).

$$Entropy = -\sum_{i=0} \sum_{j=0} P(i,j) \log P(i,j) \tag{12}$$

All selected features were validated by using the statistical hypothesis testing method, which determined whether the samples representing the normal and blast cells came from the same population, or in other words, whether the distribution was the same for both classes. Since the analyzed data did not have a normal distribution, the median and median absolute deviation (MAD) were the proper measures to describe the observations in the dataset. In general, the analyzed features can be considered to be separable in the case of sufficiently different median values and low values of MAD that describe how spread out the data are. In this work, we used the Mann–Whitney *U* test to evaluate the statistically significant differences between the two observed groups. **Table 1** shows the resulting probabilities (*p*-values) that the distributions, or in simple terms, the changes in the median values of the two classes, are not significantly different (Bodzas, 2019, p. 54).

According to **Table 1**, 15 features seem to be highly unique, with great differences between the normal and leukemic cells. Even though nucleus eccentricity results with a much lower probability, this feature is statistically significant and plays an essential role in the subsequent classification phase. Owing to the high variability of the features, which encompass a wide range of cell attributes from morphological to textural, there should not be a concern of misclassification in case of blasts with variable sizes or normal cells with size-related anomalies.

Classification

Depending on the selected classifier, the efficiency and performance of the features may vary slightly. The classification step that classifies the input data into one of the predefined classes was carried out in this work by the two most popular supervised learning algorithms, an SVM and an ANN. To achieve the best classification results, we utilized the whole range of dataset samples to determine the optimal parameters of both classifiers. The SVM as well as ANN classifiers are designed to work with the same input vector of features that we computed.

TABLE 1 | To show that the medians of the two datasets are different by the two-tailed Mann–Whitney hypothesis test, we employed the methodology of proof by contradiction, where the truth of a statement is determined by assuming that the null hypothesis is false.

Features	Normal cell		Leukemic cell		*U* test
	Median	MAD	Median	MAD	*p*-value
Morphological					
Cytoplasm area	11985.00	4894.53	4022.00	1799.24	< <0.001
Cell area	20011.00	6031.12	16255.00	2830.51	< <0.001
N/C ratio	0.75	0.21	3.15	1.15	< <0.001
Nucleus perimeter	521.00	112.64	412.00	58.51	< <0.001
Nucleus compactness	30.71	14.14	13.13	2.47	< <0.001
Nucleus form factor	0.41	0.19	0.96	0.18	< <0.001
Nucleus elongation	6.97	7.12	1.62	0.31	< <0.001
Nucleus eccentricity	0.49	0.23	0.42	0.18	0.007
Nucleus solidity	0.84	0.09	0.96	0.02	< <0.001
Statistical					
Nucleus energy	0.74	0.05	0.61	0.04	< <0.001
Cell contrast	1.85	0.16	1.53	0.13	< <0.001
Cell entropy	7.37	1.42	5.15	1.20	< <0.001
Nucleus correlation	0.82	0.08	0.89	0.05	< <0.001
Cell dissimilarity	0.56	0.08	0.40	0.07	< <0.001
Cytoplasm mean color	2.34	0.92	0.73	0.34	< <0.001
Nucleus mean color	0.37	0.20	0.57	0.22	< <0.001

In our case, the defined null hypothesis states that there is no significant difference between the observed groups. The selection of a confidence level of 95% therefore signifies that the resulting p-values less than 0.05 are considered statistically significant. This indicates that there is strong evidence against the null hypothesis, as there is less than a 5% likelihood that the null hypothesis is correct.

SVM Model Selection

SVM is a non-linear, non-parametric discriminative classifier based on the Vapnik–Chervonenkis theory. In simple terms, SVM tries to separate the data of unknown samples by finding an optimal line or hyperplane, which represents the largest margin between the classes. In the simplest two-dimensional space, this hyperplane is a line dividing a plane into two parts. Since most of the data cannot be linearly separable in a two-dimensional space, SVM projects these non-linear samples into a higher dimensional feature space by using different kernel functions (Kazemi et al., 2016). Due to this relative flexibility, SVM distinctively affords balanced predictive performance, even in studies with a limited sample size (Pisner and Schnyer, 2020).

To select an appropriate SVM classification model, we tested various kernel functions, including the most frequent linear kernel and a set of non-linear kernels, namely, Gaussian, polynomial, and radial basis function kernels. For each kernel function, we found the maximum value of the accuracy by tuning the SVM parameters using optimization techniques. To evaluate the model's performance, we employed the 10-fold cross-validation methodology, which produced the best out-of-sample estimates with a low bias and modest variance (Bodzas, 2019, p. 59). This approach involved the random division of the dataset into 10 groups called folds of approximately equal size. During the cross-validation process, the first fold is treated as a validation set while the method is fit on the remaining ninefold. The whole cross-validation process is then repeated 10 times, and each fold is used as the validation set once (James et al., 2013). As shown by the experimental results in **Table 2**, the highest classification accuracy was achieved by using the polynomial kernel function.

Neural Network Selection

ANN is a classification technique, that uses several computing units to imitate neurons in the human brain. All units are connected with each other via a weighted link, which determines the prominence of the respective input to the output. Each neuron in a structure performs a weighted sum of all inputs and finds the output using an activation function. This activation function decides whether the

TABLE 2 | Cross validation accuracy of different classification models.

Kernel function	Accuracy [%]
Linear	88.38
Polynomial	98.34
Gaussian	95.02
RBS	97.51

information is relevant or should not pass to the subsequent unit. The whole process of learning is based on altering the values of weights and biases depending on the calculated loss function between the actual and desired output (Zayegh and Bassam, 2018).

Due to the fact that there are no specific guidelines on how to determine the optimal neural network architecture parameters, in particular the number of hidden layers and neurons, we decided to select these parameters through a trial-and-error process. During this process, several architectures with different numbers of neurons and hidden layers were tried experimentally. The number of neural units in the first and last layers depends on the number of given inputs and desired outputs. In this paper, we consider 16 input neurons, where each neuron represents one of the extracted features, and two output neurons, for the leukemic and normal classes. In this phase, we additionally split the dataset into a training and validation set in the conventional ratio of 80:20. To prevent overfitting and concentration of the neural network into one domain, we trained the neural network on randomly chosen samples. Furthermore, we used identical learning rates for each learning cycle and repeated the learning process for 50 and 500 learning iterations for each training image. The overall performance of the particular neural network models is summarized in **Table 3** (Bodzas, 2019, p. 66).

The process of neural network topology verification revealed an increasing accuracy with the number of hidden layers in the case of using 50 learning iterations. We also noticed an increase of the neural network accuracy in architectures with a higher number of neurons in particular layers. On the other hand, training the neural network with a higher number of hidden layers and neurons, and 500 learning iterations, achieved greater precision and ability to classify the data correctly. In particular, the ANN models with a large difference in the number of neurons between consecutive hidden layers reached the highest accuracy, 99.91% (Bodzas, 2019, p. 68).

TABLE 3 | Experimental evaluation of the accuracy of different artificial neural network architectures.

Number of neurons in hidden layers				Accuracy [%] 50 LI*	Accuracy [%] 500 LI*
1st layer	2nd layer	3rd layer	4th layer		
50	–	–	–	92.38	99.58
90	–	–	–	92.14	99.53
100	–	–	–	92.67	99.53
500	–	–	–	90.43	99.32
50	30	–	–	93.14	99.90
70	50	–	–	93.89	99.69
100	100	–	–	91.28	99.69
400	200	–	–	93.56	99.91
200	400	–	–	91.46	98.77
100	100	100	–	91.87	98.52
200	100	200	–	90.91	99.80
500	300	100	–	94.44	99.91
500	400	300	–	92.34	99.49
100	100	100	100	91.17	99.44
700	500	300	100	95.49	99.91

**Learning iterations. Experimental evaluation of the accuracy of different artificial neural network architectures, with highlighted best performing setups.*

Classification Model Implementation

To perform the classification phase, we selected the best-performing models for both classifiers. Before the classification, all computed features were normalized by the min–max algorithm, which mapped the entire range of values to the range <0, 1>. For the binary SVM classification, we selected the C-SVM model, which utilizes a regularization parameter to penalize misclassifications during the separation of the classes. The best results of this classification model were achieved by applying the polynomial kernel function with a gamma value and regularization parameter of 1 and a degree parameter of 5. The tolerance of the maximum gradient of the quadratic function that was used to compute the support vectors was tuned to 0.001. In addition, to improve the functionality of this classification model, we implemented shrinking heuristics, which helped to reduce the number of variables used in the classification computation and therefore accelerated the optimization. The selected ANN model comprised two hidden layers with a descending number of neurons in particular layers (400, 200). The hidden layers of the neural network were fully connected layers without any inner modifications and utilized the sigmoid neuronal function for triggering. The initial weights for the proposed neural network were selected by the Xavier initialization process, which decreases the chance that the gradients will explode or vanish too quickly. The final process of training the architecture was performed by mean-squared error–based back-propagation and a stochastic gradient descent optimizer. Our neural network was trained with 8,333 epochs with a constant learning rate and randomly chosen samples. Moreover, during the learning process, when the measured error rate became saturated, the neural network was iteratively fine-tuned by changing the learning rate from 0.002 to 0.0001.

EXPERIMENTAL VERIFICATION AND RESULTS

In the final analysis, 241 extracted sub-images of 128 normal WBCs and 113 leukemic cells were used to evaluate the proposed system. Since we have to deal with a lack of medical data, we assigned 50 percent of the dataset to the training subset, which was used to build the prediction model, and the remaining fifty percent of the data to test the proposed model. To verify the proportional distribution of specific classes between the training and testing sets, we evaluated the fundamental statistical parameters for the chosen features (see **Table 4**).

Each output of the selected classifier in this work, presents a particular probability, with which the cell belongs to the leukemic and normal class. Since the output probabilities given

TABLE 4 | The separation of the dataset into a training and testing set was performed in a way that ensured the even distribution of the whole range of blood cell types.

Feature	Statistical parameter	Training set	Testing set
Form factor	Number of samples	120	121
	Maximum value	1,21	1,21
	Minimum value	0,19	0,14
	Mean	0,70	0,73
	Standard deviation	0,30	0,32
Contrast	Number of samples	119	120
	Maximum value	3,37	3,35
	Minimum value	2,17	2,13
	Mean	2,67	2,64
	Standard deviation	0,24	0,25

To verify that the split did not affect the statistical distribution, the maximum, minimum, mean, and standard deviation were compared between the two sets. Since all the statistical parameters of the two selected features seem to be well balanced, the final classification should not be burdened with significant errors.

TABLE 5 | Summarization of all correct and incorrect classifications.

	SVM		ANN	
	Disease positive	**Disease negative**	**Disease positive**	**Disease negative**
Test positive	56	3	57	3
Test negative	1	61	0	61
	Overall accuracy: 96.72%		Overall accuracy: 97.52%	

by the SVM model take into account only the probability of the corresponding class, we computed the absolute complement of the outputs to obtain an inversely proportional set. To assess the outputs of both classifiers, the winner-take-all principle was implemented in the last phase. This means that only the classification outputs with the highest score were considered to be the final results. The performance of both algorithms was subsequently estimated by constructing the confusion matrices for both implemented classifiers (see **Table 5**).

Namely, the specificity, sensitivity, accuracy, F1 score and error rate metrics of the proposed strategy were assessed using the following formulas, where TP stands for the number of true positives, TN stands for the number of true negatives and FP and FN denote the numbers of first and second error types (false positives and false negatives, respectively) (Chen et al., 2019).

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FN} + \mathrm{FP}} \tag{13}$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{15}$$

$$F_1 = \frac{\mathrm{Sensitivity} * \mathrm{Specificity}}{\mathrm{Sensitivity} + \mathrm{Specificity}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{16}$$

$$ERR = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{17}$$

The sensitivity and specificity represent warnings from two different standpoints. Whereas sensitivity indicates how often positive predictions are correct, specificity denotes the percentage of successful negative predictions. In the medical field, reaching 100% specificity is not reasonable. This value of this type of measure is reached in medical practice by the assumption that no patients have a positive diagnosis and that therefore, the test will never make an FN error. However, high values of specificity are required in cases where the main goal is to limit the number of false negatives. To achieve a better overview of diagnostic efficiency, we took into account the F1 score metric, which combines both sensitivity and specificity (Tharwat, 2018). **Table 6** shows the comparison of the implemented classifiers in terms of their prediction performance (Bodzas, 2019, p. 70).

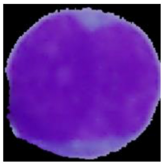
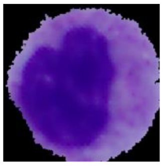
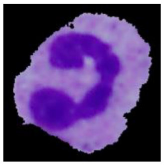
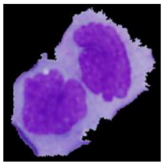
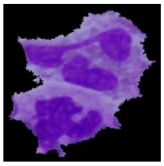
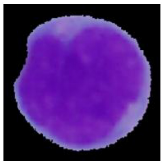
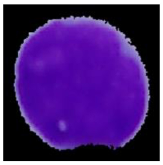
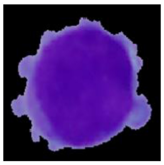
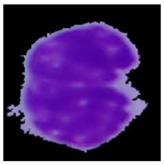
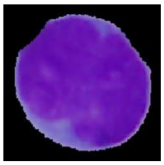
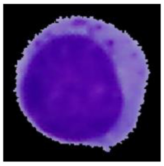
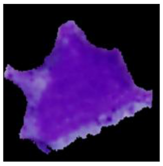
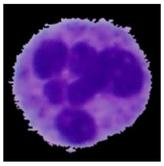
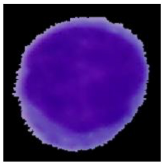
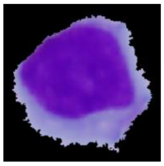
Examples of specific classification results highlighting all incorrectly classified cells are presented in **Table 7**. Two cases of incorrect classifications were caused by a flawed segmentation phase (incorrectly classified cells D and E). Nevertheless, the ANN, due to its ability to accept relatively small errors, identified one of those cells correctly with an accuracy of 98.19%. Even though the ANN proved to have a better performance, in the case of the incorrectly classified cell C, we notice overfitting, which is the major drawback of this methodology. On the contrary, overfitting is not seen in the results obtained by the SVM algorithm, which achieved better identification results in this sample. The main reason lies in the evenly distributed portions of similar cells among the learning and training sets and the small degree parameter, which decreased the flexibility of the decision boundary and therefore prevented overfitting. Other practical problems are often caused by missing image samples in the datasets. Such missing samples in the training set are sometimes indispensable for making correct predictions. This can be seen in case B among the incorrect classifications, where the lack of banded neutrophils resulted in an accuracy of 0% for both classifiers. Whereas all incorrect ANN classifications were related to the first kind of error, of predicting a positive diagnosis when the actual condition was negative, the SVM in one sample (A) resulted in the worst-case scenario (a type II error) by predicting disease absence.

It should also be noted that even though the remaining cells were classified correctly, some results do not achieve a classification probability higher than 95%, and therefore, there

TABLE 6 | Performance measures for selected supervised classifiers.

	Accuracy	Sensitivity	Specificity	F_1	Error rate
SVM	96.72	98.25	95.31	96.55	3.28
ANN	97.52	100.00	95.31	97.44	2.48

TABLE 7 | The classification probabilities of selected samples.

	A	B	C	D	E
Incorrectly classified cells					
Type	**Blast cell**	**Normal WBC**	**Normal WBC**	**Normal WBC**	**Normal WBC**
SVM accuracy	**24.15**	**0.00**	95.90	**0.00**	**21.61**
ANN accuracy	96.87	**0.01**	**24.13**	**0.19**	98.19
Blast cells					
SVM accuracy	93.74	98.80	94.29	100.00	91.47
ANN accuracy	99.90	99.88	78.09	100.00	99.94
Normal cells					
SVM accuracy	85.63	87.13	90.30	83.09	97.98
ANN accuracy	98.03	98.57	58.48	85.45	100.00

The first four rows (A–E) show examples of all incorrectly classified samples with the false positive and false negative classifications highlighted, and the rest of the rows (A–E) show the probability results of the selected examples of correct classifications.

is a high probability of the presence of overfitted areas in the vicinity of these cells.

CONCLUSION AND FUTURE PROSPECTS

In this work, we propose a method for the automated identification and classification of blast cells from microscopic peripheral blood smear images. This study introduces a novel combination of image processing methodologies and proposes extensive pre-processing to achieve high classification accuracy. In particular, the selected combination of 16 features carrying morphological and statistical information demonstrated an excellent ability to distinguish between cancerous and non-cancerous blood cells. We selected most of the features on the basis of their similarity with the visual information, on which the domain experts focus during manual examination. These features were extracted from 241 WBCs segmented from 31 peripheral blood smear images from a local dataset. To perform the classification, we selected the two most popular classifiers in the literature, the ANN and the SVM algorithm. The neural network model yielded better results, reaching a sensitivity of 100% and an overall accuracy of 97.52%. Unlike previous studies, we also presented some of the specific classification probabilities of the correctly identified cells and conducted a reverse analysis to identify the pivotal classification failures. These observations indicated that even when the published accuracies reach the highest values, a classification method may not provide clarity or sufficiently high reliability, and therefore, further examination is required.

One of the greatest problems we encountered was a lack of medical data and extensive datasets. In particular, expanding the learning set of the data would reduce overfitting and increase the probability of particular classifications. Moreover, the classification errors caused by incomplete datasets with missing cell samples would be suppressed. It should be noted that many authors have verified their proposed systems by employing small local and publicly unavailable datasets. Due to this fact, it was impossible to compare our findings with the results obtained by the previously proposed algorithms. Furthermore, this has a negative impact on the possibility of reproducing recent trends and converging toward better technical solutions. The results obtained in this work indicate that future research should be mainly devoted to the development of a more robust segmentation algorithm with the possibility of adaptive parameter adjustment, which would unify the functionality of the system under diverse conditions. Moreover, researchers should focus on improving particular classification probabilities and minimizing false negative classifications. Such a system could be then used as a medical support tool that would facilitate manual examination and save tremendous time. Using the results of particular classifications with a defined high decision limit will allow us to achieve higher identification reliability. Nevertheless, cells with lower probability should be still verified by hematological specialists.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of FN Ostrava, University Hospital Ostrava. The participants provided written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AB conceived and designed the study and drafted the manuscript. JZ coordinated the study and provided useful suggestions. PK performed searches, analyses, interpretations, and edited the manuscript. AB and PK developed machine learning algorithms. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the European Regional Development Fund in the Research Centre of Advanced Mechatronic Systems project, Project Number CZ.02.1.01/0.0/0.0/16_019/0000867 within the Operational Program Research, Development and Education. This work was also supported by the Student Grant System of VSB Technical University of Ostrava, grant number SP2020/151.

REFERENCES

Ahmed, A. S., Morsy, M., and Abou-Elsoud, M. E. A. (2016). Microscopic digital image segmentation and feature extraction of acute Leukemia. *Int. J. Sci. Eng. Appl.* 5, 228–233. doi: 10.7753/IJSEA0505.1001

Aljaboriy, S., Sjarif, N., and Chuprat, S. (2019). Segmentation and detection of acute leukemia using image processing and machine learning techniques: a review. *AUS* 26, 511–531. doi: 10.4206/aus.2019.n26.2.60

Alsalem, M. A., Zaidan, A. A., Zaidan, B. B., Hashim, M., Madhloom, H. T., Azeez, N. D., et al. (2018). A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations. *Comput. Methods Programs* 158, 93–112.

American Dental Association [ADA] (2012). *The ADA Practical Guide to Patients with Medical Conditions*, ed. L. L. Patton (New York, NY: Wiley).

Bagasjvara, R. G., Candradewi, I., Hartati, S., and Harjoko, A. (2016). Automated detection and classification techniques of Acute leukemia using image processing: A review. *Paper Presented at the 2nd International Conference on Science and Technology-Compute*, Yogyakarta. 35–43. doi: 10.1109/ICSTC.2016.7877344

Batchelor, B. G., and Waltz, F. M. (2001). *Intelligent machine vision: techniques, implementations, and applications*. New York, NY: Springer.

Bodzas, A. (2019). *Diagnosis of Malignant Haematopoietic Diseases based on the Automation of Blood Microscopic Image Analysis*. Master's thesis, Technical University of Ostrava, Ostrava, CZ.

Chan, Y. K., Tsai, M. H., Huang, D. C. H., Zheng, Z. H., and Hung, K. D. (2010). Leukocyte nucleus segmentation and nucleus lobe counting. *BMC Bioinformatics* 11:558. doi: 10.1186/1471-2105-11-558

Chen, J., Ying, H., Liu, X., Gu, J., Feng, R., Chen, T., et al. (2020). A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (in press). doi: 10.1109/TCBB.2020.2991173

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., et al. (2019). Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine. *IEEE Access.* 7, 150960–150968. doi: 10.1109/ACCESS.2019.2946980

Chiaretti, S., Zini, G., and Bassan, R. (2014). Diagnosis and Subclassification of Acute Lymphoblastic Leukemia. *Mediter. J. Hematol. Infect. Dis.* 6:e2014073. doi: 10.4084/MJHID.2014.073

Daniels, R., and Nicoll, L. H. (2012). *Contemporary Medical Surgical Nursing*, 2nd Edn. New York, NY: Cengage Learning.

Díaz, G., and Manzanera, A. (2009). "Automatic Analysis of Microscopic Images in Hematological Cytology Applications," in *Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques*, eds F. A. González and E. Romero (Landisville, PA: Yurchak Printing Inc), 167–196.

Fairchild, M. D. (2005). *Color Appearance Models*, 2nd Edn. Chichester: John Wiley & Sons.

Gao, W., Zhu, Y., Zhang, W., Zhang, K., and Gao, H. (2019). A hierarchical recurrent approach to predict scene graphs from a visualion-oriented perspective. *Comput. Intellig.* 35, 496–516. doi: 10.1111/coin.12202

Hariprasath, S., Dharani, T., and Santh, M. (2019). Detection of acute lymphocytic leukemia using statistical features. *Paper Presented at the 4th International Conference on Current Research in Engineering Science and Technology*, Trichy. Available online at: http://www.internationaljournalssrg.org/uploads/specialissuepdf/ICCREST/2019/ECE/IJECE-ICCREST-P102-JRCE1119.pdf

Inaba, H., Greaves, M., and Mullighan, C. G. (2013). Acute lymphoblastic leukaemia. *Lancet* 381, 1943–1955. doi: 10.1016/S0140-6736(12)62187-4

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* New York, NY: Springer.

Joshi, M. D., Karode, A. H., and Suralkar, S. R. (2013). White blood cells segmentation and classification to detect acute Leukemia. *Int. J. Emerg. Trends Technol. Computer Sci.* 2, 147–151.

Katz, A. J., Chia, V. M., and Schoonen, W. M. (2015). Acute lymphoblastic leukemia: an assessment of international incidence, survival, and disease burden. *Cancer Causes Control* 26, 1627–1642. doi: 10.1007/s10552-015-0657-6

Kazemi, F., Najafabadi, T., and Araabi, B. (2016). Automatic Recognition of Acute Myelogenous Leukemia in Blood Microscopic Images Using K-means Clustering and Support Vector Machine. *J. Med. Signals Sens.* 6, 183–193.

Manisha, P. (2012). Leukemia: a review article. *Int. J. Adv. Res. Pharm. Bio Sci.* 2, 397–407.

Moradiamin, M., Samadzadehaghdam, N., Kermani, S., and Talebi, A. (2015). Enhanced recognition of acute lymphoblastic leukemia cells in microscopic images based on feature reduction using principle component analysis. *Front. Biomed. Technol.* 2:128–136.

Nailon, W. H. (2010). "Texture analysis methods for medical image characterisation," in *Biomedical Imaging*, ed. Y. Mao (London: Intech Publishing), 75–100.

Neoh, S., Srisukkham, W., Zhang, L., Todryk, S., Greystoke, B., Lim, C., et al. (2015). An intelligent decision support system for leukaemia diagnosis using microscopic blood images. *Sci. Rep.* 5:14938. doi: 10.1038/srep14938

Pisner, D. A., and Schnyer, D. M. (2020). "Chapter 6 - Support vector machine," in *Machine Learning*, eds A. Mechelli and S. Vieira (Cambridge, MA: Academic Press), 101–121.

Putzu, L., Caocci, G., and Di Ruberto, C. (2014). Leucocyte classification for leukaemia detection using image processing techniques. *Artif. Intellig. Med.* 62, 179–191. doi: 10.1016/j.artmed.2014.09.002

Rogers, K. (Ed.) (2011). *Blood: Physiology and Circulation*. Edinburgh: Britannica Educational Publishing.

Rose, M. (2013). *Oncology in Primary Care*, 1st Edn. Philadelphia: Lippincott Williams & Wilkins.

Serfontein, W. (2011). *Cancer Diagnosed: What Now?* 2nd Edn. Bloomington: Xlibris.

Shafique, S., and Thesin, S. (2018). Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technol. Cancer Res. Treatment* 17:1533033818802789. doi: 10.1177/1533033818802789

Tharwat, A. (2018). Deep belief networks and cortical algorithms: A comparative study for supervised classification. *Appl. Comput. Inform.* 15, 81–93. doi: 10.1016/j.aci.2018.08.003

Wan, S., and Mak, M. W. (2015). *Machine Learning for Protein Subcellular Localization Prediction*. Boston: De Gruyter.

Wang, M., Zhou, X., Li, F., Huckins, J., King, R., and Wong, S. (2008). Novel Cell Segmentation and Online SVM for Cell Cycle Phase Identification in Automated Microscopy. *Bioinformatics.* 24, 94–101. doi: 10.1093/bioinformatics/btm530

Wang, Q., Bi, S., Sun, M., Wang, Y., Wang, D., and Yang, S. (2019). Deep learning approach to peripheral leukocyte recognition. *PLoS ONE.* 14: e0218808. doi: 10.1371/journal.pone.0218808

Wiernik, P. H. (2001). *Adult Leukemia (Atlas of Clinical Oncology)*. Hamilton: BC Decker Inc.

Zayegh, A., and Bassam, N. (2018). "Neural Network Principles and Applications," in *Digital Systems*, Ed. R. J. Tocci (London: Pearson), doi: 10.5772/intechopen.80416

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

frontiers
in Bioengineering and Biotechnology

REVIEW
published: 04 September 2020
doi: 10.3389/fbioe.2020.01032

Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA

Aimin Yang[1], *Wei Zhang*[1*†], *Jiahao Wang*[1], *Ke Yang*[2], *Yang Han*[1*†] and *Limin Zhang*[3*†]

[1] College of Science, North China University of Science and Technology, Tangshan, China, [2] College of Yi Sheng, North China University of Science and Technology, Tangshan, China, [3] Mathmatics and Computer Department, Hengshui University, Hengshui, China

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Wei Wang,
University of Macau, China
Gunasekaran Manogaran,
University of California, Davis, United States
Syed Hassan Ahmed,
JMA Wireless, United States

***Correspondence:**
Wei Zhang
zhangwei_ncst@163.com
Yang Han
hanyang@ncst.edu.cn
Limin Zhang
limin_zhang@yeah.net

† These authors have contributed equally to this work

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

Received: *22 May 2020*
Accepted: *10 August 2020*
Published: *04 September 2020*

Citation:
Yang A, Zhang W, Wang J, Yang K, Han Y and Zhang L (2020) Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Front. Bioeng. Biotechnol. 8:1032. doi: 10.3389/fbioe.2020.01032

Deoxyribonucleic acid (DNA) is a biological macromolecule. Its main function is information storage. At present, the advancement of sequencing technology had caused DNA sequence data to grow at an explosive rate, which has also pushed the study of DNA sequences in the wave of big data. Moreover, machine learning is a powerful technique for analyzing largescale data and learns spontaneously to gain knowledge. It has been widely used in DNA sequence data analysis and obtained a lot of research achievements. Firstly, the review introduces the development process of sequencing technology, expounds on the concept of DNA sequence data structure and sequence similarity. Then we analyze the basic process of data mining, summary several major machine learning algorithms, and put forward the challenges faced by machine learning algorithms in the mining of biological sequence data and possible solutions in the future. Then we review four typical applications of machine learning in DNA sequence data: DNA sequence alignment, DNA sequence classification, DNA sequence clustering, and DNA pattern mining. We analyze their corresponding biological application background and significance, and systematically summarized the development and potential problems in the field of DNA sequence data mining in recent years. Finally, we summarize the content of the review and look into the future of some research directions for the next step.

Keywords: DNA sequence, machine learning, data mining, DNA sequence alignment, DNA sequence classification, DNA sequence clustering, DNA pattern mining

INTRODUCTION

We live in the era of the genome, advances in science have allowed humans to spy on the mysteries of life. In recent decades, the rapid expansion of biological data is a significant feature of the development of molecular biology, and a massive biological information database has rapidly formed. We must obtain useful knowledge from these huge data, and simultaneously bioinformatics was born. Bioinformatics is an interdisciplinary subject. It comprehensively uses mathematics, life sciences, and computer science to mine biological information in biological data (Chu, 2014), and further guides the relevant researches of biological researchers.

Specifically, the first step is to obtain information on the protein-coding region by analyzing the genomic DNA sequence. Then simulating and predicting the spatial structure of the protein. Finally, according to the function of the protein, the researchers make the necessary drug design.

According to statistics, the amount of biological data approximately doubles every 18 months. In 1982, GenBank's first nucleic acid sequence database had only 606 sequences, containing 680,000 nucleotide bases (Bilofsky et al., 1986). As of February 2013, its database already contains 162 million biological sequence data, containing 150 billion nucleotide bases. How to mine knowledge from these huge data and guide biological research s an important research content of bioinformatics.

For complex biological data, on the one hand, it is necessary to solve the problem of storage and management of massive data, and on the one hand, it is necessary to extract effective information from the data on the premise of ensuring that the data reflects the true meaning of biology. Machine learning is an important method to achieve artificial intelligence. It can handle the automatic learning of machines without explicit programming and has been widely used in the field of bioinformatics (Li et al., 2005; Larranaga et al., 2006).

DNA is a kind of biomacromolecule in organisms. It carries the genetic information of life and guides the development of biological development and the functioning of life functions. At present, machine learning has been widely used in sequence data analysis and has very broad application prospects in improving data processing capabilities and generating valuable biological information. The review focuses on DNA sequence data mining and machine learning. The review briefly introduces the development process of sequencing technology, DNA sequence data structure, and several sequence encoding methods in machine learning. And we clarify that sequence similarity is the basis of DNA sequence data mining. We have comprehensively analyzed the basic process of data mining and summarized the algorithms commonly used in machine learning. Then, we summarized four typical applications of machine learning in DNA sequence data: DNA sequence alignment, classification, clustering, and pattern mining. In summary, we have the following conclusions: distributed sequence alignment and parallel computing may be the research focus of DNA sequence alignment. How to effectively express sequence features and analyze DNA sequence classification is a difficult point in research. The two key points of DNA sequence clustering are how to extract characteristic subsequences in the DNA sequence. DNA sequence pattern mining will generate an explosion of candidate sequence patterns, which will consume a lot of time and space. How to design a suitable search strategy and eliminate redundant sequence patterns will be an important direction for future research.

BASIC KNOWLEDGE OF DNA

Gene sequencing is one of the most popular technologies in life sciences. At present, HiSeq X Ten is the sequencing platform with the highest sequencing throughput and the lowest cost. The introduction of equipment and its commercialization has greatly promoted the development of the sequencing industry. The rapid progress of sequencing technology and the continuous decline of sequencing costs have made sequencing more and more common.

Sequence similarity is the basis of sequence data mining, and it is a research direction where sequence similarity bioinformatics is very meaningful. Sequence similarity refers to the degree of similarity between sequences. If the similarity between two sequences exceeds 30%, it is considered that they may have homology. The homologous sequences have a common evolutionary ancestor, and their structures and functions may have similarities.

Development of Sequencing Technology

With the development of biological information technology, sequencing technology has experienced three stages of development. The chain termination method proposed by Sanger and the chain degradation method proposed by Gibert is collectively called the first generation sequencing technology. At present, Sanger sequencing is still widely used in conventional sequencing applications and verification, but the sequencing cost is extremely high and the throughput is low, which seriously affects its truly large-scale application. After more than 40 years of technological development, sequencing technology has achieved considerable progress. The progress of sequencing technology is shown in **Figure 1**.

After the continuous efforts of researchers, the second generation sequencing technology marked by 454 technology was born in 2005. These sequencing systems can analyze billions of sequencing reactions at the same time. The second-generation sequencing technology is a kind of connected sequencing, which greatly improves the speed of sequencing and greatly reduces the cost of sequencing. At present, the second-generation sequencing technology (Watson, 2014) is the main force in the scientific research market. Due to its low cost, it has been widely used.

In 2011, the third generation sequencing technology represented by Oxford single molecule sequencing technology and PacBio's SMRT technology was born. Single molecule sequencing is the biggest feature of the third-generation sequencing technology. This technology needs to be continuously adjusted and upgraded for large-scale applications. Sequencing technology is revolutionizing personalized medicine by providing high throughput options with sequence capabilities for clinical diagnosis.

Genomic big data analysis is becoming the next frontier in the field of biomedicine (Roukos, 2010), which integrates data storage, data sharing, data analysis, and data quality control. The sequencing error rate profiles of different sequencing platforms are different, so we need to know which sequencing platform is used to generate the original data, what is their error rate distribution, and whether there are certain biases and limitations. At present, the three major international biological data centers (NCBI, EBI, and DDBJ) have established a series of biological information databases and various data services, which provide strong support for biological data analysis. Biomedical data

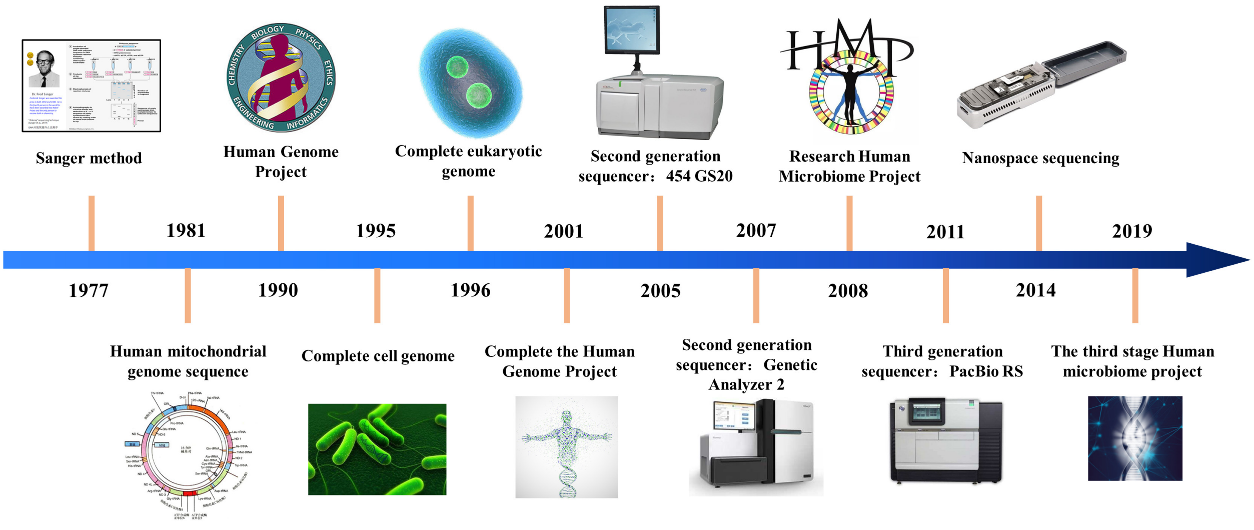

Sanger method
Human Genome Project
Complete eukaryotic genome
Second generation sequencer：454 GS20
Research Human Microbiome Project
Nanospace sequencing
1981
1995
2001
2007
2011
2019
1977
1990
1996
2005
2008
2014
Human mitochondrial genome sequence
Complete cell genome
Complete the Human Genome Project
Second generation sequencer：Genetic Analyzer 2
Third generation sequencer：PacBio RS
The third stage Human microbiome project

FIGURE 1 | History of sequencing technology.

presents the characteristics of a wide variety, high-dimensional complex internal structure, rich content, relatively scattered data, and difficulty in high-dimensional multi-level cross-sharing.

Data Structure of DNA Sequence

Biological studies have shown that biological sequences are not random and unordered strings. They consist of a linear arrangement of smaller elements. The DNA sequence is connected by four kinds of deoxyribonucleotides (bases). Base order contributes to the diversity of DNA molecules.

The structure of the DNA double helix is shown in **Figure 2**. The nitrogen-containing bases of one strand of the DNA double helix structure will only bond with specific bases of the other strand. It is generally called complementary base pairing, and a base pair is the basic unit of DNA sequence.

DNA sequence data have different characteristics from other data, mainly including:

1. DNA sequence data consists of non-numeric (A, T, C, G) characters;
2. The length of different sequences varies greatly. Some sequences have only a few dozen characters, while others are very long, up to hundreds of megabytes;
3. DNA sequence data contains its specific biological significance;
4. Due to certain errors in the sequencing process and noise in the sequence data, it is necessary to perform corresponding data preprocessing before analyzing the data.

DNA Sequence Coding

When processing the DNA sequence, it is necessary to convert the string sequence into a numerical value, so as to form a matrix input model training. Generally speaking, there are three methods for sequence encoding: sequential encoding, one-hot encoding, and k-mer encoding (Choong and Lee, 2017). The characteristics of the three DNA encoding methods are shown in **Table 1**. The performance of sequential encoding is comparable to one-hot encoding, but the training time is significantly reduced. One-hot encoding is widely used in deep learning methods and is very suitable for algorithms such as CNN (convolutional neural networks). In addition, the performance of one-hot encoding is quite consistent in different data sets, but a suitable CNN is required to get good performance. Ordinal codes represented by matrices perform best in some evaluation data sets. The performance of CNN in discovering DNA motifs depends on the proper design of sequence encoding and representation. The good performance of the ordinal coding method shows that there is still room for improvement in the single-point coding method.

DNA Sequence Similarity

The main mining modes of machine learning include data characterization and differentiation, data frequent patterns, association and correlation, classification and regression of data predictive analysis, cluster analysis, and outlier analysis. Data mining for DNA sequences is generally carried out from these aspects, and research in these areas is inseparable from similarity analysis between sequences (Pearson, 2013). It can be seen that sequence similarity is the basis of DNA sequence data mining.

Sequence similarity means that there are similar or identical sites between sequences. The sequence similarity can be a quantitative value or a qualitative description. If the degree of similarity between two sequences exceeds 30%, It is considered that the two sequences have a homologous relationship. Therefore, if the two sequences are highly similar, the two sequences are likely to have a common evolutionary ancestor. At the same time, if a sequence similar to the unknown sequence can be found from the sequences with known functions, we can further predict the function (Rogozin et al., 1996) of the unknown sequence.

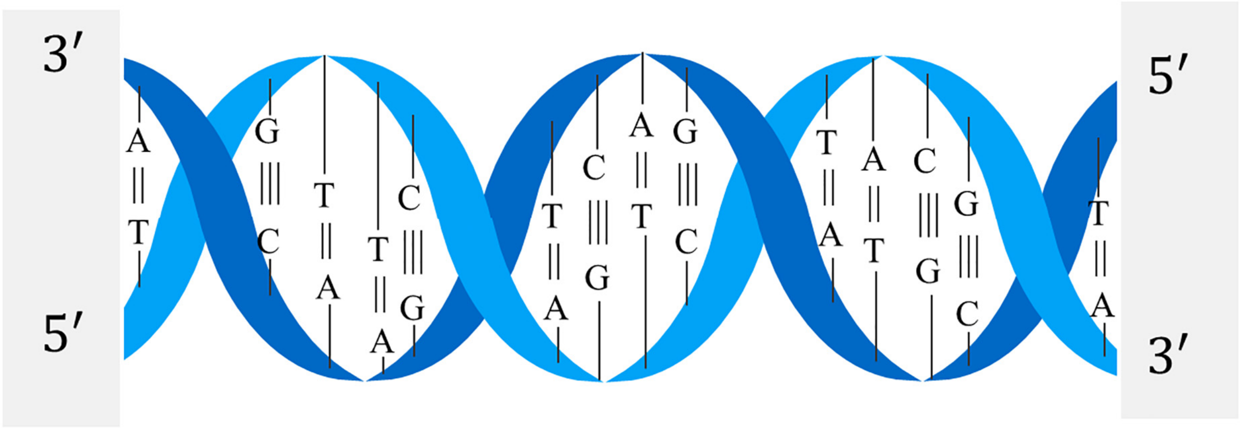


FIGURE 2 | Double helix of DNA.

One of the main problems of DNA sequence similarity research is to search for sequences whose similarity to a specified sequence exceeds a certain threshold. The most commonly used method is to establish a similarity matrix (Henikoff and Henikoff, 1992) and find the best match between sequences in consideration of possible insertions, deletions, and mutations. The study of sequence similarity is divided into global similarity research and local similarity research. The global similarity is the similarity matching of the entire sequence, which is suitable for sequences with a high degree of similarity at the global level. The Needleman-Wunsch algorithm is a typical sequence alignment algorithm (Pearson and Lipman, 1988). However, genes only account for about 2% of the DNA sequence, that is, only a few sequence fragments have a functional role. Although there is no similarity between sequences as a whole, there are similarities in some local areas. Therefore, it is more meaningful to study local similarity than global similarity. Typical local alignment algorithms include the Smith-Waterman algorithm based on dynamic programming algorithm and heuristic database similarity search algorithms FASTA and BLAST (basic local alignment search tool). In a recent study, Delibas and Arslan (2020) proposed a non-aligned sequence similarity analysis method, a new method of DNA sequence similarity analysis using the similarity calculation of texture images, which is a digital image processing method.

TABLE 1 | Common ways of encoding DNA sequences.

Encoding method	Features
Sequential encoding	This method encodes each base as a number. For example, change [A,T,G,C] to [0.25, 0.5, 0.75, 1.0], and any other character can be recorded as zero.
One-hot encoding	This method is widely used in deep learning methods. For example, [A,T,G,C] will become [0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0,0,0]. These coded vectors can be connected or turned into a two-dimensional array.
K-mer encoding	First take a longer biological sequence and decompose it into k-length overlapping fragments. For example, if we use a segment of length 6, "ATGCATGCA" will become: "ATGCAT," "TGCATG," "GCATGC," "CATGCA."

Sequence similarity is one of the key processes of DNA sequence analysis in computational biology and bioinformatics. In the study of gene function analysis, protein structure prediction and sequence retrieval, similarity calculations are required. We select the appropriate sequence similarity analysis method and improve it according to actual application requirements and biological background. This is the basis and key of DNA sequence data mining.

MACHINE LEARNING ALGORITHM

In the past few decades, we have witnessed the revolutionary development of biomedical research and biotechnology and the explosive growth of biomedical data. The problem has changed from the accumulation of biomedical data to how to mine useful knowledge from the data. On the one hand, the rapid development of biotechnology and biological data analysis methods has led to the emergence of a challenging new field: bioinformatics. On the other hand, the continuous development of biological data mining technology has produced a large number of effective and well-scalable algorithms. How to build a bridge between the two fields of machine learning and bioinformatics to successfully analyze biomedical data is worthy of attention and research. In particular, we should analyze how to use data mining for effective biomedical data analysis, and outline some research questions that may stimulate the further development of powerful biological machine learning algorithms.

Basic Process of Data Mining

Data mining is a discipline that combines classic statistical tools with computer science algorithms. This discipline aims to mine knowledge from large amounts of data for scientific, computational, or industrial use. As shown in **Figure 3**, we comprehensively describe the process of data mining from six aspects.

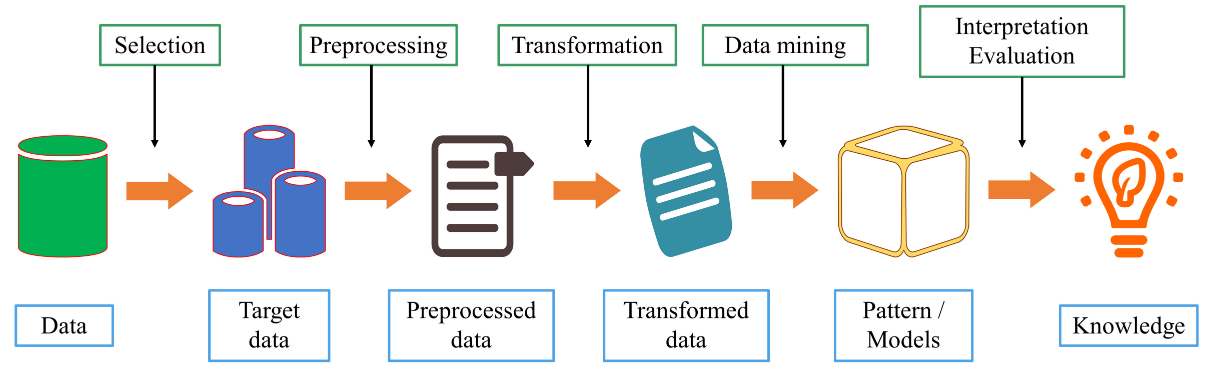


FIGURE 3 | The steps for data mining process.

1. Data cleaning. Because of the increasing amount of heterogeneous data, data sets often have missing data and inconsistent data. Low data quality will have a serious negative impact on the information extraction process. Therefore, deleting incomplete, or inconsistent data is the first step in data mining;
2. Data integration. If the source of the data to be studied is different, it must be aggregated consistently;
3. Data selection. Accurately select relevant data based on the research content;
4. Data conversion. Transform or merge data into a form suitable for mining, and integrate new attributes or functions useful for the data mining process;
5. Data mining. Select the appropriate model according to the problem and make subsequent improvements;
6. Mode evaluation. After acquiring knowledge from the data, select appropriate indicators to evaluate the model.

The main task of the data mining step is to correctly select one or a combination of these steps and find an effective and reliable method to solve the given problem. In recent years, machine learning has been widely used in bioinformatics analysis. Each step of data mining is developed independently of other steps, and each step has a large number of machine learning algorithms.

Association Rule Mining Algorithm

As one of the most important branches of data mining, association rule mining can identify the associations and frequent patterns of a set of items in a given database. It consists of two sub-problems: (1) Set the minimum support threshold and use the minimum support Find frequent itemsets from the database; (2) Use minimum confidence to find association rules that satisfy specified constraints on frequent itemsets. Association rule mining not only plays an important role in business data analysis but has also been successful in many other fields, such as virtual shopping basket analysis and medical data analysis.

The *Apriori* algorithm is a typical association rule-based mining algorithm, which has applications in sequence pattern mining and protein structure prediction. Many machine learning algorithms in data mining are derived based on *Apriori* (Zhang et al., 2014). The basic method of association rule mining is through the use of Some metrics are used to analyze the strong associations in the database. The most commonly used measurement methods are minimum support and minimum confidence. The *Apriori* algorithm uses a guided method to mine association rules between data items in the database.

Classification Algorithm

Classification is one of the most studied tasks in machine learning. The principle of classification is based on the predicted attribute to predict the class of the target attribute specified by the user. In genomics, the key issues are genome classification and sequence annotation. In the mining of biological sequences, widely used algorithms include fuzzy sets, neural networks, genetic algorithms, and rough sets. There are also many general classification models, such as naive Bayesian networks, decision trees, neural networks, and rule learning using evolutionary algorithms.

Clustering Algorithm

The clustering algorithm in machine learning can cluster together sequences with some same characteristics, and explore the effective information of unknown sequences from known functions and structures. Therefore, the clustering of biological sequences is of great significance to the research of bioinformatics. The difference from the classification is that clustering does not implement a set category. Each cluster has its own common characteristics. The purpose of cluster analysis is to divide the data with common characteristics into one category, then use other methods to analyze the data.

In recent years, with the development of artificial intelligence, the clustering algorithm has become a popular research direction in the field of machine learning. To improve the processing capacity of large scale data, domestic and foreign scholars have conducted more in-depth research on clustering algorithms. Several excellent clustering algorithms have emerged: there are mainly clustering algorithms based on granularity, clustering algorithms based on uncertainty, clustering algorithms based on entropy, clustering integration algorithms, etc.

Besides the above-mentioned ones, there are a large number of algorithms. Each algorithm has its characteristics, an algorithm cannot be applied in all situations. Understand the advantages and disadvantages of each algorithm could help us better use and research.

Challenges and Future Solutions

Machine learning is the core of data mining and the most widely used data processing method. A key advantage of machine learning algorithms is that they can be used to filter large amounts of data to explore patterns that may be overlooked. In the era of big data in biomedical research, machine learning plays a key role in discovering predictable patterns in biological systems. The current application of machine learning in biomedical data mainly has the following problems:

1. Large data sets are the key to machine learning. At present, the magnitude of most biological data sets is still too small to meet the requirements of machine learning algorithms. Although the total amount of biological data is huge and increasing day by day, the collection of data comes from different platforms. Due to the differences in technology and biology itself, it is very difficult to integrate different data sets;
2. Due to the differences in biological data itself, machine learning models trained on one data set may not be well generalized to other data sets. If the new data is significantly different from the training data, the analysis results of the machine learning model are likely to be false;
3. The black-box nature of machine learning models brings new challenges to biological applications. It is usually very difficult to interpret the output of a given model from a biological point of view, which limits the application of the model.

Machine learning presents new opportunities and challenges to the development of life sciences. In response to the above issues, we believe that future research directions should include the following:

1. The first is to collect large and well-annotated data sets;
2. A certain machine learning model cannot apply to all data sets, so any new data set should match the general attributes of the data used to train the model;
3. We urgently need to develop a means to transform the "black box" of machine learning into a biologically meaningful and interpretable "white box."

There are many opportunities at the intersection of machine learning and biomedical data integration, but there are also huge challenges to overcome. Machine learning itself is far from realizing its potential in the field of biological research, and we still have a long way to go.

APPLICATION OF MACHINE LEARNING IN DNA SEQUENCE DATA MINING

Machine learning is an important branch of computer science. On the one hand, machine learning makes it possible to mine useful knowledge from large data sets. On the other hand, many areas are also eager to obtain knowledge from data to guide practice. Machine learning also provides new opportunities and challenges for the development of these areas. The benign interaction brought about by this interdisciplinary integration has undoubtedly promoted the development and prosperity of machine learning.

DNA is a biological macromolecule and the basic unit of biological genetic material. Its main function is the storage of genetic information. The calculation and analysis of DNA sequences had undergone fundamental changes in the 1980s. As the genome sequencing system continues to develop, the study of DNA sequences has gradually shifted from the accumulation of original data onto the interpretation of data. This section summarizes the four applications of machine learning in DNA sequence data: DNA sequence alignment, classification, clustering, and pattern mining, and analyzes and discusses the corresponding biological application background and significance. Finally, we systematically summarize the research in the field of machine learning in recent years.

DNA Sequence Alignment

Sequence alignment is the comparison of two or more sequences in the order of base arrangement, mainly to compare sequences with unknown functions to sequences with known sequences. And the results of the alignment reflect the similarity between sequences and their biology Features. Sequence alignment analysis is one of the most basic and important issues in bioinformatics. Through sequence alignment analysis, the structure and function of biological sequences can be further predicted. According to the study of biology, the evolution of DNA has the possibility of gene recombination and mutation, and the evolutionary process of DNA has been unable to recover and reproduce. However, evolution can be studied to explore the homology between DNA through sequence alignment analysis.

Sequence alignment can be divided into double sequence alignment and multi-sequence alignment. Multi-sequence alignment is an extension of double sequence alignment. As the number of sequence alignments increases, the difficulty of alignment is also greater. At present, the research of biological sequence alignment is very mature, and a large number of sequence alignment tools have appeared, such as CLUSTAL, TCOFFEE, and MUSCLE. We selected three DNA sequences of equal length and used CLUSTAL software for sequence comparison. The local visualization of the comparison results is shown in **Figure 4**. The red area indicates the part of the three sequences that are completely matched. The number of completely matched bases in the figure is 25. The number of bases in the sequence fragment is 46. The sequence similarity reaches 54.35%, and it can be considered that the three sequences have local similarities. **Figure 4** is just the simplest comparison situation. In the actual sequence comparison, the situation is much more complicated.

At the early stages, research on biological sequence alignment started with dual sequence alignment. Needleman and Wunsch used dynamic programming algorithms for dual sequence alignment based on the similarity of the entire sequence,

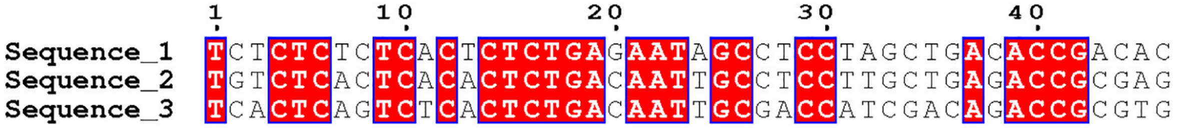

1
10
20
30
40
Sequence_1 TCTCTCTCTCACTCTCTGAGAATAGCCTCCTAGCTGACACCGACAC
Sequence_2 TGTCTCACTCACACTCTGACAATTGCCTCCTTGCTGAGACCGCGAG
Sequence_3 TCACTCAGTCTCACTCTGACAATTGCGACCATCGACAGACCGCGTG

FIGURE 4 | DNA sequence fragment alignment diagram.

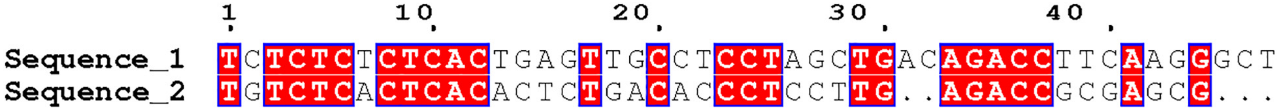

1
10
20
30
40
Sequence_1 TCTCTCTCTCACTGAGTTGCCTCCTAGCTGACAGACCTTCAAGGGCT
Sequence_2 TGTCTCACTCACACTCTGACACCCTCCTTG..AGACCGCGAGCG...

FIGURE 5 | Non-isometric DNA sequence alignment diagram.

this is the Needleman-Wunsch algorithm commonly used in sequence alignment, which is also known as global sequence comparison method and optimization matching algorithm. Smith and Waterman (1981) improved the dynamic programming algorithm to make it into a local optimal algorithm, which can search for sequence fragments with the high local similarity between two sequences. The disadvantage of the Smith-Waterman algorithm is that the comparison speed is slow. If you want to search for the maximum matching base number of two DNA sequences, you need to find the longest common substring of the two sequences. First, calculate the score matrix of the double sequence alignment, and then use the dynamic programming algorithm to obtain the matched string. We selected two DNA sequences of non-equal length and used CLUSTAL software for sequence comparison. The local visualization of the comparison results is shown in **Figure 5**. Because the number of bases in the two DNA sequences is not equal, it is necessary to insert blanks to search for the maximum number of matched bases. The number of bases for a perfect match is 25, and the local similarity is also very high.

Later, BLAST and FASTA have important applications in the query and search of biological sequence databases. With the deepening of research, the swarm intelligence algorithm and its improved algorithm gradually began to be applied to biological sequences alignment, such as the genetic algorithm, ant colony algorithm, etc. Jangam and Chakraborti (2007) proposed a double sequence alignment hybrid algorithm based on a genetic algorithm and ant colony algorithm. The algorithm combines the local feature search capability of the ant colony algorithm and the global feature search capability of the genetic algorithm. The algorithm greatly optimizes the sequence alignment results. For short and medium sequences, the algorithm has high accuracy and better performance than the basic genetic algorithm, but the search efficiency for long sequences is low. To solve the problems of slow convergence and easy local optimization of the ant colony algorithm, Zhao et al. (2008) proposed a sequence comparison method based on an improved ant colony algorithm. By adjusting the initial and final positions of the ants and modifying the pheromone at different times, the algorithm solves the problem that the result falls into a locally optimal solution, but the amount of calculation is large, and it takes a long time to solve.

Multi sequence alignment (MSA) is an extension of double sequence alignment, but when the amount of sequences is large, it will face the problem of excessive data storage space occupation and high calculation complexity. MSA has a key characteristic: Since MSA is an NP-complete problem, MSA relies on approximate alignment heuristic algorithms. These heuristic algorithms depend to a certain extent on specific data attributes. This algorithm was proposed by Hogeweg, and later researchers developed sequence alignment packages based on it, such as CLUSTAL, T-Coffee, CLUSTALW. In recent years, the research and application of iterative algorithms in MSA have become common. Huo and Xiao (2007) proposed a graph-based DNA multi-sequence alignment algorithm: MWPAlign. This algorithm expresses sequence information as a structure graph and converts the sequence alignment problem into the maximum weight path of the graph. The algorithm has a linear time complexity, which significantly reduces the problem of excessive time complexity caused by MSA. However, when the mutation rate between sequences is different, the comparison result is poor, and the algorithm itself loses sequence similarity information in the process of looping. Lee et al. (2008) proposed a multi-sequence alignment genetic algorithm (GA-ACO) with ant colony optimization. GA-ACO algorithm combined with local search. GA-ACO uses ant colony optimization (ACO) to enhance the performance of GA. In the GA-ACO algorithm, GA guarantees the diversity of comparisons, and ACO avoids the result falling into a locally optimal solution. The hybrid genetic algorithm solves the problem of large-scale calculations, but the search speed of the algorithm is relatively slow, and more accurate solutions require more training time.

Naznin et al. (2011) proposed a method of multi-sequence alignment using genetic algorithm vertical decomposition (VDGA). The algorithm uses two mechanisms to generate the initial population: (1) generate a guide with randomly selected sequences Trees; (2) Combine sequences in such trees. VDGA divides the sequence vertically into two or more subsequences, then uses the guide tree method to solve them separately, and finally combines all the subsequences to generate a new multiple sequence alignment. After statistical and experimental analysis, VDGA is an effective method to solve the problem of multiple sequence alignment. The tree model is the most widely used in the field of machine learning, and it is also a model with many variants. The tree model is easy to understand and not easy to overfit, and it consumes fewer resources during training. So tree models are also often used in the biological sequence alignment.

Many studies have focused on heuristic techniques to solve MSA problems, among which stochastic methods are very effective methods. GA is a stochastic method, which can solve this type of optimization problem well. Chowdhury and Garai (2017) summarized the DNA multiple sequence alignment from the perspective of a genetic algorithm. Genetic algorithm has the following advantages in MSA:

1. You can find the optimal solution or the suboptimal solution of the sequence alignment problem in computing time;
2. Regardless of the length of the sequence and the number of sequences, this method is applicable;
3. There is much room for improvement in the optimization of the objective function, and the description of the objective function is crucial for the optimal solution of sequence alignment.

The scale of biological sequence data continues to grow, and sequence alignment is a necessary step for sequence data analysis. Since the research of sequence alignment is very mature, a large number of excellent and open-source sequence alignment tools have appeared. At present, the research of sequence alignment focuses on improving the speed of the alignment. Faced with such a large amount of sequence data, traditional sequence comparison tools can no longer handle it, so highly distributed computers will be required. In recent years, a distributed computing framework called Hadoop can be used for big data processing and storage. It has two main components, MapReduce for programming model and Hadoop Distributed File System (HDFS) for storing data. Using the distributed platform of the MapReduce model, massive sequencing data can be effectively stored and analyzed. Mondal and Khatua (2019) proposed a distributed sequence alignment algorithm: MRaligner. The algorithm is implemented in the Apache Spark framework using MapReduce. Compared with the traditional Smith-Waterman algorithm, the sequence comparison efficiency has been significantly improved. Besides, because the framework is flexible and extensible, increasing the number of processors and good distributed HDFS management will speed up processing.

Evaluation of biological sequence alignment algorithms mainly considers the efficiency of the algorithm and the sensitivity to obtain the best alignment results. The Smith-Waterman algorithm for double-sequence alignment is highly sensitive, but its complexity is high. FASTA and BLAST are a decrease in predicted sensitivity in exchange for an increase in speed. The CLUSTALW algorithm is the most common and effective among multiple sequence alignment algorithms. The main problem in sequence alignment is whether the sensitivity of the alignment and the efficiency of the algorithm have been improved for sequences with large differences.

Next-generation sequencing technology (NGS) has brought us a lot of biological data. Sequence alignment is always an indispensable step in finding the relationship between sequences. For fairly large input sequences, sequence alignment is a difficult task Currently, traditional sequence alignment tools are inefficient in terms of computing time. In the future, in the face of high throughput, biological sequence data, distributed sequence alignment, and parallel computing may be the focus of research in this field.

DNA Sequence Classification

Classification is an important mining task in machine learning. Its purpose is to learn a classification model from the training sample set to predict the category of unknown new samples. The classification of biological sequences as a special data type is a popular problem in data mining. It is a difficult problem, due to the non-numerical attributes of the biological sequence elements, the sequence relationship between the sequence elements, and the different sequence lengths of different events, etc. Sequence classification is to predict the type of DNA sequence based on the similarity of its structure or function, and then predict the sequence function and the relationship between other sequences, and assist in the identification of genes in DNA molecules.

Levy and Stormo (1997) proposed to use circular graphs (DAWGs) to classify DNA sequences. Müller and Koonin (2003) proposed to use vector space to classify DNA sequences. Ranawana and Palade (2005) proposed a multi-classifier system for identifying *E. coli* promoter sequences in DNA sequences. He Uses four different coding methods to encode the sequence and then uses the coding sequence to train four different neural networks. The classification results of the four individual neural networks were then combined through an aggregation function, which used a variation of the logarithmic opinion pool method. Experiments show that when the same data is provided to the neural network with different encoding methods, it can provide slightly different results that can be provided. At the same time, when the results of more classifiers with the same input data are integrated into a multi-classifier, the results we can obtain are better than the single performance of the neural network. However, the main disadvantage of the neural network design is that it is difficult to obtain the optimal parameters of the neural network. This will involve the deployment of the neural network and the optimization of the encoding method used.

Ma et al. (2001) proposed a DNA sequence classification based on the combination of the expectation-maximization algorithm and a neural network, and applied the algorithm to identify the DNA sequence classification of E. coli promoters. Ma Q uses an improved expectation-maximization algorithm to locate the −35 and −10 binding sites in the E. coli promoter sequence. It is no longer assumed that the lengths of the spacers between the binding sites and between the binding sites and the transcription start site are evenly distributed. Instead, he derives the probability distribution of these lengths. According to the information contained in each E. coli promoter sequence, he selects features and uses orthogonal coding methods to represent these features. Finally, these features are input into the neural network for promoter recognition. This method obtained good performance on different data sets.

Zaki et al. (2010) proposed a variable-order hidden Markov model with the continuous state: VOGUE. VOGUE uses a variable sequence mining method to extract frequent patterns with different lengths and spacings between elements, and then he constructs a variable sequence hidden Markov model. Compared with traditional HMM, VOGUE has higher classification accuracy. However, the frequency statistical characteristics of the sub-sequences in the sequence are not considered, which affects the generalization ability of the model.

In recent years, the convolutional neural network is a widely used deep learning model. Convolutional neural networks can extract abstract features from data. Nguyen et al. (2016) used DNA sequences as text data and proposed a new method for classifying DNA sequences with convolutional neural networks. This method uses a one-stop vector to represent the sequence as the input of the model. So, it retains the information of each nucleotide in the basic position sequence. The model was evaluated in 12 DNA sequence data sets. The results show that the model has improved significantly on all these data sets. The continuous development of deep learning has also opened up new ideas for DNA sequence mining.

The machine learning method used for supervised learning classification tasks depends on feature extraction. Bosco and Di Gangi (2016) proposed two different deep learning models. He used the model for classification tasks on five datasets. It turns out that neural deep learning framework or deep learning models can automatically extract useful features from input patterns.

A key problem in genomics is the classification and annotation of sequences. In recent years, a variety of machine learning techniques have been used to complete this task. In any case, the main difficulty behind the problem is still the feature selection process. The sequence has no clear features And the general representation method easily introduces high-dimensional problems. How to effectively represent sequence features and analyze high dimensional data is the difficulty of research.

DNA Sequence Clustering

Cluster analysis is one of the most commonly used methods of machine learning. It is different from the classification that we don't know specific categories in advance. Cluster analysis is unsupervised learning of data patterns. DNA sequence clustering is based on sequence similarity analysis. Cluster analysis clusters DNA sequences with similar characteristics into a cluster and then analyzes biological sequence functions. How to determine whether there is a similarity between sequences is the key to DNA sequence clustering. At present, a lot of research in DNA sequence clustering is based on the local characteristics of DNA for clustering, and the clustering results of DNA sequences are affected by many factors Impact. If a clustering algorithm that considers the global characteristics of DNA sequences can be designed, the accuracy of clustering will be greatly improved, and it is of great significance for the further analysis of DNA sequence clusters.

Early foreign scholars Krause et al. (2000) proposed the SYSTERS algorithm, Enright et al. (2002) proposed the GENERAGE algorithm, the basic idea of the two is to calculate the similarity between sequences, and then use a hierarchical clustering algorithm to complete sequence clustering. Gerhardt et al. (2006) proposed a DNA sequence clustering tool based on the concept of graph theory. This method studies the path topology of the biological genome through a triplet network. In this network, the triplets in the DNA sequence are vertices. If two vertices appear side by side on the genome, they are connected. Then the cluster topology is measured to characterize this network topology. Finally, he aims at two main deviations: guanine-cytosine (GC) content and periodicity of DNA sequence base pairs, he constructed some test data of DNA sequences and studied the clustering method based on the constructed random network. The conclusion proves that the clustering coefficient has its research value. Based on the new distance measure DMk, Wei D proposed a new unaligned DNA sequence clustering algorithm mBKM. This method converts the DNA sequence into a feature vector. This method transforms DNA sequences into the feature vectors which contain the occurrence, location, and order relation of k-tuples in the DNA sequence. The mBKM algorithm can effectively classify DNA sequences with similar biological characteristics and discover the relationships between DNA sequences (Wei et al., 2012). However, the method did not consider edge length, and it has not addressed problems with long repeated sequences or long insertions.

Some recent studies have proposed methods for converting DNA data into genomic digital signals. These studies will provide opportunities for existing digital signal processing methods to be used in genomic data. Mendizabal-Ruiz G proposed a method for clustering analysis of DNA sequences based on GSP and K-means clustering. He chose Euclidean distance as the similarity measure to be adopted by the K-means algorithm. This method can be used to evaluate the ability of markers or genes to distinguish organisms at different levels, identify subgroups in a group of organisms, and classify fragments of DNA sequences based on known sequences (Mendizabal-Ruiz et al., 2018). Mendizabal-Ruiz G has demonstrated that it is possible to group DNA sequences based on their frequency components. The future research direction is to determine whether different pyramids occupy the weight of size in sequence clustering.

At present, the two key points of DNA sequence clustering are how to extract the characteristic subsequences in the DNA sequence, and how to design an effective similarity measure from the biological meaning. Based on the above two key points, the design of the DNA sequence clustering algorithm will get a more practical application of clustering results.

DNA Sequence Pattern Mining

During DNA evolution, its sequence patterns are well conserved, which is of great significance for biological research. The DNA sequence pattern is usually a sequence fragment in the DNA sequence that has a specific function. In the process of DNA evolution, the more conserved regions in most sequences will form specific sequence patterns, and the structure and function of these sequences play an important role. Therefore, identifying these patterns is an important research content of DNA sequence data analysis. This helps to predict DNA sequence function and explain the evolutionary relationship between sequences. The purpose of DNA sequence pattern mining is to find such sequence patterns from DNA sequences and to identify genes and their functions.

Since, Srikant and Agrawal (1996) defined rearranged sequence pattern mining in 1995, related research has become an important field of machine learning. It has attracted the attention of researchers. There are many types of replacement patterns, including interchange item sets, repetitive subsequences, and replacement substructures. DNA sequence pattern mining is to search for replacement subsequences in a sequence.

As shown in **Figure 6**, it is a schematic diagram of the sequence mode. The eight green lines in **Figure 6** represent eight sequences, and the three different colored squares represent the three patterns of the sequence. Sequences 3, 5, 6, and 7 all contain the pattern one, sequences 1, 2, 4, 5, 6, and 8 all contain pattern 2, and sequence 3 contains the only pattern three. We can find that both sequences 4 and sequence 8 contain two patterns, which can be used to further analyze the common nature of the two sequences.

The *Apriori* algorithm is commonly used to mine data association rules. It is used to find data sets that frequently appear in data values. Finding out the patterns of these sets helps us make some decisions. Srikant proposed a GSP (generalized sequential patterns mining) algorithm based on the *Apriori* algorithm. The GSP algorithm introduces time and conceptual level constraints and uses a bottom-up breadth-first strategy to mine all frequent patterns (Srikant and Agrawal, 1996). However, when the scale of the sequence database is large, a large number of candidate patterns are generated, and the sequence database needs to be scanned frequently, which leads to the overall efficiency of the algorithm. Therefore, the *Apriori* algorithm is rarely studied alone, but the *Apriori* idea is often used in combination with other algorithms, which will also produce good research results.

At present, there are two main types of calculation methods found in the study of biological sequence patterns: (1) One type uses a heuristic search strategy. This type of algorithm is usually an iterative process. The optimal solution is obtained through repeated iterations. The advantage of the solution is that the calculation complexity is reduced. This kind of method is suitable for the study of subdivided DNA sequences. The disadvantage is that its solution may fall into the local optimum; (2) Another type of algorithm uses an exhaustive search strategy to enumerate all possible solutions and evaluate them one by one to find the best solution.

Existing sequential pattern mining algorithms can be roughly divided into two categories: (1) One type of sequential pattern mining algorithm is to search for patterns in the sequence that exceed a certain threshold: mining alternative patterns. They can only mine alternating patterns in a single sequence. (2) However, in the study of biological sequence analysis, we often require simultaneous analysis of sequence patterns in sequence sets, which cannot be achieved by this method. Therefore, another type of sequence pattern mining algorithm is needed. This type of sequence pattern mining algorithm mines repeated sequence patterns in data sets. When we face massive amounts of biological data, such algorithms usually search very slowly.

Zhou et al. (2010) proposed a pattern mining algorithm: mMBioPM. The algorithm solves the problem of redundancy in the mining results by optimizing the hash table structure with pattern division features, and reduces the calculation time and improves the mining efficiency. To overcome the time complexity and memory overhead caused by a large number of projection databases and short patterns generated by the frequent pattern mining algorithm, Chen and Liu (2011) proposed a fast and efficient biological sequence frequent pattern mining algorithm: FBPM. He defined the concept of the main mode and then used the prefix tree algorithm to mine frequent main modes. At the same time, he used a pattern growth method to mine all common frequent patterns in the sequence.

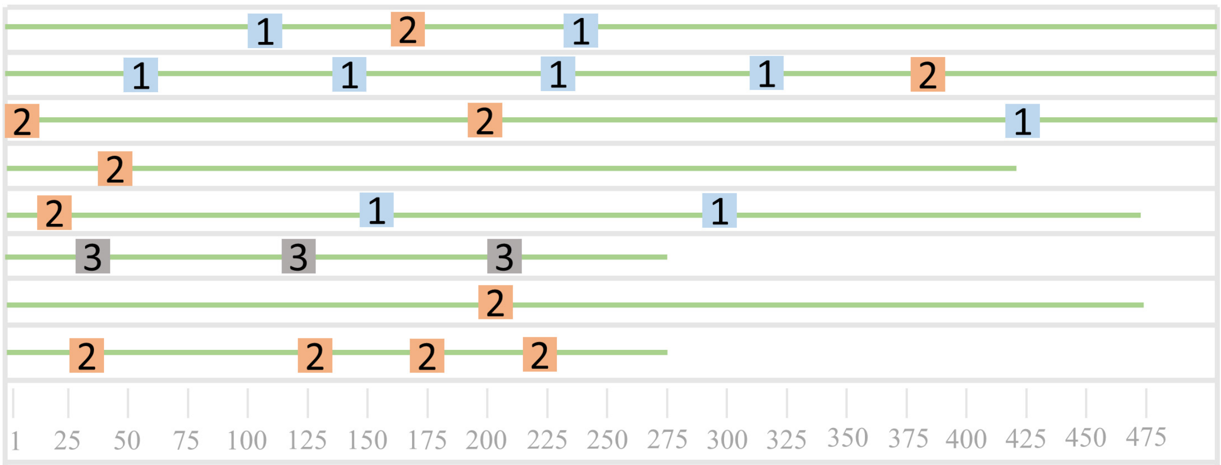


FIGURE 6 | Sequence pattern diagram.

In the biological sequence database composed of DNA sequences, the existing search algorithm is time-consuming and requires multiple scans of the database. To overcome these disadvantages, Junyan and Chenhui (2015) proposed an SPMM algorithm based on the Markov chain. The algorithm calculates the transition probability matrix of DNA sequences in the sequence database and gives the minimum support threshold as a constraint condition for mining sequence patterns. The calculated degree of support of the subsequence is compared with the threshold to finally determine the sequence pattern. The experiment proves that the SPMM algorithm not only obtains a higher mining speed, but also the mining quality of the sequence mode is higher.

Mao (2019) designed a compact data structure called an association matrix. Based on the association matrix structure, he designed an algorithm for effectively mining key fragments in DNA sequences. The correlation matrix is a novel in-memory data structure. Its structure is very compact and can handle ultra-long DNA sequences in limited storage space. By designing a compact memory data structure and a processing mechanism based on short sequences, it provides a novel idea for analyzing DNA sequences. The effective structure of the correlation matrix can help to efficiently mine key fragments from ultra-long DNA sequences.

DNA sequence pattern mining is a necessary means to study the structure and function of DNA sequences. Traditional DNA sequence pattern mining algorithms will result in a large number of redundant sequence patterns. These sequence patterns are usually short, and they have little biological significance, which makes the results of sequence pattern mining inefficient. At the same time, the long sequence always contains a considerable number of sub-sequences, so an explosive number of candidate sequence patterns will be generated, which will generate a lot of time and space consumption. How to design an appropriate search strategy and eliminate redundant sequence patterns will be an important direction for future research.

Open Issues

DNA sequence analysis provides an opportunity to explore the genetic variation of organisms. The rapid growth of DNA sequence data has continuously expanded the demand for DNA sequence analysis. At present, there are still the following problems in DNA sequence data mining:

1. There are still efficiency challenges when processing large-scale DNA sequence data;
2. For different biological needs, suitable DNA sequence data mining algorithms should be designed according to the corresponding background knowledge and sequence characteristics;
3. How to extract the sequence characteristics of DNA sequences and how to design an effective similarity measure to measure sequence similarity is very important;
4. Due to the "black box" nature of machine learning, the output of machine learning is difficult to give a reasonable explanation from a biological perspective, which limits the application of the model to a certain extent.

CONCLUSION

In the past few decades, the rapid development of hardware technology has opened up new possibilities for life scientists to collect data in various application fields, such as omics, biological imaging, medical imaging, etc. At the same time, the advancement of life science technology has brought Huge challenge. Today, how to apply numerous data mining technologies to bioinformatics analysis is a current research hotspot, including data mining architecture, machine learning algorithm development, and new data mining analysis function research suitable for biological information processing. At the same time, the interdisciplinary approach has promoted the development of machine learning. And artificial neural networks, deep learning, and reinforcement learning have made breakthroughs in machine intelligence. Besides, due to the growth of computing power, the acceleration of data storage speed and the reduction of computing costs,scientists in various fields have been able to apply these technologies to biological data. The close integration of machine learning and bioinformatics will result in more and more meaningful mining results, which will play a positive role in the progress of human society.

Based on the above research, we believe that the research of machine learning in DNA sequence analysis has two aspects that deserve attention:

On the one hand, it describes the biological significance of DNA sequences. At present, a large number of algorithms can achieve efficient performance when analyzing DNA sequences, but their mining results are highly sensitive and specific, which will make a large deviation during use. Therefore, how to integrate the biological significance of DNA sequences into the data mining process is a problem worthy of everyone's attention and research.

On the other hand, with the continuous expansion of data volume, traditional analysis tools are inefficient in terms of computing time, and how to design efficient calculation methods is an important research aspect. The integration of distributed computing and parallel computing will greatly improve mining efficiency.

At the same time, it is very necessary to choose a suitable DNA sequence coding method for a specific task. This can improve the performance of the algorithm and reduce the training time.

In summary, from the aspects of sequencing technology, DNA sequence data structure, and sequence similarity, this review comprehensively introduces the source and characteristics of DNA sequence data; we briefly summarize the machine learning algorithms and propose biological sequence data Challenges faced by machine learning algorithms in mining and possible solutions in the future. Then, we reviewed four typical applications of machine learning in DNA sequence data: DNA sequence alignment, classification, clustering, and pattern mining, analyzed and discussed their corresponding biological application background and significance, and systematically summarized recent years

Research on the field of DNA sequence data mining by domestic and foreign scholars. We put forward several key issues in the future research field of DNA sequence data mining and some future research directions and trends. In future research, I believe that the biological field and machine learning will be more closely integrated, and more effective mining results will be obtained.

AUTHOR CONTRIBUTIONS

AY proposed the idea of this manuscript. WZ searched the references. JW wrote this manuscript. KY and YH double checked this manuscript and formatted it as per the template of this journal. LZ polished the language. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Central Government Supports Local Reform and Development Fund Projects (North China University of Science and Technology 2022-5).

ACKNOWLEDGMENTS

We thank all the participants in this study.

REFERENCES

Bilofsky, H. S., Burks, C., Fickett, J. W., Goad, W. B., Lewitter, F. I., Rindone, W. P., et al. (1986). The GenBank genetic sequence databank. *Nucleic Acids Res.* 14, 1–4. doi: 10.1093/nar/14.1.1

Bosco, G. L., and Di Gangi, M. A. (2016). "Deep learning architectures for DNA sequence classification," in *Proceedings of the International Workshop on Fuzzy Logic and Applications* (Cham: Springer), 162–171. doi: 10.1007/978-3-319-52962-2_14

Chen, L., and Liu, W. (2011). "An algorithm for mining frequent patterns in biological sequence," in *Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)* (Piscataway, NJ: IEEE), 63–68. doi: 10.1109/ICCABS.2011.5729943

Choong, A. C. H., and Lee, N. K. (2017). "Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method," in *Proceedings of the 2017 International Conference on Computer and Drone Applications (IConDA)* (Piscataway, NJ: IEEE), 60–65. doi: 10.1109/ICONDA.2017.8270400

Chowdhury, B., and Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* 109, 419–431. doi: 10.1016/j.ygeno.2017.06.007

Chu, W. W. (2014). Data mining and knowledge discovery for Big Data. *Stud. Big Data* 1, 305–308. doi: 10.1007/978-3-642-40837-3

Delibas, E., and Arslan, A. (2020). DNA sequence similarity analysis using image texture analysis based on first-order statistics. *J. Mol. Graph. Model.* 99:107603. doi: 10.1016/j.jmgm.2020.107603

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Gerhardt, G. J. L., Lemke, N., and Corso, G. (2006). Network clustering coefficient approach to DNA sequence analysis. *Chaos Solitons Fractals* 28, 1037–1045. doi: 10.1016/j.chaos.2005.08.138

Henikoff, S., and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 89, 10915–10919. doi: 10.1073/pnas.89.22.10915

Huo, H. W., and Xiao, Z. W. (2007). A multiple alignment approach for DNA sequences based on the maximum weighted path algorithms. *Ruan Jian Xue Bao(Journal of Software)* 18, 185–195. doi: 10.1360/jos180185

Jangam, S. R., and Chakraborti, N. (2007). A novel method for alignment of two nucleic acid sequences using ant colony optimization and genetic algorithms. *Appl. Soft Comput.* 7, 1121–1130. doi: 10.1016/j.asoc.2006.11.004

Junyan, Z., and Chenhui, Y. (2015). "Sequence pattern mining based on markov chain," in *Proceedings of the 2015 7th International Conference on Information Technology in Medicine and Education (ITME)* (Piscataway, NJ: IEEE), 234–238. doi: 10.1109/ITME.2015.49

Krause, A., Stoye, J., and Vingron, M. (2000). The SYSTERS protein sequence cluster set. *Nucleic Acids Res.* 28, 270–272. doi: 10.1093/nar/28.1.270

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112. doi: 10.1093/bib/bbk007

Lee, Z. J., Su, S. F., and Chuang, C. C. (2008). Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Appl. Soft Comput.* 8, 55–78. doi: 10.1016/j.asoc.2006.10.012

Levy, S., and Stormo, G. D. (1997). "DNA sequence classification using DAWGs," in *Structures in Logic and Computer Science*, eds J. Mycielski, G. Rozenberg, and A. Salomaa (Berlin: Springer), 339–352. doi: 10.1007/3-540-63246-8_21

Li, J., Wong, L., and Yang, Q. (2005). Guest editors' introduction: data mining in bioinformatics. *IEEE Intell. Syst.* 20, 16–18. doi: 10.1109/MIS.2005.108

Ma, Q., Wang, J. T. L., Shasha, D., and Wu, C. H. (2001). DNA sequence classification via an expectation maximization algorithm and neural networks: a case study. *IEEE Trans. Syst.* 31, 468–475. doi: 10.1109/5326.983930

Mao, G. (2019). "Association matrix method and its applications in mining DNA sequences," in *Proceedings of the International Conference on Applied Human Factors and Ergonomics* (Piscataway, NJ: IEEE), 154–159. doi: 10.1007/978-3-030-20454-9_15

Mendizabal-Ruiz, G., Román-Godínez, I., and Torres-Ramos, S. (2018). Genomic signal processing for DNA sequence clustering. *PeerJ* 6:4264. doi: 10.7717/peerj.4264

Mondal, S., and Khatua, S. (2019). "Accelerating pairwise sequence alignment algorithm by mapreduce technique for next-generation sequencing (ngs) data analysis," in *Emerging Technologies in Data Mining and Information Security*, eds A. Abraham, P. Dutta, J. Mandal, A. Bhattacharya, and S. Dutta (Cham: Springer), 213–220. doi: 10.1007/978-981-13-1498-8_19

Müller, H. M., and Koonin, S. E. (2003). Vector space classification of DNA sequences. *J. Theor. Biol.* 223, 161–169. doi: 10.1016/S0022-5193(03)00082-1

Naznin, F., Sarker, R., and Essam, D. (2011). Vertical decomposition with genetic algorithm for multiple sequence alignment. *BMC Bioinformatics* 12:353. doi: 10.1186/1471-2105-12-353

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., et al. (2016). DNA sequence classification by convolutional neural network. *J. Biomed. Sci. Eng.* 9:280. doi: 10.4236/jbise.2016.95021

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinform.* 42, 1–8. doi: 10.1002/0471250953.bi0301s42

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85, 2444–2448. doi: 10.1073/pnas.85.8.2444

Ranawana, R., and Palade, V. (2005). A neural network based multi-classifier system for gene identification in DNA sequences. *Neural Comput. Appl.* 14, 122–131. doi: 10.1007/s00521-004-0447-7

Rogozin, I. B., Milanesi, L., and Kolchanov, N. A. (1996). Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* 12, 161–170. doi: 10.1093/bioinformatics/12.3.161

Roukos, D. H. (2010). Next-generation sequencing and epigenome technologies: potential medical applications. *Expert Rev. Med. Devices* 7, 723–726. doi: 10.1586/erd.10.68

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197. doi: 10.1016/0022-2836(81)90087-5

Srikant, R., and Agrawal, R. (1996). "Mining sequential patterns: generalization and performance improvements. Advances in Database Technology," in *Proceedings of the 15th Int'l Conf. on Extending Database Technology* (London: Springer-Verlag), 3–17. doi: 10.1007/BFb0014140

Watson, M. (2014). Illuminating the future of DNA sequencing. *Genome Biol.* 14:108. doi: 10.1186/gb4165

Wei, D., Jiang, Q., Wei, Y., and Wang, S. (2012). A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics* 13:174. doi: 10.1186/1471-2105-13-174

Zaki, M. J., Carothers, C. D., and Szymanski, B. K. (2010). VOGUE: a variable order hidden Markov model with duration based on frequent sequence mining. *ACM Trans. Knowl. Discov. Data* 4, 1–31. doi: 10.1145/1644873.1644878

Zhang, W., Ma, D., and Yao, W. (2014). Medical diagnosis data mining based on improved Apriori algorithm. *J. Netw.* 9:1339. doi: 10.4304/jnw.9.5.1339-1345

Zhao, Y., Ma, P., Lan, J., Liang, C., and Ji, G. (2008). "An improved ant colony algorithm for DNA sequence alignment," in *Proceedings of the 2008 International Symposium on Information Science and Engineering* (Piscataway, NJ: IEEE), 683–688. doi: 10.1109/ISISE.2008.82

Zhou, Q., Jiang, Q., and Li, S. (2010). "An efficient algorithm for protein sequence pattern mining," in *Proceedings of the 2010 5th International Conference on Computer Science & Education* (Piscataway, NJ: IEEE), 1876–1881. doi: 10.1109/ICCSE.2010.5593815

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 10 September 2020
doi: 10.3389/fbioe.2020.00839

Develop and Evaluate a New and Effective Approach for Predicting Dyslipidemia in Steel Workers

*Jianhui Wu[1,2], Sheng Qin[1], Jie Wang[1], Jing Li[1], Han Wang[1], Huiyuan Li[1], Zhe Chen[1], Chao Li[1], Jiaojiao Wang[1] and Juxiang Yuan[1,2]**

[1] School of Public Health, North China University of Science and Technology, Tangshan, China, [2] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, North China University of Science and Technology, Tangshan, China

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Xiaoxian Yang,
Shanghai Second Polytechnic University, China
Gunasekaran Manogaran,
University of California, Davis, United States
Jingxia Chen,
Shaanxi University of Science and Technology, Xi'an, China
Jing Zhang,
Jining Medical University, China

***Correspondence:**
Juxiang Yuan
gwxyjxb@ncst.edu.cn

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

Received: *05 April 2020*
Accepted: *30 June 2020*
Published: *10 September 2020*

Citation:
Wu J, Qin S, Wang J, Li J, Wang H, Li H, Chen Z, Li C, Wang J and Yuan J (2020) Develop and Evaluate a New and Effective Approach for Predicting Dyslipidemia in Steel Workers. Front. Bioeng. Biotechnol. 8:839. doi: 10.3389/fbioe.2020.00839

The convolutional neural network (CNN) has made certain progress in image processing, language processing, medical information processing and other aspects, and there are few relevant researches on its application in disease risk prediction. Dyslipidemia is a major and modifiable risk factor for cardiovascular disease, early detection of dyslipidemia and early intervention can effectively reduce the occurrence of cardiovascular diseases. Risk prediction model can effectively identify high-risk groups and is widely used in public health and clinical medicine. Steel workers are a special occupational group. Their particular occupational hazards, such as high temperatures, noise and shift work, make them more susceptible to disease than the general population, which makes the risk prediction model for the general population no longer applicable to steel workers. Therefore, it is necessary to establish a new model dedicated to the prediction of dyslipidemia of steel workers. In this study, the physical examination information of thousands of steel workers was collected, and the risk factors of dyslipidemia in steel workers were screened out. Then, based on the data characteristics, the corresponding parameters were set for the convolutional neural network model, and the risk of dyslipidemia in steel workers was predicted by using convolutional neural network. Finally, the predictive performance of the convolutional neural network model is compared with the existing predictive models of dyslipidemia, logistics regression model and BP neural network model. The results show that the convolutional neural network has a good predictive performance in the risk prediction of dyslipidemia of steel workers, and is superior to the Logistic regression model and BP neural network model.

Keywords: deep learning, convolutional neural network, dyslipidemia, steel worker, disease model prediction, model performance comparison

INTRODUCTION

Dyslipidemia is a chronic noncommunicable disease of lipid metabolism disorder, characterized by increased and/or decreased lipid levels in the blood. With the rapid development of China's economy and the change of life style, cardiovascular disease has become the main death disease of residents (Roth et al., 2017). In recent years, the blood lipid level of Chinese population has gradually increased, and the prevalence of dyslipidemia has increased significantly. Evidence demonstrates that dyslipidemia is an independent and modifiable major risk factor for cardiovascular disease, and its level can significantly increase the incidence and mortality of cardiovascular disease (Pikula et al., 2015; Lee et al., 2017). Studies have shown (Miller, 2009; Hendrani et al., 2016; Stevens et al., 2016) that the early detection and management of high-risk groups with dyslipidemia can effectively reduce the incidence of cardiovascular disease, which can reduce the burden of cardiovascular disease and brings great social value.

China has a huge number of steel workers. Steel workers are a special occupational group, whose occupational environment is special, such as high temperature, noise, shift system and other special occupational exposure can cause or affect the occurrence of chronic diseases (Chauhan et al., 2014; Hedén Stahl et al., 2014; Tong et al., 2017; Wu et al., 2019b). Therefore, the prediction model of dyslipidemia in the general population is not suitable for steel workers. In order to improve the quality of life and health status of steel workers, it is urgent to establish a new risk prediction model of dyslipidemia in steel workers.

Logistics regression is a traditional prediction model, which is widely used in the field of disease prediction due to its clear parameter significance and easy to understand outcome indicators (Liu et al., 2018). However, its applicable conditions are relatively strict, which often limits the accuracy of its predictions. BP neural network is a widely used artificial neural network for disease prediction (Yao et al., 2019). Its good nonlinear processing ability and flexible grid structure make it have a good self-learning ability. However, it has a slow learning speed and is liable to fall into local minima, which makes its network promotion ability limited. Convolutional neural network is a kind of feedforward neural network with deep structure and convolution computation. The convolution structure can reduce the memory occupied by the neural network and has strong adaptability. It is good at mining local features of data and extracting global training features and classification, which has some advantages that traditional technologies do not have. In addition, the three key operations of convolution kernel, "local receptive field," weight sharing and pooling, can effectively reduce the number of network parameters, significantly reduce the computational complexity, and alleviate the problem of model overfitting.

Based on thousands of physical examination data of steel workers, we established a convolutional neural network model to predict the risk of dyslipidemia of steel workers, and compared the prediction performance with the existing dyslipidemia prediction model. Overall, our study consists of three contributions:

1. Based on thousands of physical examination data of steel workers, we screened out the risk factors of dyslipidemia of steel workers, which can provide a basis for formulating early prevention strategies for dyslipidemia of steel workers.
2. Combine the characteristics of the data to set the corresponding parameters of the model, and use the convolutional neural network to predict the risk of dyslipidemia in steel workers. We found that the convolutional neural network has a good fit with the physical examination data of steel workers, and has a good prediction performance.
3. Compare the prediction performance of the convolutional neural network model with some of the existing dyslipidemia prediction models and find that the prediction performance of the convolutional neural network model is better. In this way, we can use convolutional neural networks to predict the risk of dyslipidemia of steel workers, so as to achieve the early prevention of dyslipidemia of steel workers and improve the health and quality of life of steel workers.

RELATED WORK

Disease risk prediction model is a very effective way for early detection of high-risk groups. In recent years, more and more studies on model prediction of dyslipidemia have been conducted, such as Xinghua Yang et al. (2018) established a logistics model of dyslipidemia using a longitudinal database based on Taiwanese MJ health checkups. Chongjian Wang et al. (2012) established an artificial neural network model to identify those at high risk of dyslipidemia in rural adult residents. Xiaoshuai Zhang et al. (2019) used a random forest survival model to predict the risk of dyslipidemia in Chinese Han adults. However, these studies are aimed at the general population, and there are few studies on the risk prediction of dyslipidemia in special occupational populations.

Convolutional neural network has been widely used in medical research and has shown good accuracy and generalization ability (Lee et al., 2018; Lin et al., 2018; Horiuchi et al., 2019; Wu et al., 2019a). However, no one has tried to establish and evaluate the effect of convolutional neural network model on predicting the risk of dyslipidemia in steel workers.

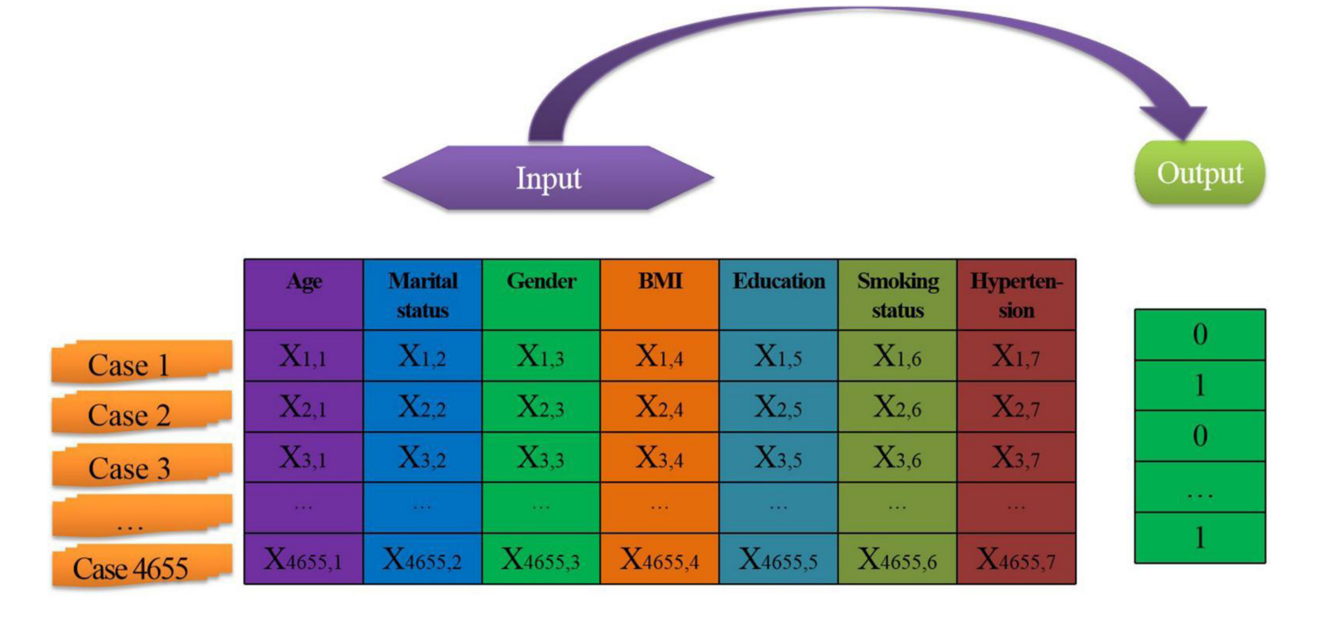
MATERIALS AND METHODS

Study Population

This study was a cross-sectional survey. Based on the baseline data of the health effects cohort study of the occupational population in the Beijing-Tianjin-Hebei region, steel workers who had undergone occupational health examinations in a steel group company hospital from March 2017 to June 2017 were selected as the research objects. To be eligible, steel workers must on-the-job for at least 1 year, aged ≤ 60 years and free from incomplete health examination data. Ultimately, a total of 4655 steel workers were included in the study. All steel workers included in the study received written informed consent. According to the 2016 Chinese guidelines for the management of dyslipidemia in adults (Joint committee for guideline revision, 2018), the steel workers were divided into the dyslipidemia group and the non-dyslipidemia group. Dyslipidemia refers to the total cholesterol (TC) $\geq$ 6.2 mmol/L, and/or triglyceride (TG) $\geq$ 2.3 mmol/L, and/or low-density lipoprotein cholesterol (LDL-c) $\geq$ 4.1 mmol/L, and/or high-density lipoprotein cholesterol (HDL-c) $<$ 1.0 mmol/L.

Data Collection

The basic personal information of the steel workers was collected in a one-to-one questionnaire by trained investigators, which mainly included ethnicity, age, gender, education leve, marital status, income, family history of hyperlipidemia, drinking status, smoking status, physical activity, diet, etc. Anthropometric data are measured and collected by doctors and professional trainers in the physical examination center according to the unified standards, which mainly include weight, height, hip circumference, waist circumference, blood pressure, etc. Then the body mass index (BMI) is calculated by dividing the weight



	Age	Marital status	Gender	BMI	Education	Smoking status	Hyperten-sion	
Case 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	$X_{1,7}$	0
Case 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$	$X_{2,6}$	$X_{2,7}$	1
Case 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$	$X_{3,6}$	$X_{3,7}$	0
…	…	…	…	…	…	…	…	…
Case 4655	$X_{4655,1}$	$X_{4655,2}$	$X_{4655,3}$	$X_{4655,4}$	$X_{4655,5}$	$X_{4655,6}$	$X_{4655,7}$	1

FIGURE 1 | Sample data structure.

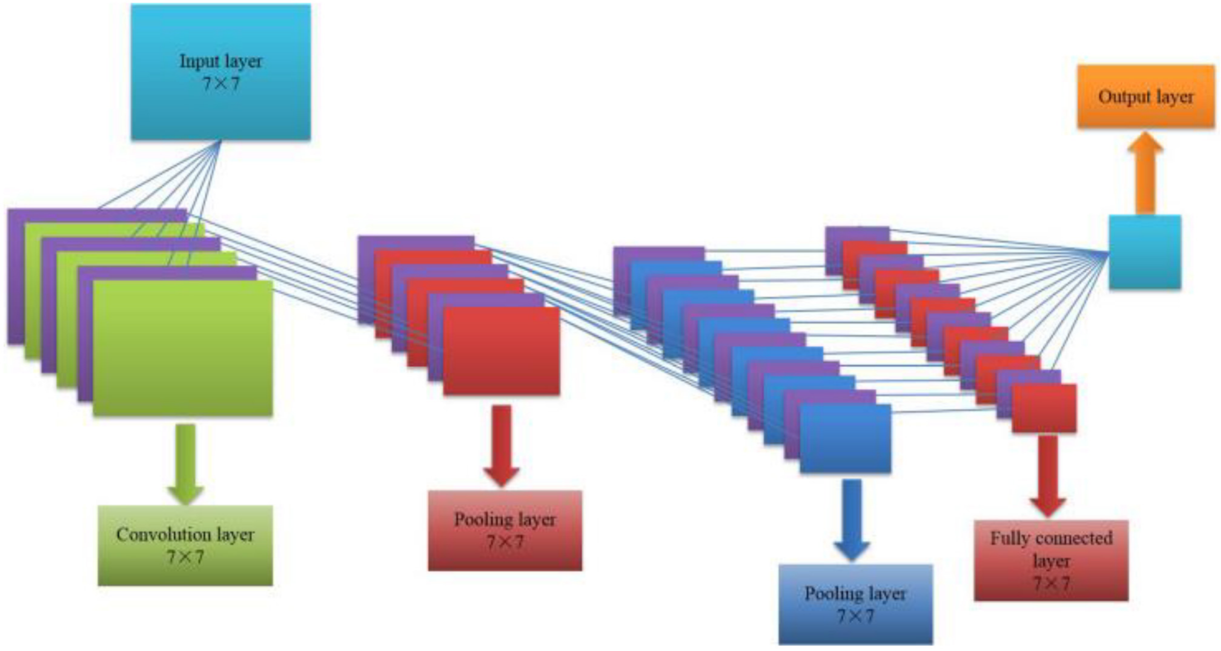


FIGURE 2 | CNN algorithm structure.

(kg) by the square of height (m), and the waist to hip ratio (WHR) is calculated by dividing the waist circumference (cm) by the hip circumference (cm). Laboratory test data were obtained by analyzing fasting blood samples of steel workers collected by doctors or nurses in the hospital, which mainly included total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), fasting blood glucose, etc. Occupational factors are provided by steel companies, which mainly include high temperature exposure, noise exposure, shift exposure, etc. Hypertension is defined as blood pressure ≥ 140/90 mmHg or diagnosed as hypertension by doctors. Diabetes is defined as FPG ≥ 7.0 mmol / L or diagnosed as diabetes by doctors.

Model Independent Variable Filtering Method

We established an Excel database based on questionnaires and physical examination data, and screened out independent variables of risk factors for dyslipidemia of steel workers for model prediction. Measurement data was presented as $\bar{X} \pm S$ for normal distribution or M(P25, P75) for non-normal distribution, and we used t-test or the rank sum test for comparison between groups, respectively. The classification data was expressed by numbers and percentages, and the comparison between groups was performed by Chi-square test. The rank data was presented by numbers and composition ratio, and the rank sum test was used for inter group comparison. Unconditional Logistic regression analysis was used for multivariate analysis of influencing factors. Differences were deemed significant when $p < 0.05$. Factors influencing dyslipidemia of steel workers were screened out by univariate analysis and multi-factor logistics regression analysis. In order to avoid the influence of data multicollinearity, the screened influencing factors were diagnosed by multicollinearity. Combined with expert consultation and literature inquiry to determine the appropriate model independent variables. The statistical analysis was performed by SPSS 25.0. ROC curves were drawn using MedCalc.

The Construction of Sample Set

After screening (the screening results will be introduced later), a total of 4655 steel workers' physical examination data constitute the sample set, as shown in **Figure 1**. There are seven independent variables, and the output target value is the presence or absence of dyslipidemia (Dyslipidemia is represented by 1 and non-dyslipidemia is represented by 0). Meanwhile, 4655 sample data were randomly assigned into 70% training set (n = 3258), 20% verification set (n = 931), and 10% test set (n = 466).

Convolutional Neural Network Configuration

Convolutional neural network is an important algorithm in the field of deep learning, including five parts of input layer, convolution layer, activation function, pooling layer and fully connected layer. It continuously adjusts the bias and connection weights between various neurons by combining forward propagation of information and backward propagation of error (Arun et al., 2018; Keshari et al., 2018). Its algorithm structure is shown in **Figure 2**.

To predict whether steel workers are dyslipidemia by convolutional neural network, setting reasonable complex structure is an important premise to ensure the accuracy of the prediction model. According to the characteristics of the collected data, the network structure of convolutional neural network model designed is shown in **Figure 3**. We set up 1 input layer, 3 convolution layers, 3 pooling layers, 1 fully connected layer and 1 output layer in convolutional neural network. The size of the convolution kernel set by the convolution layer 1∼3 is 2 × 2, and the number of convolution nuclei is 20. All the three poolings are maximized sampling (Xu et al., 2015; Iqbal et al., 2018), the core size is 2 × 2. The activation functions are all Relu functions. The number of neurons in the whole connective layer is 25.

Convolutional Neural Network Algorithm Solution

In this paper, we use the data of thousands of physical examination questionnaires of steel workers and convolution neural network algorithm to analyze and predict whether individuals have dyslipidemia.

A convolutional neural network for data processing rules that input data will pass through one or more hidden layers. In the hidden layer, each data is assigned a weight and bias, so the input data is assigned a new output value. If these new output

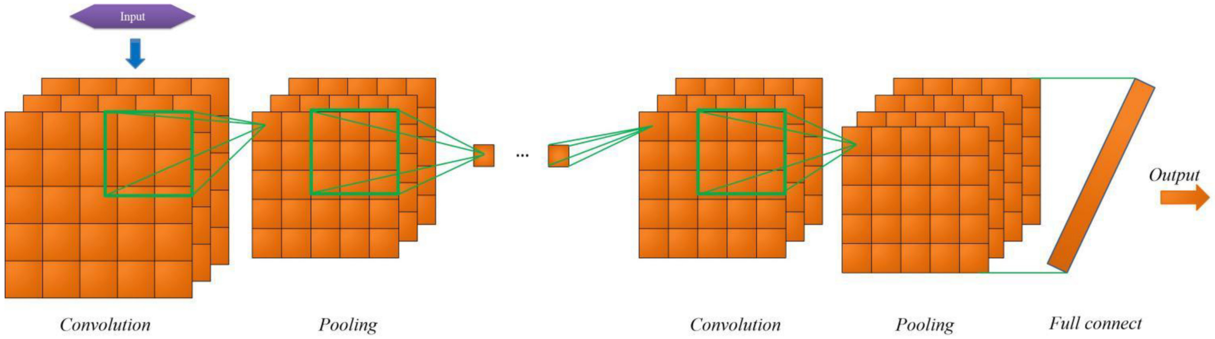


FIGURE 3 | Structure of CNN prediction model for dyslipidemia in steel workers.

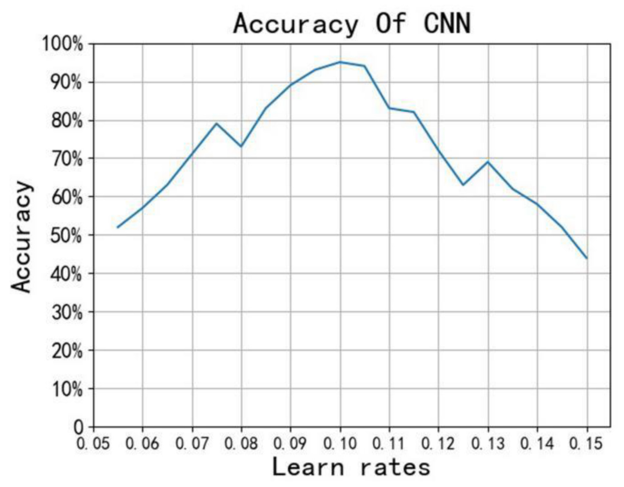


FIGURE 4 | Learning rate and accuracy.

values do not meet expectations, they are also assigned new weights and bias, and the process is repeated to produce the final output. The process mainly includes forward propagation and backward propagation.

The forward propagation calculation formula of convolutional neural network is:

$$a_j^{(l)} = f\left(u^l\right) \tag{1}$$

$$u^l = W^l a^{(l-1)} + b^l \tag{2}$$

Where $a_j^{(l)}$ represents the output of layer l, W^l represents weights, b^l represents biases, f is the activation function.

Since the feature map input in the forward propagation of the convolutional layer is convolved with the convolution kernel, the forward propagation formula of the j-th convolution kernel in the l-th layer of the convolutional neural network is as follows:

$$a_j^{(l)} = f\left(\sum_{i\in N_j} a_j^{(l-1)} \cdot k_{ij}^{(l)} + b_j^{(l)}\right) \tag{3}$$

Where k is the convolution kernel, $a_j^{(l-1)}$ is the output of the j-th convolution kernel of the l-1 layer, N_j is a choice of input features. $b_j^{(l)}$ is the bias Shared by each convolution layer, f is the activation function of the convolution layer.

The forward propagation of the pooling layer requires the pooling calculation of the input features and then other calculations. The calculation formula of the pooling layer is as follows:

$$a_j^{(l)} = f\left(\beta_j^{(l)} pooling\left(a_j^{(l-1)}\right) + b_j^{(l)}\right) \tag{4}$$

Where $a_j^{(l)}$ is the result of pooling the j-th characteristic map of the l-th convolution. *Pooling* is the pooling operation, $\beta_j^{(l)}$ is

TABLE 1 | Comparison of baseline characteristics of dyslipidemia and non-dyslipidemia patients in steel workers.

Variable	**Dyslipidemia *N* (%)/*M* (P25,P75)**		**χ^2/Z**	***p***
	No (*N* = 2860)	**Yes (*N* = 1795)**		
Age	46 (38,50)	46 (39,50)	0.739	0.46
Gender			54.494	<0.001
Male	2549 (89.1)	1711 (95.3)		
Female	311 (10.9)	84 (4.7)		
Nation			1.834	0.176
Han	2801 (97.9)	1747 (97.3)		
Other	59 (2.1)	48 (2.7)		
Marital status			2.0046	<0.001
Unmarried	119 (4.2)	34 (1.9)		
Married	2670 (93.4)	1702 (94.8)		
Other	71 (2.5)	59 (3.3)		
Education			18.67	<0.001
Elementary and below	37 (1.3)	18 (1.0)		
Middle and high school	2142 (74.9)	1408 (79.4)		
Junior college and undergraduate	639 (22.3)	363 (20.2)		
Graduate and above	42 (1.5)	6 (0.3)		
Monthly income	6000 (4000,7000)	5000 (4000,7000)	0.259	0.796
Family history of hyperlipidemia			0.976	0.323
No	2724 (95.2)	1698 (94.6)		
Yes	136 (4.8)	97 (5.4)		
Smoking status			93.918	<0.001
No smoking	1374 (48.0)	605 (33.7)		
Quit smoking	149 (5.2)	105 (5.8)		
smoking	1337 (46.7)	1085 (60.4)		
Drinking situation			11.509	0.003
No drinking	1762 (61.6)	1016 (56.6)		
Quit drinking	58 (2.0)	40 (2.2)		
Drinking	1040 (36.4)	739 (41.2)		
Physical activity			0.247	0.884
Mild	616 (21.5)	378 (21.1)		
Moderate	1233 (43.1)	786 (43.8)		
Severe	1011 (35.3)	631 (35.2)		
High fat diet score	12 (11,13)	12 (11,13)	0.916	0.36
Vegetable Fruit Score	6 (6,7)	6 (5,7)	1.748	0.08
BMI	24.9 (22.7,27.2)	26.7 (24.5,29.1)	17.277	<0.001
WHR	0.875 (0.831,0.917)	0.901 (0.862,0.939)	13.4	<0.001
Diabetes			10.448	0.001
No	2751 (96.2)	1690 (94.2)		
Yes	109 (3.8)	105 (5.8)		
Hypertension			34.381	<0.001
No	2480 (86.7)	1441 (80.3)		
Yes	380 (13.3)	354 (197)		
Shift work			6.276	0.043
Never shift	500 (17.5)	269 (15.0)		
Once shifts	509 (17.8)	306 (17.0)		
Now shifts	1851 (64.7)	1220 (68.0)		
Occupation noise			3.02	0.082
No	2279 (79.7)	1392 (77.5)		
Yes	581 (20.3)	403 (22.5)		
Occupation high temperature			7.288	0.007
No	2373 (83.0)	1433 (79.8)		
Yes	487 (17.0)	362 (20.2)		

TABLE 2 | Multicollinearity diagnostic table.

	Collinearity statistics	
	Tolerance	**VIF**
(Constant)		
Age	0.711	1.407
Marital status	0.91	1.099
Gender	0.873	1.146
BMI	0.936	1.069
WHR	0.979	1.021
Education	0.758	1.319
Smoking status	0.849	1.178
Drinking situation	0.88	1.136
Diabetes	0.966	1.035
Hypertension	0.91	1.098
Shift work	0.966	1.035
Occupation high temperature	0.991	1.009

the multiplicative bias of the pooling layer, and $b_j^{(l)}$ is the additive bias of the pooling layer.

The back propagation of convolutional neural network. Suppose that the loss function J_{mse} defined as the convolution neural network is the mean square error, and the formula is as follows:

$$J_{mse} = \frac{1}{2} \sum_{i=1} \left(Y_i - y_i \right)^2 \tag{5}$$

Where Y_i is the actual value and y_i is the output value.

Backpropagation of the fully connected layer in convolutional neural network is obtained by BP algorithm. For the convolution layer of the convolution neural network, if the next layer of the convolution layer l is the fully connected layer, then the sensitivity $\delta_j^{(l)}$ of the j-th convolution kernel can be obtained by the BP algorithm. If it is the pooling layer, then the calculation formula of the error sensitivity is:

$$\delta_j^{(l)} = \beta_j^{(l+1)} \left(f' \left(u_j^{(l)} \right) \circ up \left(\delta_j^{(l+1)} \right) \right) \tag{6}$$

Where $\beta_j^{(l+1)}$ represents the multiplier bias of the corresponding pooling layer, and up represents the anti-pooling operation. After the error sensitivity of the convolution layer is obtained, the convolution kernel and bias of the convolution layer are updated, and the formula is as follows:

$$\frac{\partial J_{mse}}{\partial k_{ij}^{(l)}} = \sum_{m,n} \delta_j^{(l)} p_j^{(l-1)} \tag{7}$$

$$\frac{\partial J_{mse}}{\partial b_j} = \sum_{m,n} \left(\delta_j^{(l)} \right)_{m,n} \tag{8}$$

Where $p_j^{(l-1)}$ is the value of $a_j^{(l-1)}$ multiplied by each element of the convolution kernel $k_{ij}^{(l)}$, m and n are the location information of the element in the input feature.

Similarly, the pooling layer is similar to the convolution layer. When the pooling layer is followed by the fully connected layer, the error sensitivity can be obtained by BP algorithm. When the pooling layer is the convolution layer, the error sensitivity is:

$$\delta_j^{(l)} = f' \left(u_j^{(l)} \right) \circ conv2 \left(\delta_j^{(l+1)}, rot180 \left(k_j^{l+1} \right), 'full' \right) \tag{9}$$

Where $conv2$ represents the convolution calculation, $rot180$ represents the rotation of the matrix by 180 degrees, and $full$ represents the missing data in the matrix replaced by 0. After the error sensitivity of the pooling layer is obtained, the gradient calculation formula of $b_j^{(l)}$ and $\beta_j^{(l)}$ is as follows:

$$\frac{\partial J_{mse}}{\partial b_j} = \sum_{m,n} \left(\delta_j^{(l)} \right)_{m,n} \tag{10}$$

$$\frac{\partial J_{mse}}{\partial \beta_j} = \sum_{m,n} \left(\delta_j^{(l)} \circ pooling \left(a_j^{(l-1)} \right) \right)_{m,n} \tag{11}$$

Platform and Parameter Settings

In this paper, TensorFlow modules in Python are used to construct the convolutional neural network model. TensorFlow is fully open source and available to anyone with minimal device configuration requirements. It can run models automatically on all platforms, from mobile phones, a single CPU/GPU, to distributed systems consisting of hundreds of GPU CARDS.

TABLE 3 | Multivariate logistics regression analysis of risk factors of dyslipidemia in steel workers.

Variable	B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I. for Exp (B)	
							Lower	**Upper**
Marital status (others)								
Unmarried	−0.963	0.287	11.285	1	0.001	0.382	0.218	0.67
Gender (female)	−0.479	0.137	12.193	1	0	0.619	0.473	0.81
BMI	0.13	0.009	198.068	1	0	1.139	1.118	1.159
Education (graduate and above)								
Middle and high school	1.072	0.452	5.612	1	0.018	2.921	1.203	7.091
Junior college and undergraduate	1.035	0.452	5.253	1	0.022	2.815	1.162	6.821
Smoking status (smoking)			44.924	2	0			
No smoking	−0.473	0.071	44.821	1	0	0.623	0.542	0.716
Hypertension	0.187	0.088	4.521	1	0.033	1.206	1.015	1.434

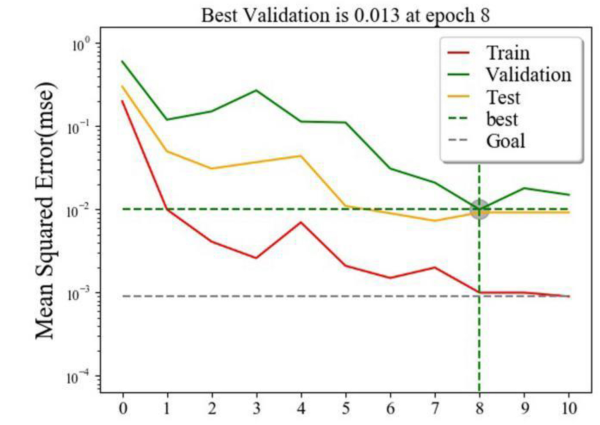


FIGURE 5 | Effect error graph of CNN learning.

We use the random initialization function to set the weight and bias. The smaller the learning rate, the longer the model takes to converge, but it can improve the accuracy of the model. In order to find the best learning rate of convolutional neural network, we first use random function in Python to initialize the learning rate at random, and then use Python to traverse different learning rates in steps of 0.01. Finally, use Matplotlib module to make some corresponding images as shown in **Figure 4**. It can be seen that when the learning rate is about 0.1, the accuracy is the highest. Therefore, choosing 0.1 as the learning rate can make the convolutional neural network achieve better prediction effect.

Performance Metrics

In this paper, five performance metrics including accuracy, sensitivity, specificity, F1-score and ROC curve were selected to evaluate the performance of the convolutional neural network model. Meanwhile, the prediction performance of training set and test set of convolutional neural network model, Logistics regression model and BP neural network model was compared. The calculation method of the above metrics are as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{13}$$

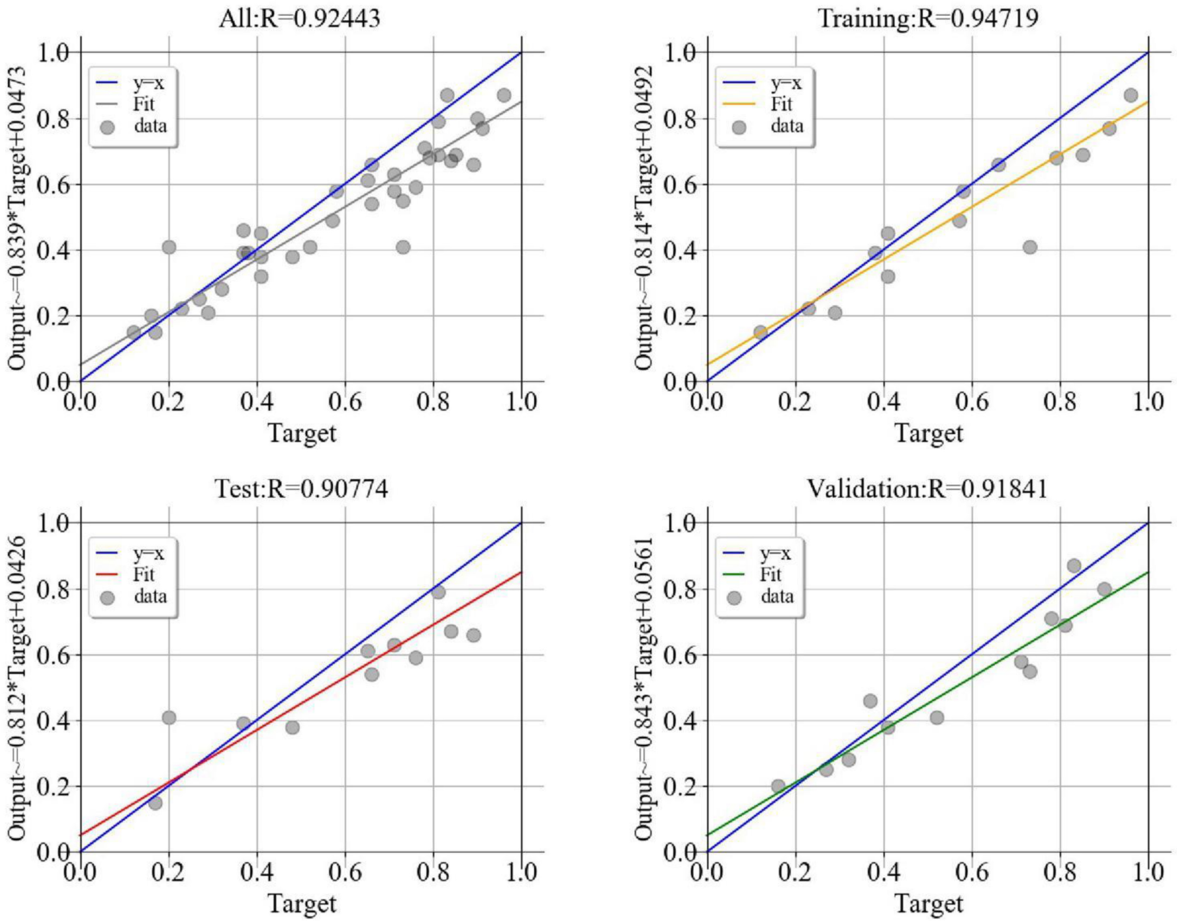


FIGURE 6 | CNN model goodness of fit test chart.

TABLE 4 | Comparison of performance metrics of each model.

Model	Training set				Test set			
	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 score	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 score
Logistics	72.45	76.47	74.92	0.69	71.11	70.30	70.6	0.65
BP neural network	86.7	88.96	88.09	0.85	81.11	83.57	82.62	0.78
CNN	93.23	95.65	94.72	0.93	90.00	91.26	90.77	0.88

TABLE 5 | Performance metrics of convolutional neural network.

Performance metrics	Training set	Test set	Validation set
Sensitivity (%)	93.23	90.00	89.97
Specificity (%)	95.65	91.26	93.01
Accuracy (%)	94.72	90.77	91.84
F1 score	0.93	0.91	0.89
AUC (95% CI)	0.944 (0.936–0.952)	0.906 (0.876–0.931)	0.915 (0.895–0.932)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{14}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{15}$$

$$precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

TP represents true positions, TN represents true negatives, FP represents false positions, FN represents false negatives. Sensitivity reflects the model's ability to find patients, specificity reflects the model's ability to find non-patients, and accuracy represents the model's overall predictive ability. The F1 score is a harmonic average of the accuracy and recall rates and is used as a final measurement. In addition, ROC curves and AUC are often used to test the balance between true and false positive rates.

RESULTS

Baseline Characteristics

A total of 4655 subjects were included in this study, including 1795 cases of dyslipidemia (38.56%) and 2860 cases of non-dyslipidemia (61.43%). The characteristics of baseline data and the results of univariate analysis are shown in **Table 1**. Univariate analysis showed that there were statistically significant differences ($p < 0.05$) between the dyslipidemia group and the non-dyslipidemia group in gender, educational level, marital status, smoking status, drinking situation, hypertension, diabetes, BMI, waist-to-hip ratio, shift work, and occupational high temperature. Unexpectedly, no significant differences ($p > 0.05$) were observed in ethnicity, age, income, diet, physical activity, family history of hyperlipidemia and occupation noise between the two groups.

Independent Variable Selection

The significant variables of univariate analysis were used for multicollinearity diagnosis, and age as an influential factor of disease was also included in the analysis. The results show (**Table 2**) Tolerance > 0.1 and VIF < 10, so there is no multicollinearity among the variables. Then, these variables were analyzed by multivariate unconditional logistic regression. The results showed (**Table 3**) that marital status and educational level were the influencing factors of dyslipidemia. Meanwhile, hypertension is a risk factor for dyslipidemia, male workers have lower risk than female workers, the steel workers who don't smoke have a lower risk. The higher the BMI, the higher the risk of dyslipidemia. Literature supports (Ni et al., 2015; Pereira et al., 2015; Qi et al., 2015) that age is an influential factor of dyslipidemia, so age was included in the model as an independent variable. Finally, according to factor analysis, literature inquiry and expert consultation, seven independent variables were selected to enter the model. The seven independent variables are age, gender, marital status, educational, BMI, smoking status and hypertension.

Convolutional Neural Network Model Results

The effect error chart (**Figure 5**) of the dyslipidemia convolutional neural network prediction model shows that the minimum verification error is 0.013 when training in step 8. The goodness of fit test results of the convolutional neural network prediction model for dyslipidemia (**Figure 6**) show that the training set is 0.974, the verification set is 0.918, and the test set is 0.908. The performance metrics of the convolutional neural network model of dyslipidemia in steel workers are shown in **Table 4**. The sensitivity is 93.23, 90.00, and 89.97% in training set, test set and verification set, respectively. The specificity is 95.65, 91.26, and 93.01% in training set, test set and verification set, respectively. The accuracy is 94.72, 90.77, and 91.84% in training set, test set and verification set, respectively. The F1 score is 0.93, 0.91, and 0.89 in training set, test set and verification set, respectively. The AUC (95% CI) is 0.944 (0.936–0.952), 0.906 (0.876–0.931) and 0.915 (0.895–0.932) in training set, test set and verification set, respectively. The above results show that the convolutional neural network model is very suitable for the physical examination data of steel workers with dyslipidemia. In addition, the convolutional neural network model has a good ability to find patients with dyslipidemia and non-dyslipidemia, and has high prediction accuracy.

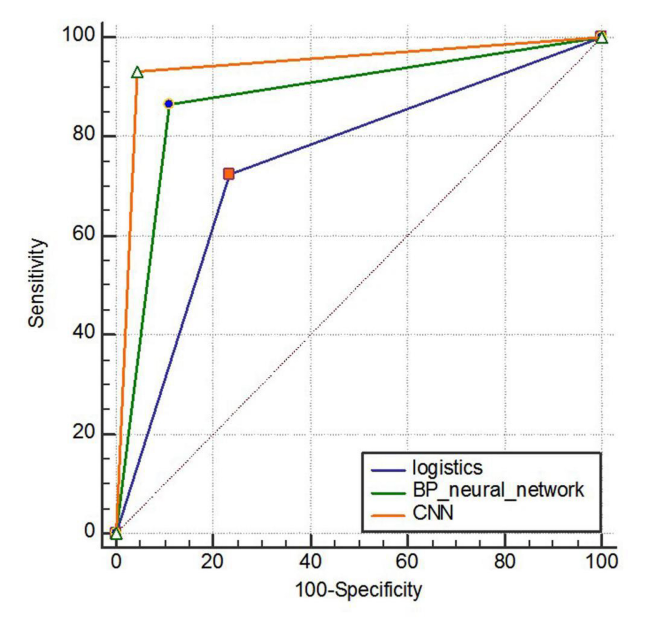


FIGURE 7 | ROC curve comparison of three model training sets.

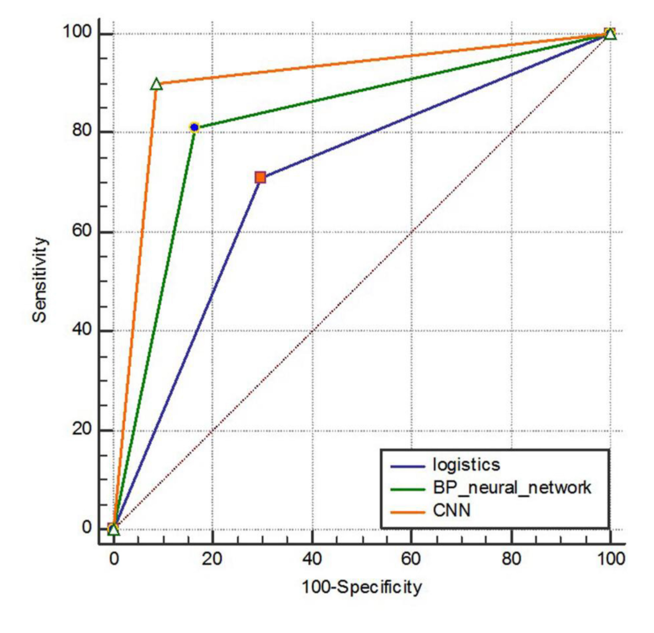


FIGURE 8 | ROC curve comparison of three model test sets.

Model Effect Comparison

We compared the prediction performance of the convolutional neural network model for dyslipidemia in steel workers with that of the Logistics regression model and BP neural network model. The comparison results of performance metrics are shown in **Table 5**.

Predictive performance results of three model training sets samples. The sensitivity of the logistic regression model, BP neural network model and convolutional neural network model is 72.45%, 86.7% and 92.23%, respectively. The specificity is 76.47, 88.96, and 95.65%, respectively. The accuracy is 74.92, 88.09, and 94.72%, respectively. The F1 score is 0.69, 0.85, and 0.93, respectively. The area under the ROC curve is shown in **Figure 7**, and the AUC (95%CI) is 0.745 (0.729–0.760), 0.878 (0.867–0.889), and 0.944 (0.936–0.952), respectively, with statistically significant differences ($P < 0.001$).

Predictive performance results of three model test sets samples. The sensitivity of the logistic regression model, BP neural network model and convolutional neural network model is 71.11, 81.11, and 90.00%, respectively. The specificity is 70.30, 83.57, and 91.26%, respectively. The accuracy is 70.60, 82.62, and 90.77%, respectively. The F1 score is 0.65,0.78 and 0.88, respectively. The area under the ROC curve is shown in **Figure 8**, and the AUC (95%CI) is 0.707 (0.663–0.748), 0.823 (0.786–0.857), and 0.906 (0.876–0.931), respectively, with statistically significant differences ($P < 0.001$).

In combination with the above performance metrics, in the prediction of dyslipidemia of steel workers, the convolutional neural network is optimal in terms of sensitivity, specificity, accuracy, F1 score and AUC. Therefore, in the prediction of dyslipidemia in steel workers, the convolutional neural network has better prediction performance.

CONCLUSION

In this work, we constructed a convolutional neural network model to predict dyslipidemia in steel workers, a special occupational group. At the beginning, we screened the data and found out the risk factors for dyslipidemia in steel workers to construct a prediction model. Subsequently, we tested the fitting degree of the model and data, and the goodness of fit in the training set, test set and verification set were 94.72, 90.77, and 91.84%, respectively. In addition, we evaluate the prediction performance of the convolution neural network model. In the training set, test set and verification set, the sensitivity is 93.23, 90.00, and 89.97%, respectively. The specificity is 95.65, 91.26, and 93.01%, respectively. The accuracy is 94.72%, 90.77% and 91.84%, respectively. The F1 score is 0.93, 0.91, and 0.89, respectively. The AUC (95% CI) is 0.944 (0.936–0.952), 0.906 (0.876–0.931) and 0.915 (0.895–0.932), respectively. The results prove that the convolutional neural network is very suitable for the prediction of dyslipidemia of steel workers and has high accuracy.

Finally, we compared the predictive performance of the convolutional neural network with the logistics model and BP neural network model of common models of dyslipidemia. We found that the predictive performance of the convolutional neural network model was better than that of the Logistics regression model and BP neural network model in the risk prediction of dyslipidemia of steel workers.

In the current study, the convolutional neural network model can accurately predict the risk of dyslipidemia in steel workers, and is superior to some existing predictive models of dyslipidemia. Therefore, the convolutional neural network model can be used to predict the risk of dyslipidemia in steel workers, and provide a basis for the formulation of early prevention strategies for dyslipidemia in steel workers, so as to improve the health status and quality of life of steel workers. In this paper, we only use the traditional convolutional neural network algorithm. So in the future, we will further study new algorithms to improve the predictive performance of the model.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of moral restrictions. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

This research was approved by the Ethics Committee of North China University of Science and Technology. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HW and SQ contributed to conception and design of the study. JW and JL organized the database. HW and HL performed the statistical analysis. SQ wrote the first draft of the manuscript. ZC, CL, and JJW wrote sections of the manuscript. JY contributed to manuscript revision. All authors agreed to submit this article.

FUNDING

This work was supported by the National Key R&D Program of China (No. 2016YFC0900605).

ACKNOWLEDGMENTS

The authors thank all the participants in this study and all the members who participated in the baseline data collection.

REFERENCES

Arun, P. V., Buddhiraju, K. M., and Porwal, A. (2018). CNN based sub-pixel mapping for hyperspectral images. *Neurocomputing* 311, 51–64. doi: 10.1016/j.neucom.2018.05.051

Chauhan, A., Anand, T., Kishore, J., Danielsen, T. E., and Ingle, G. K. (2014). Occupational hazard exposure and general health profile of welders in rural Delhi. *Indian J. Occup. Environ. Med.* 18, 21–26. doi: 10.4103/0019-5278.134953

Hedén Stahl, C., Novak, M., Hansson, P. O., Lappas, G., Wilhelmsen, L., and Rosengren, A. (2014). Incidence of Type 2 diabetes among occupational classes in Sweden: a 35-year follow-up cohort study in middle-aged men. *Diabet. Med.* 31, 674–680. doi: 10.1111/dme.12405

Hendrani, A. D., Adesiyun, T., Quispe, R., Jones, S. R., Stone, N. J., Blumenthal, R. S., et al. (2016). Dyslipidemia management in primary prevention of cardiovascular disease: current guidelines and strategies. *World J. Cardiol.* 8, 201–210. doi: 10.4330/wjc.v8.i2.201

Horiuchi, Y., Aoyama, K., Tokai, Y., Hirasawa, T., Yoshimizu, S., Ishiyama, A., et al. (2019). Convolutional neural network for differentiating gastric cancer from gastritis using magnified endoscopy with narrow band imaging. *Dig. Dis. Sci.* 65, 1355–1363. doi: 10.1007/s10620-019-05862-6

Iqbal, S., Ghani, M. U., Saba, T., and Rehman, A. (2018). Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN). *Microsc. Res. Tech.* 81, 419–427. doi: 10.1002/jemt.22994

Joint committee for guideline revision (2018). 2016 Chinese guidelines for the management of dyslipidemia in adults. *J. Geriatr. Cardiol.* 15, 1–29. doi: 10.11909/j.issn.1671-5411.2018.01.011

Keshari, R., Vatsa, M., Singh, R., and Noore, A. (2018). "Learning structure and strength of cnn filters for small sample size training," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ.

Lee, J. H., Kim, D. H., Jeong, S. N., and Choi, S. H. (2018). Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *J. Periodont. Implant Sci.* 48, 114–123. doi: 10.5051/jpis.2018.48.2.114

Lee, J. S., Chang, P. Y., Zhang, Y., Kizer, J. R., Best, L. G., and Howard, B. V. (2017). Triglyceride and HDL-C dyslipidemia and risks of coronary heart disease and ischemic stroke by glycemic dysregulation status: the strong heart study. *Diabetes Care* 40, 529–537. doi: 10.2337/dc16-1958

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., et al. (2018). Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Front. Neurosci.* 12:777. doi: 10.3389/fnins.2018.00777

Liu, M. M., Wen, L., Liu, Y. J., Cai, Q., Li, L. T., and Cai, Y. M. (2018). Application of data mining methods to improve screening for the risk of early gastric cancer. *BMC Med. Inform. Decis. Mak.* 18:121. doi: 10.1186/s12911-018-0689-4

Miller, M. (2009). Dyslipidemia and cardiovascular risk: the importance of early prevention. *QJM* 102, 657–667. doi: 10.1093/qjmed/hcp065

Ni, W. Q., Liu, X. L., Zhuo, Z. P., Yuan, X. L., Song, J. P., Chi, H. S., et al. (2015). Serum lipids and associated factors of dyslipidemia in the adult population in Shenzhen. *Lipids Health Dis.* 14:71. doi: 10.1186/s12944-015-0073-7

Pereira, L. P., Sichieri, R., Segri, N. J., da Silva, R. M., and Ferreira, M. G. (2015). Self-reported dyslipidemia in central-west Brazil: prevalence and associated factors. *Cien Saude Colet.* 20, 1815–1824. doi: 10.1590/1413-81232015206.16312014

Pikula, A., Beiser, A. S., Wang, J., Himali, J. J., Kelly-Hayes, M., Kase, C. S., et al. (2015). Lipid and lipoprotein measurements and the risk of ischemic vascular events: framingham study. *Neurology* 84, 472–479. doi: 10.1212/WNL.0000000000001202

Qi, L., Ding, X., Tang, W., Li, Q., Mao, D., and Wang, Y. (2015). Prevalence and risk factors associated with dyslipidemia in chongqing, China. *Int. J. Environ. Res. Public Health* 12, 13455–13465. doi: 10.3390/ijerph121013455

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., et al. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* 70, 1–25. doi: 10.1016/j.jacc.2017.04.052

Stevens, W., Peneva, D., Li, J. Z., Liu, L. Z., Liu, G., Gao, R., et al. (2016). Estimating the future burden of cardiovascular disease and the value of lipid and blood pressure control therapies in China. *BMC Health Serv. Res.* 16:175. doi: 10.1186/s12913-016-1420-8

Tong, J., Wang, Y., Yuan, J., Yang, J., Wang, Z., Zheng, Y., et al. (2017). Effect of interaction between noise and A1166C site of AT1R Gene polymorphism on essential hypertension in an iron and steel enterprise workers. *J. Occup. Environ. Med.* 59, 412–416. doi: 10.1097/JOM.0000000000000970

Wang, C., Li, Y., Wang, L., Li, L., Guo, Y., Zhang, L., et al. (2012). Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents. *PLoS One* 7:e43834. doi: 10.1371/journal.pone.0043834

Wu, J., Li, J., Wang, J., Zhang, L., and Yuan, J. (2019a). Risk prediction of type 2 diabetes in steel workers based on convolutional neural network. *Neural Comput. Appl.* 3, 1–16.

Wu, J., Wei, W., Zhang, L., Wang, J., Robertas, D., Li, J., et al. (2019b). Risk assessment of hypertension in steel workers based on LVQ and fisher-SVM deep excavation. *IEEE Access.* 7, 23109–23119. doi: 10.1109/access.2019.2899625

Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., and Wang, X. (2015). Word embedding composition for data imbalances in sentiment and emotion classification. *Cognit. Comput.* 7, 226–240. doi: 10.1007/s12559-015-9319-y

Yang, X., Xu, C., Wang, Y., Cao, C., Tao, Q., Zhan, S., et al. (2018). Risk prediction model of dyslipidaemia over a 5-year period based on the Taiwan MJ health check-up longitudinal database. *Lipids Health Dis.* 17:259. doi: 10.1186/s12944-018-0906-2

Yao, L., Zhong, Y., Wu, J., Zhang, G., Chen, L., Guan, P., et al. (2019). Multivariable logistic regression and back propagation artificial neural network to predict diabetic retinopathy. *Diabetes Metab. Syndr. Obes.* 12, 1943–1951. doi: 10.2147/DMSO.S219842

Zhang, X., Tang, F., Ji, J., Han, W., and Lu, P. (2019). Risk prediction of dyslipidemia for chinese han adults using random forest survival model. *Clin. Epidemiol.* 11, 1047–1055. doi: 10.2147/CLEP.S223694

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 Public Health

ORIGINAL RESEARCH
published: 03 November 2020
doi: 10.3389/fpubh.2020.584387

Alzheimer's Disease Classification With a Cascade Neural Network

Zeng You[1,2†], *Runhao Zeng*[2†], *Xiaoyong Lan*[1†], *Huixia Ren*[1,3], *Zhiyang You*[1,2], *Xue Shi*[1], *Shipeng Zhao*[1,2], *Yi Guo*[1*], *Xin Jiang*[4*] *and Xiping Hu*[2*]

[1] Department of Neurology, Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology, The Second Clinical Medical College of Jinan University, Shenzhen, China, [2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, [3] The First Affiliated Hospital, Jinan University, Guangzhou, China, [4] Department of Geriatrics, Shenzhen People's Hospital, The First Affiliated Hospital of Southern University of Science and Technology, The Second Clinical Medical College of Jinan University, Shenzhen, China

OPEN ACCESS

Edited by:
Ying Li,
Zhejiang University, China

Reviewed by:
Laisen Nie,
Northwestern Polytechnical University, China
Weijing Qi,
Chongqing University of Posts and Telecommunications, China

****Correspondence:***
Yi Guo
xuanyi_guo@163.com
Xin Jiang
jiangxinsz@163.com
Xiping Hu
xp.hu@siat.ac.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *17 July 2020*
Accepted: *28 September 2020*
Published: *03 November 2020*

Citation:
You Z, Zeng R, Lan X, Ren H, You Z, Shi X, Zhao S, Guo Y, Jiang X and Hu X (2020) Alzheimer's Disease Classification With a Cascade Neural Network. Front. Public Health 8:584387. doi: 10.3389/fpubh.2020.584387

Classification of Alzheimer's Disease (AD) has been becoming a hot issue along with the rapidly increasing number of patients. This task remains tremendously challenging due to the limited data and the difficulties in detecting mild cognitive impairment (MCI). Existing methods use gait [or EEG (electroencephalogram)] data only to tackle this task. Although the gait data acquisition procedure is cheap and simple, the methods relying on gait data often fail to detect the slight difference between MCI and AD. The methods that use EEG data can detect the difference more precisely, but collecting EEG data from both HC (health controls) and patients is very time-consuming. More critically, these methods often convert EEG records into the frequency domain and thus inevitably lose the spatial and temporal information, which is essential to capture the connectivity and synchronization among different brain regions. This paper proposes a cascade neural network with two steps to achieve a faster and more accurate AD classification by exploiting gait and EEG data simultaneously. In the first step, we propose attention-based spatial temporal graph convolutional networks to extract the features from the skeleton sequences (i.e., gait) captured by Kinect (a commonly used sensor) to distinguish between HC and patients. In the second step, we propose spatial temporal convolutional networks to fully exploit the spatial and temporal information of EEG data and classify the patients into MCI or AD eventually. We collect gait and EEG data from 35 cognitively health controls, 35 MCI, and 17 AD patients to evaluate our proposed method. Experimental results show that our method significantly outperforms other AD diagnosis methods (91.07 vs. 68.18%) in the three-way AD classification task (HC, MCI, and AD). Moreover, we empirically found that the lower body and right upper limb are more important for the early diagnosis of AD than other body parts. We believe this interesting finding can be helpful for clinical researches.

Keywords: Alzheimer's disease, deep learning, automatic diagnosis, gait, EEG

1. INTRODUCTION

Alzheimer's disease (AD) is the most common cause of cognitive impairment and is one of the diseases with the highest incidence among the elderly. In 2006, 26.6 million people on the earth suffered from AD, and the number is still rapidly increasing every year (1). More critically, AD has become the seventh leading cause of death (2). Conventional AD diagnosis methods often use scale screening and brain imaging equipment such as functional Magnetic Resonance Imaging (fMRI), Computer Tomography (CT), and Positron Emission Tomography (PET). These methods require experienced clinicians as well as exhaustive examinations.

Recently, many studies (3–9) have been conducted to reduce the diagnosis cost and shorten the diagnosis time by designing an AD classification system that is able to detect and classify AD automatically. However, it is challenging to classify AD precisely for the following reasons: on the one hand, the prodromal stage of AD, namely mild cognitive impairment (MCI), has a light symptom, making it hard to detect; On the other hand, extracting robust features for AD detection is very challenging due to the limited volume of medical data.

Previous studies on AD classification exploit gait data (3, 10–17) due to the strong relationship between gait features and cognitive function (18–25). They often extract hand-crafted features from the input gait data (e.g., skeleton) and classify AD relying on these features. However, designing hand-crafted features for AD classification requires expert knowledge, and it is difficult to generalize the hand-crafted features to other tasks. Recently, some researchers (12, 13, 15, 16, 26, 27) attempt to conduct AD classification using EEG data. However, existing EEG-based methods often (6, 7) need to convert EEG data into frequency domain information and calculate the Power Spectral Density (PSD) features for classification. In this sense, these methods will inevitably lose the information in the spatial and time domains of EEG data, which, however, is very important for capturing the coherence and synchronizations among different brain regions. It is worth noting that existing methods use one modal only (gait or EEG data) and suffer from the following limitations: (1) as discussed in (28, 29), using gait data can accurately distinguish HC and patients but often fails to classify MCI and AD, and (2) using EEG data can classify MCI and AD more accurately, but it is time-consuming to collect EEG data from both HC and patients.

We contend that considering the two modalities (i.e., gait and EEG data) simultaneously helps achieving faster and more accurate classification. To this end, we propose a cascade neural network with two steps for the early diagnosis of AD using both gait data and EEG data simultaneously. **In the first step**, we use gait data to classify HC and patients. For the purpose of reducing the psychological disturbance to the subject, we follow (10) to use the Kinect devices as the acquisition equipment to capture skeleton sequences. Regrading the non-Euclidean skeleton data, we propose to use attention-based spatial temporal graph convolutional networks (AST-GCN) to model the relationships among body key points and automatically extract powerful features for distinguishing between HC and patients. **In the second step**, we use the original EEG data to distinguish MCI and AD patients further. Unlike other methods that convert EEG data to the frequency domain, we propose spatial temporal convolutional networks (ST-CNN) to directly extract the spatial and temporal features from original EEG data and use them to classify MCI and AD. In this manner, the EEG data from HC are no longer required, saving a lot of data collection time. We collect a data set consisting of gait and EEG data from 35 cognitively health controls, 35 MCI patients, and 17 AD patients to evaluate our proposed method.

Our main contributions are summarized as follows:

- We propose a cascade neural network that uses both gait and EEG data to classify AD, which achieves a high accuracy rate with less manual participation. This is the first attempt to consider two modalities for AD classification to the best of our knowledge.
- We propose attention-based spatial temporal graph convolutional networks to automatically extract the features from gait data and leverage them to classify AD.
- Moreover, we also propose spatial temporal convolutional networks to fully extract the spatial and temporal features from the original EEG data in both space and time domains.
- The accuracy rate of our proposed cascade neural network in the three-way classification of HC, MCI, and AD reaches 91.07%, which is much higher than the method using one modal (68.18%). The accuracy of HC vs. MCI/AD is up to 93.09%.

The rest of the paper is arranged as follows: Related work is concentrated on section 2; Section 3 details the proposed framework and the modules in it; Experimental results are exhibited in section 4; Section 5 concludes this paper.

2. RELATED WORK

Gait data has been used extensively to classify AD. Wang et al. (3) developed a device to collect the inertial signals of subjects. They designed an algorithm to leverage the inertial signals to detect and calculated the features of the stride. Then they selected the salient features to classify HC and AD. The classification accuracy rates in the female and the male groups are 70.00 and 63.33%, respectively. Choi et al. (29) compared the gait and cognitive function between the HC group and MCI/AD groups. They found that gait features can distinguish MCI and HC, while cognitive tests are suitable for distinguishing AD and HC. The average detection rate of AD and MCI from HC using gait variables is 75%. Seifallahi et al. (10) used Kinect to collect gait data, extracted, and screened the features. Then they used Support Vector Machines (SVM) to classify AD and HC. The classification accuracy rate is 92.31%. Varatharajan et al. (4) used IoT devices to collect gait data and then extracted the features using the dynamic time warping (DTW) algorithm. The accuracy rate of classification is about 70%. Although the above works achieve good performance, they all rely on handcrafted feature extraction, which cannot guarantee the full use of the implicit information in gait data, and the features designed for specific tasks cannot be applied to other general tasks. The attention-based spatial temporal graph convolutional networks we proposed can automatically extract gait data features and exploit the relationships among body joints.

EEG data is another important information that can be used to diagnose AD. Existing methods for the early diagnosis of AD using EEG data can be categorized into **handcrafted feature based-methods** and **deep learning methods**. Anderer et al. (12) and Pritchard et al. (13) input EEG markers into an ANN to perform a binary classification between AD and HC with an accuracy rate of 90%. Trambaiolli et al. (15) extracted features based on coherence and used Support Vector Machines(SVM) to classify AD and HC, with 79.9% accuracy. Rossini et al. (16) tested the IFAST procedure to classify HC and MCI, achieving 93.46% accuracy. These methods all require handcrafted feature extraction. In recent years, more and more deep learning methods have been applied to the classification of AD. Ieracitano et al. (6) calculated the PSD features of the subject's EEG data. They converted the PSD features into images, and then used the convolutional neural networks for the early diagnosis of AD, achieving an accuracy of 89.8% in the binary classification and 83.3% in three-way classification. Bi and Wang (7) calculated the PSD features of EEG data, then used the feature representation method proposed by (30) to convert the PSD features into images. They designed a DCssCDBM with a multi-task learning framework, achieving an accuracy of up to 95.05%. These deep learning methods all need to convert EEG data into frequency domain information. This way will lose the information in the spatial and temporal domains of EEG data, which is essential for capturing coherence and synchronization among different brain regions. We directly use the original EEG data containing both spatial and temporal information. We propose spatial temporal convolutional networks to extract the temporal and spatial implicit features of EEG data.

The methods mentioned above leveraged either gait data or EEG data only for the early diagnosis of AD. The gait data collection procedure is simple, short in time, and easy to operate, but there is no significant difference in gait features between MCI and AD (29), and thus method relying on gait data cannot classify AD and MCI precisely. Conversely, EEG data can provide promising cues to classify AD and MCI, but the acquisition process is complicated and takes a long time. We consider gait and EEG data simultaneously to achieve a fast and accurate classification of AD.

3. PROPOSED METHOD

Notation. Let $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$ be the subject set that includes N_s subjects, where s_i represents the i^{th} subject. Let $\mathcal{G}_i = \{g_i^j\}_{j=1}^{N_g}$ denote clip set where N_g clips are sampled from the gait data of the i^{th} subject s_i, where g_i^j represents the j^{th} clip. Let $\mathcal{E}_i = \{\varepsilon_i^e\}_{e=1}^{N_e}$ denote the epoch set containing N_e epochs sampled from the EEG data of the i^{th} subject s_i, where ε_i^e represents the e^{th} epoch.

Problem Definition. Given gait clip set $\mathcal{G}_i$ and EEG epoch set $\mathcal{E}_i$ of subject s_i, the classification of AD aims to map physiological signals, $\mathcal{G}_i$ and $\mathcal{E}_i$, into HC, MCI, and AD groups corresponding to the state of subject s_i. This task is very challenging due to the limited volume of data and the subtle differences among the three groups, especially for HC and MCI.

3.1. Pipeline Overview

Existing methods used either gait data or EEG data only for the classification of AD. However, as discussed in (28, 29), using gait data can accurately distinguish HC and patients, but the methods using gait data only often fail to classify MCI and AD. For the EEG data that are more sensitive to the differences between MCI and AD, some studies used EEG data to classify AD. However, collecting EEG data from both HC and patients takes a lone time. We believe that combining the two is able to make the early diagnosis of AD faster and more accurately. This drives us to propose a cascade neural network for the early diagnosis of AD with both gait and EEG data.

Given gait clip g_i^j and EEG epoch ε_i^e of subject s_i, we conduct the classification in two steps. Firstly, we use gait data to distinguish HC and MCI/AD patients. In this step, we select key points from g_i^j to form key-point skeleton sequences first. Then we input the key-point skeleton sequences into attention-based spatial temporal graph convolutional networks (AST-GCN) to extract features. Finally, we use these features to classify HC and MCI/AD by a standard SoftMax classifier. We further distinguish AD from MCI with EEG epoch ε_i^e in the second step. We input ε_i^e into the spatial temporal convolutional networks (ST-CNN) to extract the implicit features in spatial and temporal domain. We then used the features extracted by ST-CNN for the binary classification of MCI vs. AD. In our method, the EEG data from HC are not required. The architecture of our proposed framework is shown in **Figure 1**.

3.2. Attention-Based Spatial Temporal Graph Convolutional Networks

Existing methods that use gait data for the early diagnosis of AD rely on handcrafted features, which are inefficient and cannot fully use implicit information in gait data. We need to automatically extract the implicit features in gait data for the early diagnosis of AD, which is the strength of deep learning. Our gait data is composed of skeleton sequences recognized by the Kinect devices. Traditional deep learning methods such as convolutional networks cannot handle such non-Euclidean data. The ST-GCN proposed by (31) shows an excellent performance in extracting the features from skeleton sequences. We apply it as our basic model to the classification of AD and propose attention-based spatial temporal graph convolutional networks (AST-GCN) according to our task and data characteristics. Based on clinical experience and experimental comparison results, we found that different body parts have different importance in the classification of AD. For this reason, given skeleton sequences, we first perform key point filtering to form our key-point skeleton sequences and then input it into the proposed attention-based spatial temporal graph convolution networks. The extracted spatial and temporal features are finally used for classification. In the next few subsections, we will first briefly introduce ST-GCN, then we will introduce how we do key point filtering and the proposed attention-based spatial temporal graph convolutional networks.

3.2.1. Spatial Temporal Graph Convolutional Networks

Firstly, a spatial temporal graph is constructed from skeleton sequences, as shown in **Figure 2A**. The edges of the spatial temporal graph consist of two parts. One part is the natural

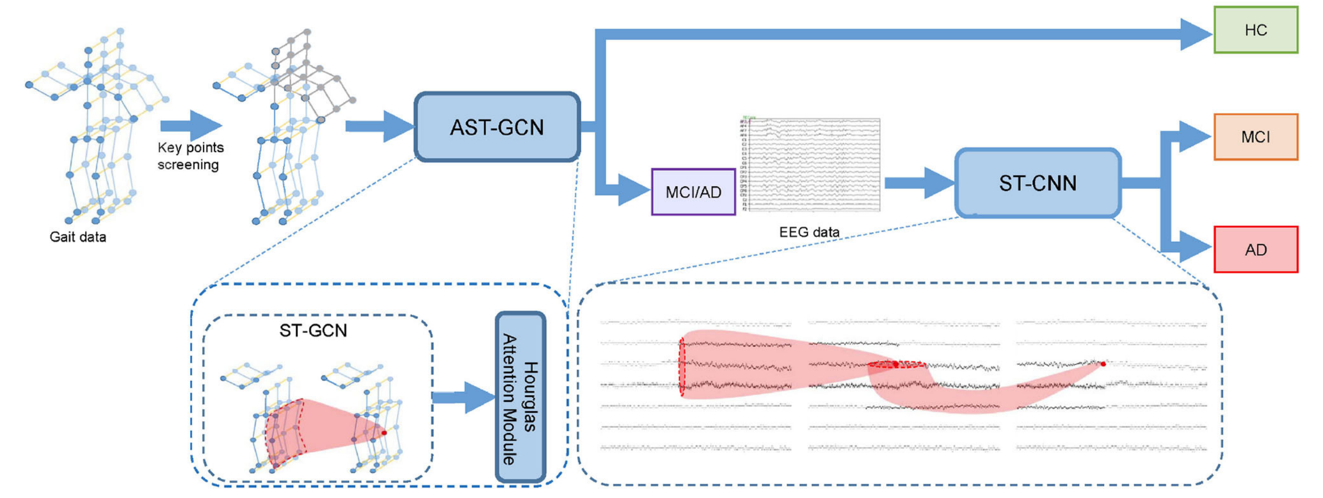


FIGURE 1 | Cascade neural network for the early diagnosis of AD. We perform key point screening on gait data to form key-point skeleton sequences. Then we use attention-based spatial temporal graph convolutional networks (AST-GCN) to extract features and classify the subject into HC or MCI/AD with features. If the subject is classified into MCI/AD, we will input the EEG data into spatial temporal convolutional networks (ST-CNN) to extract features and perform MCI vs. AD binary classification.

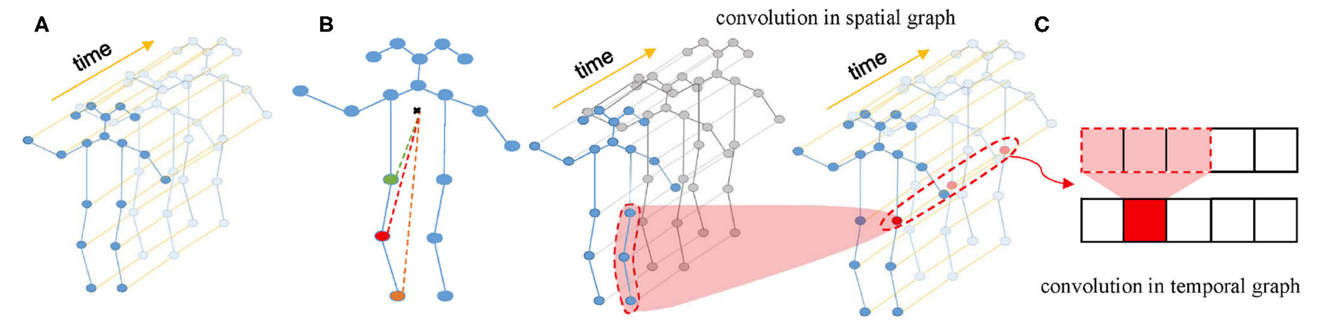


FIGURE 2 | (A) Spatial temporal graph of skeleton sequences. **(B)** The "Spatial Configuration" strategy. **(C)** The architecture of ST-GCN.

connections between joint points of the human skeleton in a single frame called spatial edges, and the other part is the time edges formed by connecting the same joint points between adjacent frames. Then, the input features composed of the coordinate vectors of the nodes in the graph are inputted into multiple layers of spatial-temporal graph convolution. Defining the weight function of the graph convolution operation can be realized by a variety of strategies for partitioning each node's neighborhood point set. Experiments show that the "Spatial Configuration" strategy, as shown in **Figure 2B**, works best. According to this strategy, the neighborhood point set of the root node (red node) is divided into three subsets, namely: (1) The root node itself (red node); (2) The centripetal group (orange node): the nodes closer to the gravity center of the skeleton than the root node; (3) centrifugal group (green node): the nodes that are farther from the gravity center of the skeleton than the root node. The formula of space graph convolution can be written as:

$$f_{out} = \Lambda^{-\frac{1}{2}} \left(\mathbf{A}+\mathbf{I}\right) \Lambda^{-\frac{1}{2}} f_{in}\mathbf{W}, \qquad (1)$$

where f_{in} denotes the feature map of the clip composed of the coordinates of input skeleton sequences, which is a $D \times T \times V$ matrix, where $D = 3$ corresponds to Three coordinates (x, y, z), T represents the time points i.e., the number of frames of the skeleton sequences, V is the number of nodes that constitute the spatial graph in each frame. **W** is the weight function; Λ is the degree matrix of the spatial graph; **A** is the adjacency matrix of the spatial graph; **I** is the self-connection matrix. Moreover, **M** is proposed as a learnable edge weight, which has the same size as the adjacency matrix. It is used in every layer of spatial temporal graph convolution. Then the Equation (1) can be written as:

$$f_{out} = \Lambda^{-\frac{1}{2}} \left(\left(\mathbf{A}+\mathbf{I}\right) \bigodot \mathbf{M}\right) \Lambda^{-\frac{1}{2}} f_{in}\mathbf{W}, \qquad (2)$$

where $\bigodot$ notes the element-wise multiply. Spatial temporal convolution module consists of a convolution in the spatial graph and a convolution in the temporal graph. The structure of spatial temporal convolution module is shown in **Figure 2C**.

3.2.2. Key Points Filtering

Several studies (18, 20–24, 32) found that the AD group has significant differences with the HC group in gait speed, gait cadence, stride et al. This means that the joints of the lower body, such as the ankles, are more critical for the early diagnosis of AD. Besides, Most subjects are right-handed. It is clinically believed that the left hemisphere of right-handed patients is more sensitive to AD and more likely to be affected. When we observe the learnable parameter **M** of the basic model after it converges, we find that the connections among the joint points of the lower body and the right upper limb are given higher weights, which means that these joint points are more important than other parts. Through experimental comparison, we also verified that performance classification with the skeleton sequences composed of the joint points of the lower body and the right upper limb are better than that with the skeleton sequences composed of other parts. Therefore, we select the joint points of the lower body and the right upper limb to form key-point skeleton sequences.

3.2.3. Hourglass Attention Module

From the description above, we can see that different parts are of different importance for the early diagnosis of AD. We argue that even in the key-point skeleton sequences we construct, joints in some parts are more important than other parts, such as ankles and wrists. Therefore, to drive the model further focus on important joints, we introduced an hourglass attention module with a structure similar to the attention module in (33). However, we replaced the pooling layer with a convolutional layer in the time domain with a stride of 4. The structure of the hourglass attention module is shown in **Figure 3**.

3.3. Spatial Temporal Convolutional Networks

Existing deep learning methods that use EEG data for the classification of AD convert EEG data into frequency domain information, then calculate PSD features and convert them into images. This way will lose the information in the time domain or even the spatial domain, which is essential to capture coherence and synchronization among different brain regions. The EEGnet proposed by (34) extracts the temporal and spatial features of original EEG data to recognize task-state EEG and shows good performance. However, its feature extraction in the spatial domain of EEG data simply uses a convolution layer to map the data to a single value. We believe that this is not able to fully extract the spatial features of EEG data. We propose the spatial temporal convolutional networks to extract features from original EEG data. Each ST-CNN module consists of a spatial convolution layer with a kernel size of $K_s \times 1$ and a temporal convolution layer with a kernel size of $1 \times K_t$ similar to (31). In this way, the EEG data is alternately convoluted in the space domain and the time domain through multiple ST-CNN layers to fully extract the implicit features in space and time. The structure of spatial temporal convolutional networks is shown in **Table 1**.

TABLE 1 | The structure of spatial temporal convolutional networks, where K_s and K_t are the size of the kernel used in the spatial convolution layer and the temporal convolution layer in a ST-CNN module, respectively.

Layer	Input channels	Operation	Kernel size	Stride	Output channels
0	3	Batch normalization	–	–	3
1	3	ST-CNN	$K_s = 1, K_t = 33$	1	4
2	4	ST-CNN	$K_s = 15, K_t = 33$	4	4
3	4	ST-CNN	$K_s = C, K_t = 33$	1	16
4	16	ST-CNN	$K_s = 1, K_t = 33$	4	8
6	8	Flatten	–	–	$T/2$
Classifier	$T/2$	Full connection	–	–	N
	N	SoftMax	–	–	N

C is the number of EEG channels. T is the number of time points. N is the number of classes. In the second layer, we use depthwise separable convolutions. In the 2nd and 4th ST-CNN module, we set stride to 4 as the pooling layer. The residual mechanism is used in each ST-CNN module.

TABLE 2 | The grouping criteria for HC, MCI, AD.

	HC	MCI	AD
MoCA	> 30	$18 \sim 30$	$0 \sim 17$
MMSE	≥ 24	≥ 24	< 24

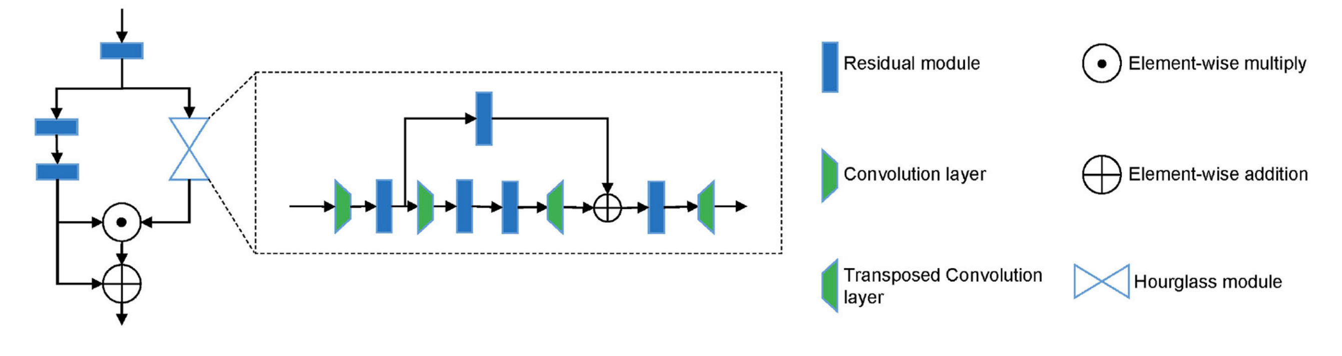


FIGURE 3 | The structure of hourglass attention module.

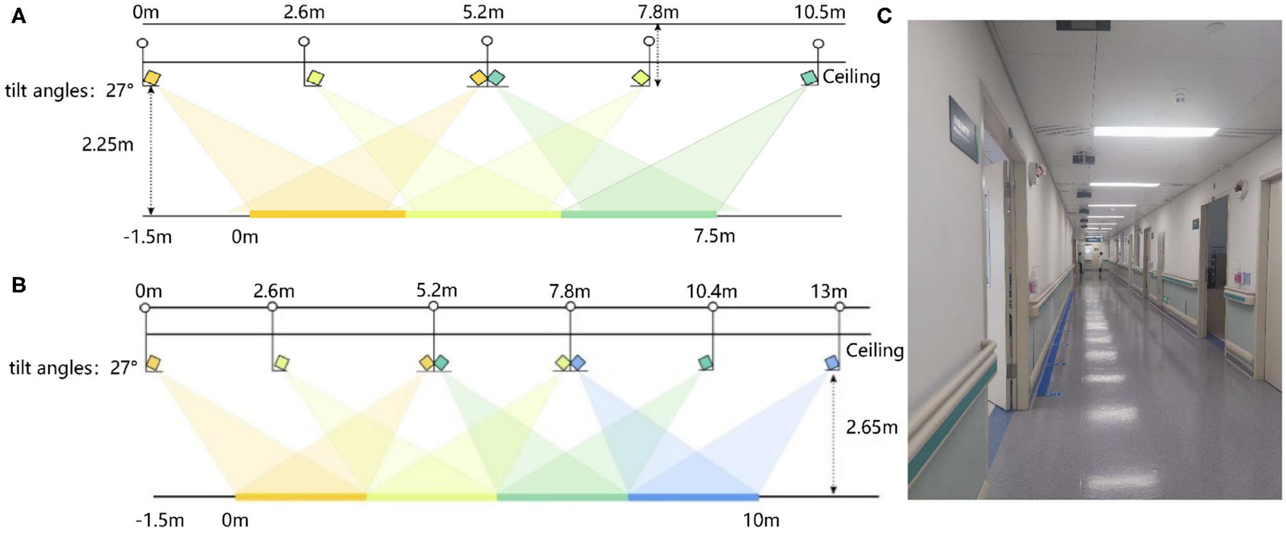


FIGURE 4 | The deployment diagram of Kinect V2.0 devices: **(A)** The deployment diagram of devices in the Neurology Department. **(B)** The deployment diagram of devices in the Geriatrics Department. **(C)** The diagram of the actual data acquisition scene.

4. EXPERIMENTS

4.1. Data Acquisition and Preprocessing

We collect gait data in cooperation with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences and the Shenzhen People's Hospital, and the EEG data are collected by the Shenzhen People's Hospital. All MCI and AD patients are diagnosed by experienced neurologists based on the Montreal Cognitive Assessment(MoCA) and Mini-Mental State Examination (MMSE). We divide the subjects into three groups: HC, MCI, and AD. These groups include 35 cognitively healthy controls, 35 MCI patients, and 17 AD patients with mild-to-severe AD, respectively. The grouping criteria are shown in **Table 2**. We collect both gait and EEG data for MCI and AD patients, and only collect gait data for cognitively healthy controls.

4.1.1. Gait Data

4.1.1.1. Data Acquisition

Gait data of 52 MCI and AD patients and 35 control subjects are collected in the Neurology and the Geriatrics Departments of Shenzhen People's Hospital, respectively. Our data collection settings are similar to (35). We use Microsoft Kinect V2.0 cameras as our data acquisition devices. The subjects are asked to walk at their natural and comfortable speed and posture under the devices. They walk a round trip on a straight path about 10 m. We deploy 8 and 6 devices in the Neurology and Geriatrics Department, respectively. The deployment diagram is shown in **Figure 4**. The tilt angle of all devices was set 27°.

4.1.1.2. Data Preprocessing

Our gait data consists of the skeleton sequences recognized by the devices. Each skeleton is composed of three-dimensional coordinates of 25 joints. Their indexes are shown in **Figure 5A**. In each recording, the devices estimate the skeleton joint coordinates from both the front and back views. However, the skeletons estimated from the back view are less accurate than those from the front view. Therefore, we only select the skeletons from the front view as gait data.

Due to the venue restrictions, the data acquisition devices for patients and the devices for heath controls are deployed in different environments, which may cause differences in absolute coordinates of key points. To eliminate these differences, we follow (36) to perform the following coordinate transformation on the collected gait data in the data preprocessing stage. Since our devices are mounted on the ceiling, and there is an angle of 27° with the horizontal, we first rotate the coordinates $\left[x, y, z\right]$ around the x-axis by $-27°$ by calculating

$$\begin{vmatrix} x' \\ y' \\ z' \end{vmatrix} = \mathbf{R_x} \times \begin{vmatrix} x \\ y \\ z \end{vmatrix}, \text{where } \mathbf{R_x} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & cos\theta & -sin\theta \\ 0 & sin\theta & cos\theta \end{vmatrix}, \ \theta = -27°. \quad (3)$$

In this way, the skeleton sequences are in a horizontal position relative to the cameras. We then move the origin of the coordinates to the base of the human spine, namely point 0, by computing

$$\mathrm{v}'_{\tau p} = \mathrm{v}_{\tau p} - \mathrm{v}_{\tau 0}, \quad (4)$$

where $\mathrm{v}_{\tau p}$ is a coordinate vector of pth joint point of the skeleton in τth frame. Moreover, the time lengths of gait records are different. Similar to (37), we intercept several clips of data from each gait record through a sliding window to make the number of clip frames consistent. We set the sliding window with a size of 60 frames and a stride of three frames. In this way, we have a total of 5,519 clips, and each gait clip g_i^j is a matrix with a dimension of $D \times T \times V$, where $D = 3$, $T = 60$, $V = 25$.

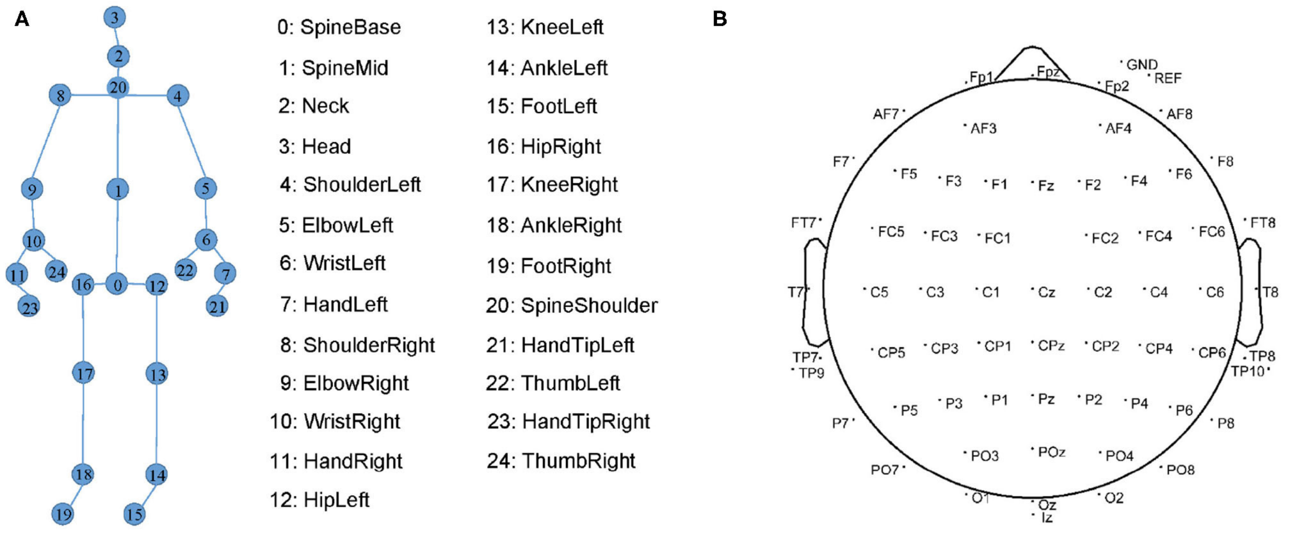


FIGURE 5 | (A) The 25 markers on human skeleton recognized by Kinect. **(B)** 64 EEG electrode locations in the International 10-20 System.

4.1.2. EEG Data

4.1.2.1. Data acquisition

The EEG data are collected by the Neurology Department, Shenzhen People's Hospital. Due to a large mount of artifacts (e.g., myoelectricity) during human walking, the collected EEG data are in low quality. We follow (6, 38) to collect higher-quality resting EEG data. We collect the EEG data of the patients with eyes closed and with eyes open for 8 min each. We place 64-channel EEG electrodes on the patient's scalp at the standard locations during data acquisition as shown in **Figure 5B**. The EEG signals are recorded at a sampling frequency of 5,000 Hz.

4.1.2.2. Data preprocessing

After EEG records are collected, we first remove artifacts from EEG records, such as electrooculograms and myoelectricity. Then we re-reference the data. The EEG signals of the Ref and Gnd electrodes are removed, and the average value of the remaining 62 channels is used as a reference value to recalculate the value of the EEG data. Using the original EEG data with a sampling rate of 5,000 Hz in our ST-CNN will inevitably incur large computation cost. Specifically, the input size is 5,000 × 62 when the epoch duration is set to one second. In this paper, we follow Toll et al. (38) to downsample the EEG data to 250 Hz, aiming to reduce the computation cost and improve the inference speed. Similar to (7), we then intercept 120 epochs from each subject's EEG data by a sliding window without overlapping. We set the sliding window with a size of 256, which is about 1 s. The epochs sampled from the data collected with the eyes open and the eyes closed are concatenated in the time dimension. Finally, we copy it for three times in depth dimension. In this way, we have a total of 5,519 epochs, and each epoch ε_i^e is a $3 \times C \times 2T$ matrix, where $C = 62$ is the number of channels of EEG data, $T = 256$ denotes the number of time points.

4.2. Implementation Details

We randomly select 75% of the subjects. We use their corresponding data clips as our training set, including 3,277 data clips. The remaining data clips serve as our test set, including 2,242 data clips. We train the model for 50 epochs, using a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.05 and a batch size of 64. All experiments are conducted on a single GTX 1060 GPU.

As for EEG data, we randomly select 75% of the EEG epochs as our training set, containing 4,680 epochs, and the remaining EEG epochs serve as our test set, including 1,560 epochs. We train the model for 70 epochs, using a stochastic gradient descent optimizer with an initial learning rate of 0.005, and with a batch size of 64. All experiments are conducted on a single GTX 1060 GPU.

4.3. Comparisons With Other AD Diagnosis Methods

We compare our proposed method with other existing methods. The results is listed in **Table 3**. Firstly, we compare our proposed attention-based spatial temporal graph convolutional networks with the methods using handcrafted features. We extract the same features as (10) from gait data and feed them into a SVM classifier with the Gaussian (RBF) kernel and a random forest classifier, respectively. The accuracy of the two classifiers are much lower than our proposed attention-based spatial temporal graph convolutional networks (93.09%). These results demonstrate that our proposed attention-based spatial Temporal graph convolutional networks is able to extract more powerful features for the diagnosis of AD.

Then we compare the proposed spatial temporal convolutional networks with several baselines on the collected EEG dataset. The baselines include EEGnet (34), ResNet-18 (39), VGG-13 (40), and the standard convolution networks. standard

TABLE 3 | Comparison with other methods.

Methods	Data		Accuracy		
	Gait	EEG	HC vs. MCI/AD (%)	MCI vs. AD (%)	Three-way classification (%)
Handcrafted features + SVM	✓		63.64	57.73	55.45
Handcrafted feature + RF	✓		81.82	57.14	68.18
AST-GCN(ours)	✓		93.09	58.41	68.51
standard CNN		✓	–	69.66	–
EEGnet		✓	–	97.85	–
ResNet 18		✓	–	97.59	–
VGG 13		✓	–	96.48	–
ST-CNN(ours)		✓	–	98.63	–
cascade neural network(ours)	✓	✓	**93.09**	**98.63**	**91.07**

Standard CNN represents the model we substitute 2D convolution layers with a kernel size of $K_s \times K_t$ for ST-CNN modules. "Handcrafted features + SVM" and "Handcrafted features + RF" indicate the methods using different classifiers with the handcrafted features same as (10). The bold values indicates the best performance that method obtain in that experiment.

TABLE 4 | Ablation study of key point filtering and hourglass attention module on gait data.

Components		Accuracy (%)
Key point filtering	Hourglass attention module	
×	×	88.18
✓	×	91.97
×	✓	90.14
✓	✓	93.09

convolutional networks share the same architecture as the spatial temporal convolutional networks but all ST-CNN modules are replaced with 2D convolution layers with a kernel size of $K_s \times K_t$. It is observed that our model achieves the best performance on our data set. We believe that the reason is that ST-CNN can extract the spatial and temporal features from EEG data better. Finally, we test our proposed neural network on our test set. The accuracy of binary classification is 93.09%, and the accuracy of the three-way classification is 91.07%. In addition, we introduce a voting mechanism to improve the fault tolerance of the entire framework. We randomly select a subject s_i from the test set and input his gait clip set $\mathcal{G}_i$ into AST-GCN for classification. If more than 50% of the clips are classified into MCI and AD, all the EEG epochs in $\mathcal{E}_i$ will be inputted into ST-CNN to perform binary classification of MCI vs. AD. Otherwise, s_i is finally classified into HC. If more than half of the epochs are classified into MCI(AD), then s_i is finally classified into MCI(AD). With the voting mechanism, the framework can achieve an accuracy of 100% on the binary classification of HC vs. MCI/AD, and accuracy of 99.14% on the three-way classification of HC, MCI, and AD.

4.4. Ablation Studies

4.4.1. The Effectiveness of the Proposed Component

We conduct experiments on gait data to study the effectiveness of key point filtering and the hourglass attention module. In **Table 4**, we observe that these two components increase the accuracy from 88.18 to 91.97% and 90.14%, respectively. With both components, we achieve the best performance with an accuracy rate of 93.09%. We believe the reason is that both components can guide the model to focus more on the points more critical to the diagnostic task. Key point filtering removes insignificant points and noise points, and the attention module drives the model to further focus on the important points in key points.

4.4.2. Which Key Points Are Essential for AD Diagnosis?

In **Figure 6A**, we compare the performance of the skeleton sequences of the lower body, the upper body, and the whole body. We find that the whole body joint performs best. We consider that this is because all joints can provide more information for diagnosis. In addition, we observe that the lower body joints perform better than upper body joints. We believe the reason is that the behavior of lower body is more relative to early AD diagnosis.

Clinically, it is believed that the left hemisphere of right-handed patients is more sensitive and easier to be affected by AD. As the left hemisphere controls the movement of the right body part, for the right-handed patients, their behaviors of the right body part may provide more information for AD diagnosis. To study this empirically, we further divide the body joints into two more fine-grained groups, namely "lower body + right upper limb" and "lower body + left upper limb." All subjects in the collected dataset are right-handed. In **Figure 6B**, "lower body + right upper limb" performs best. these results are consistent with the clinical perspective. Based on such observation, we select the skeleton sequence of "lower body + right upper limb" as a default setting in all experiments.

4.4.3. Where Should We Use the Hourglass Attention Module?

We explore the performance of our model with different placements of the attention module. We try to add the hourglass attention module after the third, sixth, and ninth layer of the basic model, respectively, and add three hourglass attention modules after the 3rd, 6th, and 9th layers. The experimental results are shown in **Table 5**. We see that using three attention modules additionally includes 67.78% parameters more than

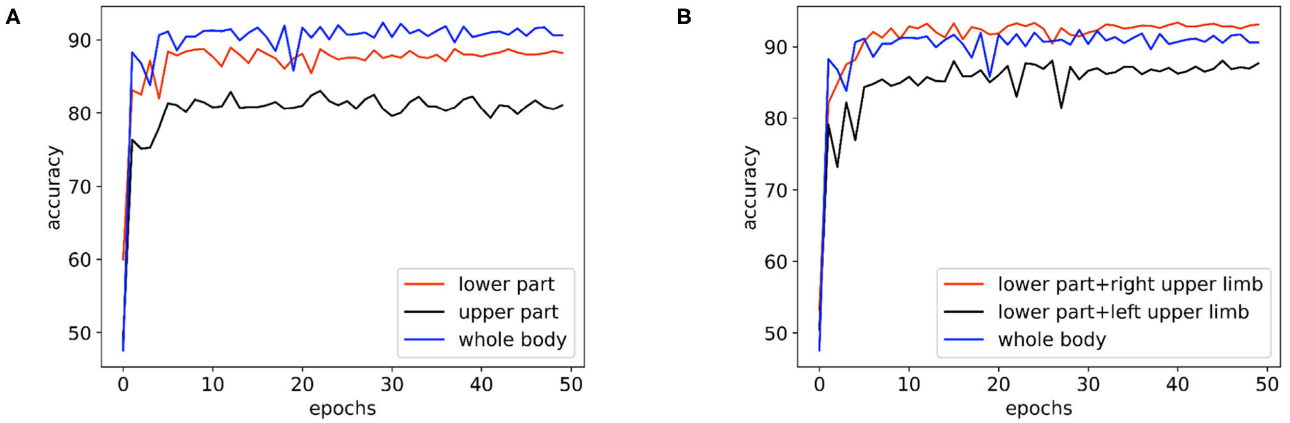


FIGURE 6 | The performance comparison of the basic model on the skeleton sequences composed of different parts: **(A)** The performance of the basic model on the skeleton sequences composed of the lower body, the upper body, and the whole body. **(B)** The performance of the basic model on the datasets of skeleton sequences composed of the whole body, the lower body + the right upper limb, and the lower body + the left upper limb.

TABLE 5 | Performance comparison of the models with different hourglass attention module locations.

	Basic model (%)	After 3rd layer (%)	After 6th layer (%)	After 9th layer	After 3rd,6th,9th layers (%)
Accuracy	88.18	88.76	88.22	**90.14**	87.97

The bold values indicates the best performance that method obtain in that experiment.

TABLE 6 | Comparison of the performance and inference speed with different models.

Cascade stage		Accuracy(%)	No. of parameters	Inference speed (ms)
Stage 1	Stage 2			
AST-GCN (gait)	AST-GCN (gait)	74.46	9.42M	7.06
AST-GCN (gait)	ST-CNN (EEG)	**91.07**	4.72M(4.71M+0.01M)	3.99

The bold values indicates the best performance that method obtain in that experiment.

using one attention module while decreasing the performance. It is worth nothing that the model with three attention modules outperforms that with one attention module (99.75 vs. 98.04%) in the training phase, but it leads to a worse accuracy (87.97 vs. 90.14%) in the testing phase. We conjecture that adding three attention modules may incur the overfitting issue since a larger network is more likely to lead to overfitting in the case of a limited amount of data (41). We see that adding one attention module after the ninth layer of the basic model achieves the best performance. Therefore, we use the model with an attention module after 9th as the default setting.

4.4.4. The Efficiency of Our Method

We conduct an ablation study to validate the effectiveness and efficiency of our method. We replace ST-CNN (classification model with EEG data) in our cascade network with AST-GCN (classification model with gait data). The experimental results are shown in **Table 6**. Our proposed method with two models significantly outperforms the baseline with one modal (i.e., gait data) while enjoying a faster inference speed (3.99 vs. 7.06 ms) and less parameters (4.72 vs. 9.42M). Since we do not have the EEG data collected from HC regarding the difficulty of collecting them in our experimental environments, we did not compare our method with the EEG-based method, and we leave it for our future work.

5. CONCLUSION

In this paper, we have exploited both the gait and EEG data to achieve a faster and more accurate classification of AD. To this end, we have proposed a cascade neural network. Our proposed neural network consists of two parts. In the first part, we used gait data to distinguish HC from patients. For the purpose of modeling the natural connection among the human joints, we have proposed attention-based spatial temporal graph convolutional networks to extract features to classify the HC and patients. In the second part, we further classify MCI and AD patients with EEG data. Compared with the methods that convert EEG data into the frequency domain, we extract the spatial and temporal features from the original EEG data to distinguish the AD patients from MCI patients. The proposed cascade network has the following advantages: (1) The EEG data from HC are not required in our method, which saves a lot of data collection time. (2) The accuracy of our proposed framework in the three-way classification of HC, MCI, and AD is 91.07%, which is much higher than the method using one modal only (68.18%), and the accuracy in the binary classification of HC vs. MCI/AD reaches 93.09%. It would be interesting to extend this framework to the diagnosis task of other neurological diseases, and we leave it for future work.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Shenzhen People's Hospital Medical Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

FUNDING

This research was funded by the Science and Technology Planning Project of Shenzhen Municipality (JCYJ20170818111012390), Sanming Project of Medicine in Shenzhen (SYJY201905 and SYJY201906), and the Shenzhen Health Committee Project (SZXJ2017034).

REFERENCES

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement*. (2007) 3:186–91. doi: 10.1016/j.jalz.2007.04.381
2. Patterson CA. *World Alzheimer Report 2018.* London: Alzheimer's Disease International (2018).
3. Wang WH, Hsu YL, Pai MC, Wang CH, Wang CY, Lin CW, et al. Alzheimer's disease classification based on gait information. In: *2014 International Joint Conference on Neural Networks (IJCNN)*. Beijing (2014) p. 3251–7. doi: 10.1109/IJCNN.2014.6889762
4. Varatharajan R, Manogaran G, Kumar PM, Sundarasekar R. Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Comput*. (2017) 21:681–90. doi: 10.1007/s10586-017-0977-2
5. Gao H, Liu C, Li Y, Yang X. V2VR: reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability. *IEEE Trans Intell Transport Syst*. (2020). doi: 10.1109/TITS.2020.2983835. [Epub ahead of print].
6. Ieracitano C, Mammone N, Bramanti A, Hussain A, Morabito FC. A Convolutional Neural Network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing*. (2019) 323:96–107. doi: 10.1016/j.neucom.2018.09.071
7. Bi X, Wang H. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural Netw*. (2019) 114:119–35. doi: 10.1016/j.neunet.2019.02.005
8. Bennasar M, Setchi R, Hicks Y, Bayer A. Cascade classification for diagnosing dementia. In: *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. San Diego, CA (2014) p. 2535–40. doi: 10.1109/SMC.2014.6974308
9. Ning Z, Zhang K, Wang X, Guo L, Hu X, Huang J, et al. Intelligent edge computing in internet of vehicles: a joint computation offloading and caching solution. *IEEE Trans Intell Transport Syst*. (2020). doi: 10.1109/TITS.2020.2997832. [Epub ahead of print].
10. Seifallahi M, Soltanizadeh H, Mehraban AH, Khamseh F. Alzheimer's disease detection using skeleton data recorded with Kinect camera. *Cluster Comput*. (2019) 23:1469–81. doi: 10.1007/s10586-019-03014-z
11. Yu Y, Liu S, Guo L, Yeoh PL, Vucetic B, Li Y. CrowdR-FBC: a distributed fog-blockchains for mobile crowdsourcing reputation management. *IEEE Intern Things J*. (2020) 7:8722–35. doi: 10.1109/JIOT.2020.2996229
12. Anderer P, Saletu B, Klöppel B, Semlitsch HV, Werner H. Discrimination between demented patients and normals based on topographic EEG slow wave activity: comparison between z statistics, discriminant analysis and artificial neural network classifiers. *Electroencephalogr Clin Neurophysiol*. (1994) 91:108–17. doi: 10.1016/0013-4694(94)90032-9
13. Pritchard WS, Duke DW, Coburn KL, Moore NC, Tucker KA, Jann MW, et al. EEG-based, neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by non-linear EEG measures. *Electroencephalogr Clin Neurophysiol*. (1994) 91:118–30. doi: 10.1016/0013-4694(94)90033-7
14. Ning Z, Dong P, Wang X, Hu X, Guo L, Hu B, et al. Mobile edge computing enabled 5G health monitoring for internet of medical things: a decentralized game theoretic approach. *IEEE J Select Areas Commun*. (2020). [Epub ahead of print].
15. Trambaiolli LR, Lorena AC, Fraga FJ, Kanda PAM, Anghinah R, Nitrini R. Improving Alzheimer's disease diagnosis with machine learning techniques. *Clin EEG Neurosci*. (2011) 42:160–5. doi: 10.1177/155005941104200304
16. Rossini PM, Buscema M, Capriotti M, Grossi E, Babiloni C. Is it possible to automatically distinguish resting EEG data of normal elderly vs. mild cognitive impairment subjects with high degree of accuracy? *Clin Neurophysiol*. (2008) 119:1534–45. doi: 10.1016/j.clinph.2008.03.026
17. Gao H, Xu Y, Yin Y, Zhang W, Li R, Wang X. Context-aware QoS prediction with neural collaborative filtering for internet-of-things services. *IEEE Intern Things J*. (2020) 7:4532–42. doi: 10.1109/JIOT.2019.2956827
18. Callisaya ML, Launay CP, Srikanth V, Verghese J, Allali G, Beauchet O. Cognitive status, fast walking speed and walking speed reserve-the Gait and Alzheimer Interactions Tracking (GAIT) study. *GeroScience*. (2017) 39:231–9. doi: 10.1007/s11357-017-9973-y
19. Ning Z, Zhang K, Wang X, Obaidat MS, Guo L, Hu X, et al. Joint computing and caching in 5G-envisioned internet of vehicles: a deep reinforcement learning-based traffic control system. *IEEE Trans Intell Transport Syst*. (2020). doi: 10.1109/TITS.2020.2970276. [Epub ahead of print].
20. Beauchet O, Launay CP, Sekhon H, Montembeault M, Allali G. Association of hippocampal volume with gait variability in pre-dementia and dementia stages of Alzheimer disease: results from a cross-sectional study. *Exp Gerontol*. (2019) 115:55–61. doi: 10.1016/j.exger.2018.11.010
21. Elbaz A, Artaud F, Singh-Manoux A, Dumurgier J. Gait speed and decline in gait speed as predictors of incident dementia. *Innov Aging*. (2017) 1:75. doi: 10.1093/geroni/igx004.310
22. Ardle RM, Morris R, Wilson JB, Galna B, Thomas AJ, Rochester LR. What can quantitative gait analysis tell us about dementia and its subtypes? A structured review. *J Alzheimer's Dis*. (2017) 60:1295–312. doi: 10.3233/JAD-170541
23. Morris R, Lord S, Lawson RA, Coleman S, Galna B, Duncan GW, et al. Gait rather than cognition predicts decline in specific cognitive domains in early Parkinson's disease. *J Gerontol Ser A*. (2017) 72:1656–62. doi: 10.1093/gerona/glx071
24. Hsu YL, Chung PC, Wang WH, Pai MC, Wang CY, Lin CW, et al. Gait and balance analysis for patients with Alzheimer's disease using an inertial-sensor-based wearable instrument. *IEEE J Biomed Health Informatics*. (2014) 18:1822–30. doi: 10.1109/JBHI.2014.2325413

25. Gao H, Kuang L, Yin Y, Guo B, Dou K. Mining consuming behaviors with temporal evolution for personalized recommendation in mobile marketing apps. *Mobile Netw Appl*. (2020) 25:1233–48. doi: 10.1007/s11036-020-01535-1
26. Wang X, Ning Z, Guo S. Multi-agent imitation learning for pervasive edge computing: a decentralized computation offloading algorithm. *IEEE Trans Parallel Distrib Syst*. (2020) 32:411–25. doi: 10.1109/TPDS.2020.3023936
27. Yu Y, Liu S, Yeoh P, Vucetic B, Li Y. LayerChain: a hierarchical edge-cloud blockchain for large-scale low-delay IIoT applications. *IEEE Trans Indus Informatics*. (2020) doi: 10.1109/TII.2020.3016025. [Epub ahead of print].
28. Yan JH, Rountree SD, Massman PJ, Doody R, Li H. Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. *J Psychiatr Res*. (2008) 42:1203–12. doi: 10.1016/j.jpsychires.2008.01.006
29. Choi JS, Oh HS, Kang DW, Mun KR, Choi MH, Lee SJ, et al. Comparison of gait and cognitive function among the elderly with Alzheimer's disease, mild cognitive impairment and healthy. *Int J Precis Eng Manufact*. (2011) 12:169–73. doi: 10.1007/s12541-011-0024-9
30. Bashivan P, Rish I, Yeasin M, Codella N. Learning representations from EEG with deep recurrent-convolutional neural networks. *CoRR*. (2016) abs/1511.06448.
31. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *AAAI*. New Orleans, LA (2018).
32. Wang X, Ning Z, Guo S, Wang L. Imitation learning enabled task scheduling for online vehicular edge computing. *IEEE Trans Mobile Comput*. (2020). doi: 10.1109/TMC.2020.3012509. [Epub ahead of print].
33. Yang Z, Li Y, Yang J, Luo J. Action recognition with spatio-temporal visual attention on skeleton image sequences. *IEEE Trans Circuits Syst Video Technol*. (2019) 29:2405–15. doi: 10.1109/TCSVT.2018.2864148
34. Lawhern V, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance B. EEGNet: a compact convolutional network for EEG-based brain-computer interfaces. *J Neural Eng*. (2018) 15:056013. doi: 10.1088/1741-2552/aace8c
35. Chattopadhyay P, Sural S, Mukherjee J. Frontal gait recognition from occluded scenes. *Pattern Recogn Lett*. (2015) 63:9–15. doi: 10.1016/j.patrec.2015.06.004
36. Fang J, Wang T, Li C, Hu X, Ngai ECH, Seet BC, et al. Depression prevalence in postgraduate students and its association with gait abnormality. *IEEE Access*. (2019) 7:174425–37. doi: 10.1109/ACCESS.2019.2957179
37. Beyrami SMG, Ghaderyan P. A robust, cost-effective and non-invasive computer-aided method for diagnosis three types of neurodegenerative diseases with gait signal analysis. *Measurement*. (2020) 156:107579. doi: 10.1016/j.measurement.2020.107579
38. Toll RT, Wu W, Naparstek S, Zhang Y, Narayan M, Patenaude B, et al. An electroencephalography connectomic profile of posttraumatic stress disorder. *Am J Psychiatry*. (2020) 177:233–43. doi: 10.1176/appi.ajp.2019.18080911
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV (2016) p. 770–8. doi: 10.1109/CVPR.2016.90
40. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *CoRR*. (2015) abs/1409.1556.
41. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. (2014) 15:1929–58.

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 Public Health

ORIGINAL RESEARCH
published: 23 November 2020
doi: 10.3389/fpubh.2020.584430

Application of Structural and Functional Connectome Mismatch for Classification and Individualized Therapy in Alzheimer Disease

Huixia Ren [1,2], Jin Zhu [3], Xiaolin Su [4], Siyan Chen [4], Silin Zeng [4], Xiaoyong Lan [4], Liang-Yu Zou [4], Michael E. Sughrue [5] and Yi Guo [4]**

[1] Department of Neurology, The Second Clinical Medical College, Shenzhen People's Hospital, Jinan University, Shenzhen, China, [2] The First Affiliated Hospital, Jinan University, Guangzhou, China, [3] Department of Medical Imaging, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China, [4] Department of Neurology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China, [5] Centre for Minimally Invasive Neurosurgery, Prince of Wales Hospital, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Wenbing Zhao,
Cleveland State University,
United States

Reviewed by:
Ying Han,
Capital Medical University, China
Junhong Zhou,
Harvard Medical School,
United States
Wenda Wang,
Guangzhou Medical University, China

***Correspondence:**
Michael E. Sughrue
Sughruevs@gmail.com
Yi Guo
xuanyi_guo@163.com

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: *17 July 2020*
Accepted: *09 October 2020*
Published: *23 November 2020*

Citation:
Ren H, Zhu J, Su X, Chen S, Zeng S, Lan X, Zou L-Y, Sughrue ME and Guo Y (2020) Application of Structural and Functional Connectome Mismatch for Classification and Individualized Therapy in Alzheimer Disease.
Front. Public Health 8:584430.
doi: 10.3389/fpubh.2020.584430

While machine learning approaches to analyzing Alzheimer disease connectome neuroimaging data have been studied, many have limited ability to provide insight in individual patterns of disease and lack the ability to provide actionable information about where in the brain a specific patient's disease is located. We studied a cohort of patients with Alzheimer disease who underwent resting state functional magnetic resonance imaging and diffusion tractography imaging. These images were processed, and a structural and functional connectivity matrix was generated using the HCP cortical and subcortical atlas. By generating a machine learning model, individual-level structural and functional anomalies detection and characterization were explored in this study. Our study found that structural disease burden in Alzheimer's patients is mainly focused in the subcortical structures and the Default mode network (DMN). Interestingly, functional anomalies were less consistent between individuals and less common in general in these patients. More intriguing was that some structural anomalies were noted in all patients in the study, namely a reduction in fibers involving parcellations in the right anterior cingulate. Alternately, the functional consequences of connectivity loss were cortical and variable. Integrated structural/functional connectomics might provide a useful tool for assessing AD progression, while few concerns have been made for analyzing the mismatch between these two. We performed a preliminary exploration into a set of Alzheimer disease data, intending to improve a personalized approach to understanding individual connectomes in an actionable manner. Specifically, we found that there were consistent patterns of white matter fiber loss, mainly focused around the DMN and deep subcortical structures, which were present in nearly all patients with clinical AD. Functional magnetic resonance imaging shows abnormal functional connectivity different within the patients, which may be used as the individual target for further therapeutic strategies making, like non-invasive stimulation technology.

Keywords: brain connectivity, diffusion tractography imaging, Alzheimer's disease, brain parcellation, functional MRI, machine learning

INTRODUCTION

Alzheimer disease (AD) is characterized as the most common cause of dementia with non-stop developing progression and effective strategies, even to date. It is well-known that conventional magnetic resonance imaging (MRI) imaging provides very limited insight into dementia patients (1). While patterns of atrophy can provide some indirect diagnostic evidence for one type of degenerative disease vs. another, this is relatively limited and often can be non-specific. Furthermore, individuals can have substantial age-related atrophy and not exhibit clinical signs of dementia, again suggesting that structural brain MRI has only limited ability to diagnose, stage, or guide treatment in any meaningful way for these patients (1). Growing evidence supports the idea that AD is associated with disruptions in brain activity and networks that may target specific functionally connected neuronal networks (2, 3). These facts drive interest in more sophisticated neuroimaging, such as positron emission tomography–based studies, which are able to image the amyloid and tau proteins (4), and connectomic-based approaches, leveraging imaging studies such as functional magnetic resonance imaging (fMRI) and diffusion tractography imaging (DTI) (5). A growing number of researchers work on the development of personalized, reproducible, non-invasive, and neuroscientifically interpretable biomarkers for early diagnosis or prediction of AD even on the subjective cognition decline (SCD) stage (6–8), yet most of which is focused on the consistent abnormal connection within the multimodal imaging as the combination with DTI and fMRI (9, 10). Given the subtle and often diffuse nature of dementing disorders, machine learning–based approaches provide the most realistic method for complex imaging datasets (11, 12).

Machine learning is an application of artificial intelligence that allows computers to learn automatically and improve from experience. It is one of today's most rapidly growing technical fields (13), which performs throughout science including health care (14) such as identification and classification for diseases like AD (15–17), traffic programming (18), and marketing apps designing (19), which allows us to process large-scale, multidimensional, complex datasets in this information explosion of an era. Machine learning–based analysis of connectomic data created from neuroimaging studies in patients AD has been extensively studied in the literature (5, 9, 12, 20, 21). Most such efforts utilize a method for modeling features of either DTI and/or fMRI studies, which allow a model to differentiate between some combination of healthy controls, patients with mild cognitive impairment, and those with AD. While early identification of patients who will progress to clinical AD would provide a clinically critical patient cohort who are the best candidates for disease-modifying therapies (8), models that provide a yes vs. no answer ignore the possibility of heterogeneity of phenotypes, have limited ability to provide insight into stages of the disease, and lack the ability to provide actionable information about where in the brain a specific patient's disease is located and what specifically is happening. Treatments such as repetitive transcranial magnetic stimulation (rTMS) provide a safe and potentially useful tool that may palliate symptoms in patients even if not disease-modifying, but for which it is unclear what the appropriate target is (22).

TABLE 1 | Demographic and clinical characteristics of participants.

	Healthy control (n = 41)	AD (n = 21)	P
Age (years)	70.25 (0.77)	67.43 (2.35)	0.14
Gender (% female)	22 (50%)	17 (76%)	0.001**
Education (years)	16.56 (0.40)	10.71 (1.02)	<0.0001****
Handedness (% right handed)	40 (100)	21 (100)	0.99
MMSE	29.00 (0.18)	24.29 (1.05)	0.002**

means a significant difference with P = 0.001; * means a significant difference with p < 0.0001.*

In this pilot study, we presented a different approach using machine learning to study AD which focused on characterizing the site of a structural and functional anomaly at the single-subject level. Not only did this approach provide potentially actionable information, for therapies such as rTMS, but our data suggested that specific anomalies were remarkably consistent between individuals regardless of disease staging, which suggested that they might represent fundamental steps in early symptomatology of AD, and others became increasing less consistent which indicated the possibility of heterogeneous subgroups or stages of the disease.

MATERIALS AND METHODS

Participants

The study included 21 patients with clinically diagnosed AD between the ages of 50 and 90 years who presented to Shenzhen's People's Hospital for evaluation and 41 healthy controls with similar age and intact cognition. All research testing was performed with the approval of the local institutional review board (Shenzhen People's Hospital Medical Ethics Committee) and with informed consent from the patient and/or designated surrogate. The research has registered in the Chinese Clinical Trial Registry (ChiCTR1800019199). The demographic characteristics of the participants are listed in **Table 1**.

Clinical and Neurocognitive Assessments

We administered the same standardized neurocognitive test to participants in both the AD and HC groups. All patients underwent standard neurologic testing in addition to the Mini-Mental Status Examination (MMSE) (23) and the Montreal Cognitive Assessment (24) to confirm the diagnosis. MMSE was used for the comparison between the AD and HC groups, based on the correction of educational level; patients were classified as cognitive decline where ≤18 MMSE. In the AD group, 17 of 21 patients were female, which had a significant difference with HC ($P = 0.001$), despite we included equal proportions of gender in HC, in clinical setting; two-thirds of persons diagnosed with AD are women. There was also a notable difference in education between two groups ($P < 0.0001$), which was consistent with the research that older adults with at least 16 years of education had less of the progressive neurodegeneration associated with AD. The MMSE in the AD group was decreased significantly compared with HC ($P = 0.002$). The participants had suffered approximately 3.2 years from AD or a noticeable cognition decline with a variation from 2 up to 10 years.

Inclusion and Exclusion Criteria for AD

For inclusion criteria, (1) a diagnosis of probable AD according to the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) (NINCDS-ADRDA) (25), (2) at age 50 to 90 years, (3) with ≤18 MMSE score, and (4) current symptomatic treatment of AD.

And for the exclusion criteria, any other causes for cognitive decline (1) prior or current neurological or central nervous system disorders, (2) psychiatric disorder such as schizophrenia, major depression, or any other psychiatric condition, (3) abnormalities on MRI like lacunar infarcts, cerebral lesions, etc., and (4) the presence of associated disorders, immune, metabolic, or endocrine disorders and a history of cancer, etc., (5) use of prohibited medication or alcohol abuse, and (6) a diagnosis of AD and concomitant cerebrovascular disease.

MRI Data Acquisition

For the HC group, we obtained 36 normal subject images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) from the ADNI2 study collected on the Philips Achieva and GE Discovery MR 750 3.0-T MRI scanner. DTI was acquired on with 5 $b = 0$ baseline image and a $b = 1{,}000$ shell with 41-direction acquisition, field of view (FOV) $= 350 * 350$ mm, slice thickness 2.7 mm, 0-mm gap between slices with no overlap, full brain coverage, isotropic voxels, square $256 * 256$ matrix.

Resting-state fMRI (rsfMRI) images were acquired on a 3.0-T MRI scanner, $3.312 \times 3.312 \times 3.312$-mm voxels, 140 volumes/run, TR $= 2{,}020$ ms, TE $= 30$ ms, field of view $= 224 \times 224$ mm, flip angle $= 80^\circ$, 7-min run time.

For AD patients, Siemens Skyra 3.0-T MRI scanner was used for data acquisition; all patients underwent a pretreatment standard structural T1- and T2-weighted images, as well as diffusion-weighted image, and MR angiography to rule out secondary explanations for their clinical dementia.

DTI with the following parameters: with 10 $b = 0$ baseline image and a $b = 1{,}000$ shell with 64 direction acquisition, FOV $= 224 * 224$ mm, slice thickness 2 mm, 0-mm gap between slices with no overlap, full brain coverage, isotropic voxels, square $112 * 112$ matrix.

rsfMRI was performed with the following parameters: T2-star EPI sequence, $3.5 \times 3.5 \times 3.5$-mm voxels, 240 volumes/run, TR $= 2{,}020$ ms, TE $= 30$ ms, field of view $= 224 \times 224$ mm, flip angle $= 90^\circ$, 8-min run time.

To eliminate the difference made by MRI scanners in this study, a preprocessing step using tangent space normalization and whitening method was applied to correct the influence of the bias field to reduce misdiagnosis and improve the accuracy of diagnosis before segmentation or classification.

rsfMRI Preprocessing

The rsfMRI images were processed using standard processing steps: (1) motion correction was performed on the T1 and BOLD images using a rigid body alignment; (2) slices with excess movement [defined as DVARS> 2 sigma (26) from the mean slice] were eliminated; (3) the T1 image was skull stripped using a convolutional neural net (CNN); this was inverted and aligned to the resting state bold image using a rigid alignment, which was then used as a mask to skull strip the rsfMRI image, (4) slice time correction and global intensity normalization was performed, (5) gradient distortion correction were performed using a diffeomorphic warping method which aimed to locally similarize the rsfMRI and T1 images, (6) High variance confounds were calculated using the CompCor method (27) as well as motion confounds were regressed out of the rsfMRI image, and the linear and quadratic signals were detrended, (7) spatial smoothing was performed using a 4-mm full width at half maximum Gaussian kernel. The personalized atlas created in previous steps was registered to the T1 image and localized to the gray matter regions. Thus, it was ideally positioned for extracting an average BOLD time series from all 379 areas (180 parcellations $\times$ 2 hemispheres, additionally with 19 subcortical structures), which yielded 143,641 correlations.

Diffusion Tractography Preprocessing

The diffusion tractography (DT) images were processed using the Omniscient software, which employs a standard processing steps in the Python language (28): (1) the diffusion image was resliced to ensure isotropic voxels; (2) motion correction was performed using a rigid body alignment; (3) slices with excess movement (defined as DVARS >2 sigma from the mean slice) were eliminated; (4) the T1 image was skull stripped using a convolutional neural net (CNN); this was inverted and aligned to the DT image using a rigid alignment and then used as a mask to skull strip the DT; (5) gradient distortion correction was performed using a diffeomorphic warping method which aimed to locally similarize the DT and T1 images; (6) eddy current correction was performed; (7) fiber response function was estimated and the diffusion tensors were calculated using constrained spherical deconvolution; and (8) deterministic tractography was performed with random seeding, usually creating about 300,000 streamlines per brain.

Machine Learning–Based Parcellation

Not only the ML has been largely used in the prediction for internet-of-Things services (29) and traffic control system (30), which also been applied to the neurological science. To create a personalized brain atlas, the structural adjacency matrix was extracted as a set of fibers running between each pair of parcellations. To minimize the effects of brain atrophy, we created a machine learning–based, subject-specific version of the HCP-MMP1 (31) atlas based on DTI structural connectivity. This was created by training a machine learning model on 200 normal adult subjects by first processing T1 and DT images as above. A HCP-MMP1 atlas in NIFTI MNI space was then warped onto each brain and the structural connectivity was calculated between every pair of this atlas and a set of ROI's containing 8 subcortical structures per hemisphere as well as the brainstem based on the streamlines, which terminated within an ROI. This step both allowed the generation of feature vectors that basically a 379 × 379 structural connectivity based adjacency matrix, and generated a centroid of the parcellation, which was utilized to constrain the voxels studied for assignment to a given parcellation to a plausible area near its typical position. These feature vectors for each region were then used as a training set and the data were modeled using the eXtreme Gradient Boosting (XGBoost) method.

This model was then applied to the new subject by first warping the HCP-MMP1 atlas to the new brain and collecting a set of feature vectors of the connectivity of each voxel (32–35). The feature vectors were then used to determine if each voxel belongs to a parcellation or region. This created a version of the HCP-MMP1 atlas with subcortical components, which was not dependent on brain shape or pathologic distortion but specific for this subject while comparable between subjects.

Personalized Anomaly Detection

Instead of trying to fit a machine learning model to the raw data, we studied these patients on an individual level by utilizing machine learning to direct us to areas that were abnormal in AD patients compared to age-similar controls. To do this, we utilized the ADNI2 dataset to generate a training set, which was processed using the same technique. We then performed a tangent space connectivity transformation, whitening, and normalization (36) to determine the range of normal correlations for each functional connectivity and structural connectivity pair in the matrix. We then excluded the one-third of pairs in both structural and functional with the highest between subject variance in the normal cohort (37), under the hypothesis that these areas might be prone to false discovery, possible due to inter-individual variability in normal subjects. Abnormal connectivity for each connection was determined as a 3-sigma outlier for that structural or functional entry. Assignment of parcellations to various large-scale brain networks was based on several previous coordinates based meta-analyses, which have been previously published research (38–41).

The illustration of the data processing and model forming is shown in **Figure 1**.

Statistical Analyses

All statistical analyses were conducted in SPSS software (IBM Corporation), for the comparison of demographic and clinical characteristics of participants, independent sample T-test analyses using two-sided tests in continuous data and a Chi-square was assessed for the discreet data.

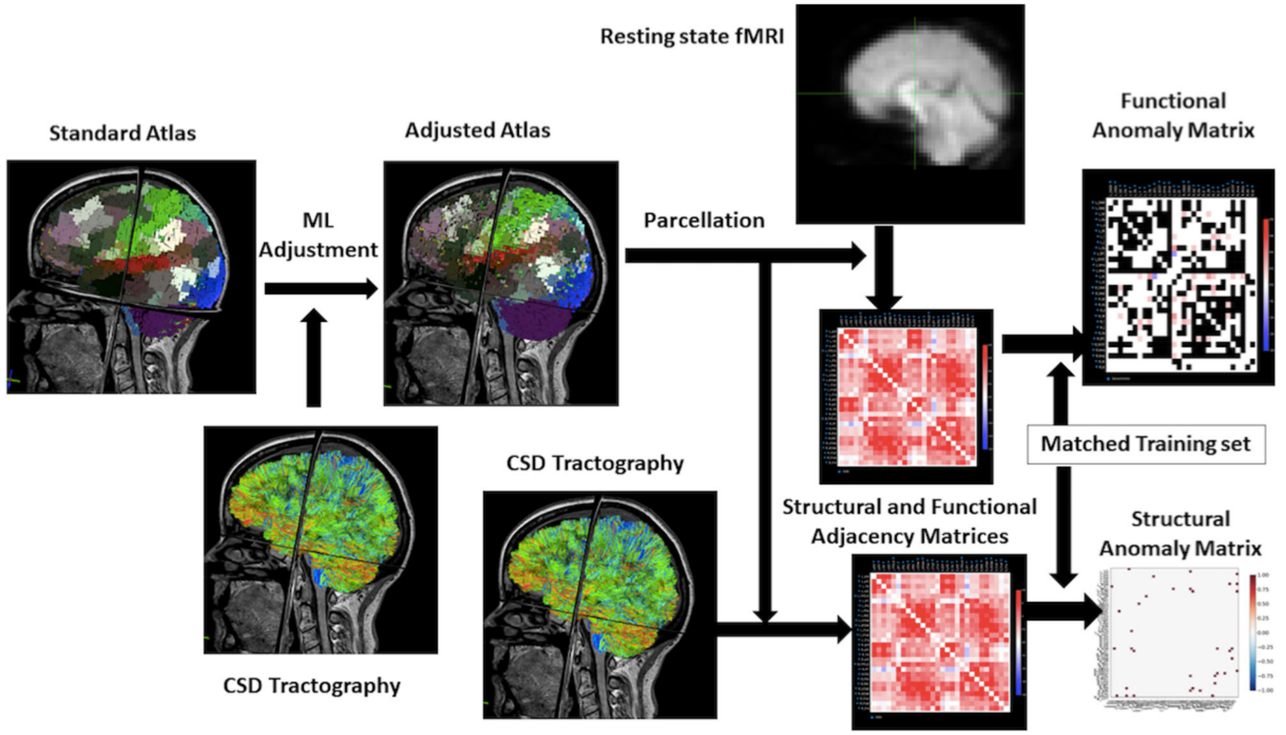


FIGURE 1 | Workflow for the research. From the upper left to the right of this flowchart: the research starts with a standard atlas warped onto the brain, the boundaries are smooth because it is not machine learning–based. Then using the constrained spherical deconvolution–based tractography to adjust the atlas to personalize it. Process the rsfMRI to a functional matrix and structural MRI to a structural matrix by taking parcellation of atlas. The final step will be utilizing a training set in machine learning to make an anomaly matrix of structural and functional connectivity for further analysis.

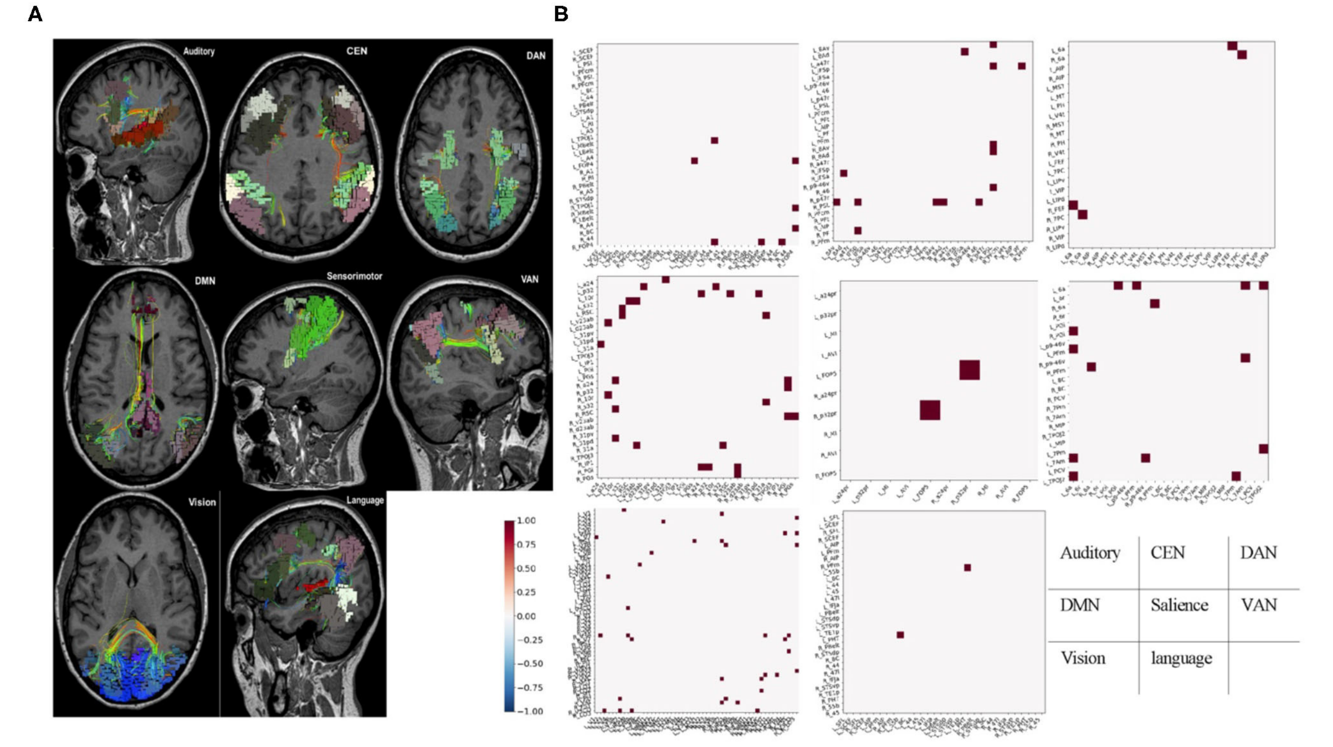


FIGURE 2 | Fiber tracts and fMRI-based brain network. **(A)** Parcellations and fiber tracts–based brain network pulled out from the machine learning algorithms. Three-dimensional rendering of parcellations and tractography-based MRI images for identified set of seven canonical brain connectivity networks that Only shows tracts within areas of the network. **(B)** Example submatrices of structural anomalies for the same patient based on affiliation in the same brain-network with **(A)**. Normal or high variances (excluded areas) were indicated in white. Dots represent areas with less diffusion tractography fibers traces between them and normal, age-similar subjects. These maps provided a network-by-network fingerprint. CEN, central executive network; DAN, dorsal attention network; DMN, default mode network; VAN, ventral attention network.

RESULTS

Anomaly Detection–Based Fingerprinting of AD-Based Anomalies

Parcellations and fiber tracts–based brain network pulled out from the machine learning algorithms and an example of this matrix subset based on the affiliation of a parcellation with one of the known large-scale brain networks. This example showed the form of data these algorithms provide about specific brain networks (**Figure 2**). Note when we visually inspected all 21 brains, we did not note any consistent patterns between patients except that the default mode network was always abnormal in some way. It was important to note that white entries include both connections that were within normal limits compared to age-similar controls, and those connections are highly variable in the control group, suggesting that they were too interindividual variable to be meaningfully called an anomaly.

Structural Disease Burden in AD Is Mainly in the Subcortical Structures and in DMN

To understand the behavior of data produced by our approach, we first analyzed the overall frequency of anomalies in all areas we studied to get an estimate of which areas were most frequently part of pair with a decreased number of white matter fibers on the diffusion tractography study of these patients compared to the age-similar controls. Note that two aspects of the methodology were worth reiterating. First, we parcellated the brains of both groups using a machine learning model that assigns voxels to a parcellation of subcortical structure based on which other voxels they connect to on the DTI. This means that the basic patterns of connections are held relatively consistent, and should not greatly vary due to alignment of the atlas or other similar problems. Second, while white matter connections decrease with age dependent ways, which do not necessarily cause dementia, the comparison with age-similar controls implies that this comparison should select out AD-specific connection loss.

Table 2 demonstrates the areas with the highest fraction of their possible anomalies in all 21 patients who had an anomaly. We noted that that the top 23 areas had decreased numbers of fibers between the area and 7.6 and 13.85% of all possible target areas in all 21 patients studied (at least among the low variance options). **Figure 3** shows this structural anomaly burden as a series of bar graphs. This demonstrates two natural inflection points where the burden drops, suggesting somewhat significant changes in behavior. As **Table 1**, shows, the majority of the high anomaly burden areas are subcortical and include basal ganglia

TABLE 2 | Structural anomaly burden.

Parcellation	No. of anomalies	No. of subjects with at least one anomaly	No. of low variance connections	Total potential anomalies	Percentage of total %
R_8BL	634	21	218	4,578	13.85
L_pallidum	592	21	204	4,284	13.82
R_pallidum	694	21	249	5,229	13.27
R_ventralDC	294	21	112	2,352	12.50
R_9m	543	21	211	4,431	12.25
R_caudate	362	21	148	3,108	11.65
R_10v	714	21	302	6,342	11.26
L_ventralDC	203	21	87	1,827	11.11
Brain stem	36	21	16	336	10.71
L_putamen	225	21	104	2,184	10.30
L_thalamus	240	21	114	2,394	10.03
L_8BM	288	21	143	3,003	9.59
R_thalamus	207	21	103	2,163	9.57
R_8BM	338	21	175	3,675	9.20
L_10v	416	21	230	4,830	8.61
R_p24	560	21	333	6,993	8.01
R_OFC	462	21	276	5,796	7.97
R_cerebellum	108	21	65	1,365	7.91
R_10pp	301	21	184	3,864	7.79
R_a24	498	21	307	6,447	7.72
L_caudate	229	21	142	2,982	7.68
L_TGd	170	21	106	2,226	7.64
R_accumbens	417	21	261	5,481	7.61

structures, the dorsal diencephalon, and areas 8BL, and 8BM. Also notable are several parts of the anterior portion of the default mode network. Note that patients had at least one structural anomaly in every parcellation and subcortical area compared to healthy age-similar controls; these areas have the most frequent anomalies. Of note, neither hippocampus was among the most frequent sites of structural anomalies.

Structural–Functional Mismatch Characterizes the Anomalies in AD

Table 2 shows a similar analysis of Functional anomalies in AD. Note that the highest-burden areas are generally not subcortical regions. The default mode areas, such as p24 and 10v are on both lists as are frontal areas 8BM and 8BL. Also note that with the exception of the right hippocampus, all of the highest functional anomaly burden areas are cortical. In other words, even though the deep structures frequently show decreased numbers of white matter fibers on with different brain regions, the less commonly show observable functional connectivity disturbances with those areas.

Disease Defining Anomalies in AD Were Structural Changes in the Right Anterior Cingulate

To see how consistent the anomalies seen in AD occurred, and specifically if there were any connection, which was usually abnormal. **Table 4** demonstrates the results of this frequency analysis on the structural connectomes of these patients. Interestingly, two anomalies were seen in all 21 patients, and 3 anomalies were seen in 20/21 patients. These involved the anterior and middle cingulate gyrus on the right as one or both pairs of abnormal structural connections. As we looked through the connections of decreasing frequency, the most consistent connections were overrepresented by right-sided and DMN anomalies, consistent with many other studies.

The Functional Consequences of Connectivity Loss Were Cortical and Variable

Table 5 demonstrates a similar analysis of the most common functional anomalies in AD patients. Two obvious differences were notable. First, functional anomalies were far less consistent with the most common anomaly in functional connectivity only occurring in 8 patients. Second, these anomalies are corticocortical or corticohippocampal, and none appear to be corticobasal or corticothalamic. Interestingly, the abnormal functional connectivity, which was common between subjects spread into numerous networks, as opposed to mainly the DMN, and it was mostly areas that were interhemispheric or not immediately adjacent to each other. The Dorsolateral prefrontal cortex (DLPFC) and dorsomedial prefrontal cortex (DMPFC) were particularly affected, with 8BM and 8BL notable inclusions.

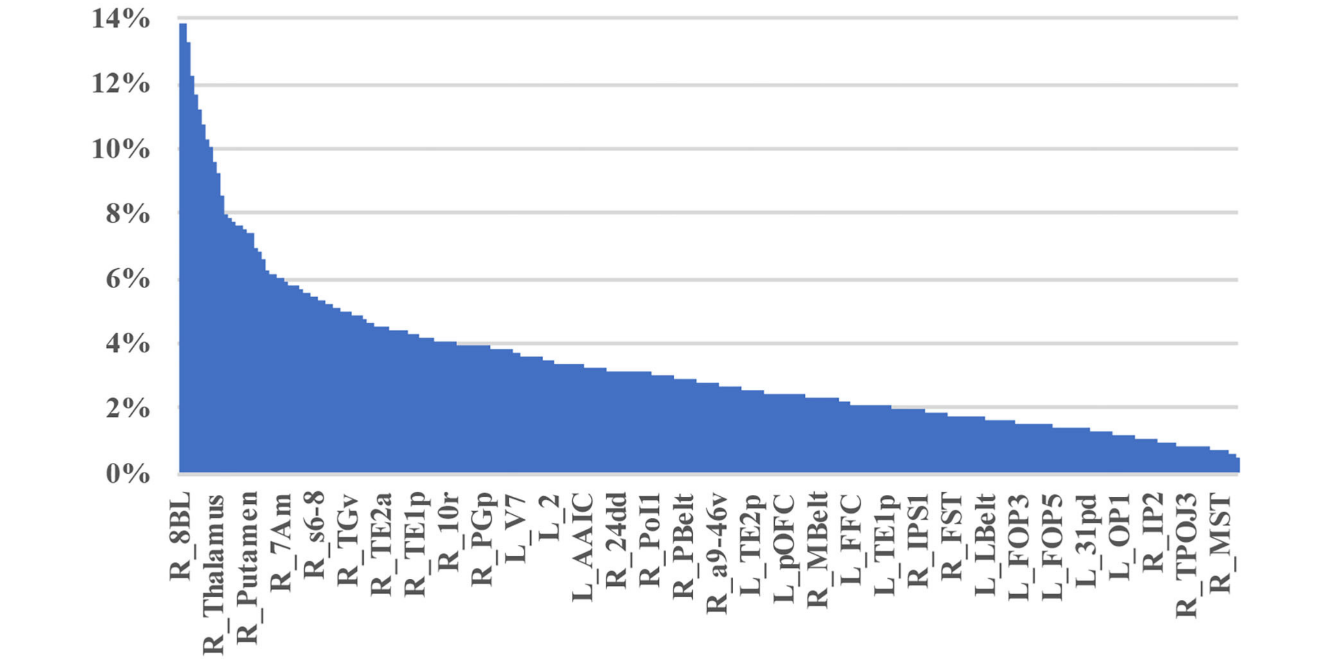


FIGURE 3 | A visual depiction of structural anomaly burden in these 21 subjects. This is a set of 377 bar graphs representing the total fractions of anomalies noted in each of the cortical parcellations and subcortical regions of interest expressed as a total % of possible anomalies. This gives a sense of which connections are most consistently abnormal compared to normal age-similar but healthy controls in non-variable areas. Note there are two inflection points in this graph that demonstrate steep transitions in the data. Areas to the left of the first inflection point are mostly subcortical structures, including the putamen, caudate, and thalamus, among others, and areas 10v, right 9M, bilateral 8BM, and right area 8BL. Areas between the two inflection points mainly include regions within the anterior cluster of the Default mode network. Most other areas have a lower anomaly burden and are to the right of the second inflection point.

DISCUSSION

The development of personalized, reproducible, non-invasive, and neuroscientifically interpretable biomarkers are urgently needed for AD precision medicine (16, 42), yet despite remarkable advances, few such biomarkers are available. Neuroimaging using DTI and fMRI in conjunction provides objective information on the structure and function that for assessing network connectivity of the brain. In this study, we performed a preliminary exploration into a set of AD data with a goal of revising a heuristic for analyzing these patients with the goal of improving a personalized approach to understanding individual connectomes in an actionable manner. Specifically, we found that there were consistent patterns of white matter fiber loss, mainly focused around the DMN and deep subcortical structures, which were present in nearly all patients with clinical AD (**Tables 2**, **4**). Additionally, these structural anomalies were frequent, but not universal. We also found an obvious mismatch between the structural and functional anomalies in these patients, with the latter being most cortical, and mostly areas separated at long distances from each other.

The fact that DTI found white matter fiber anomalies, which were consistent between individuals, even being present in all patients, was a surprising finding, but aligns with other machine learning approaches (5) aimed at making the diagnosis of AD vs. normal, suggesting that these changes are early and disease defining. In other words, it is difficult to have clinical AD with a DMN with normal structural connectivity.

As important as this is, it implies that these problems are not useful for personalizing treatment approaches, or for staging. To that effect, the parcellations in the less common, but not rare groups e.g., being present in 50–65% of patients, seem like better candidates, as these might track the course of the disease better. Previews studies showed that the combination fMRI or/with DTI can be used for identification of the early stage of AD (9, 43) and classification from various manifestations dementia (15), while revealed only the abnormalities in large-scale network connectivity in several brain regions such as right hippocampal, left middle frontal gyrus, posterior cingulate, and middle cingulate gyrus on the right, which is consistent with the structural abnormal assessed with DTI in our study. The mismatch between structural and functional anomalies in our research was striking (**Tables 2–5**). It is interesting to speculate why this would be the case, but given the physical distance between areas common on this list, we suggest that loss of corticobasal and corticothalamic fibers, common in these patients, reduce the ability of these structures to facilitate communication with distant areas. It highlights the need to look at areas beyond the large-scale brain networks when we try to understand functional-phenotypic relationships.

It was well-known that DMN was considered as the most affected network in neurological and neuropsychiatric disorders,

TABLE 3 | Functional anomaly burden.

Parcellation	No. of anomalies	No. of subjects with at least one anomaly	No. of low variance connections	Total potential anomalies	Percentage of total %
L_8BM	577	19	377	7,163	8.06
R_PFt	533	19	360	6,840	7.79
R_V1	540	19	374	7,106	7.60
L_9-46d	535	19	379	7,201	7.43
L_10v	500	19	378	7,182	6.96
R_hippocampus	453	19	370	7,030	6.44
L_AAIC	437	19	378	7,182	6.08
R_8BL	389	19	378	7,182	5.42
R_13l	384	19	374	7,106	5.40
L_IFJa	318	19	360	6,840	4.65
R_VMV3	327	19	373	7,087	4.61
L_PIT	306	19	360	6,840	4.47
R_MIP	314	19	371	7,049	4.45
R_PHT	290	19	345	6,555	4.42
L_IFJp	316	19	376	7,144	4.42
L_9p	310	19	371	7,049	4.40
R_PIT	303	19	367	6,973	4.35
L_s32	289	19	351	6,669	4.33
R_p24	304	19	374	7,106	4.28
L_PHA1	289	19	357	6,783	4.26
L_V4t	290	19	362	6,878	4.22
R_PoI2	264	19	334	6,346	4.16
R_2	282	19	359	6,821	4.13

TABLE 4 | Frequency of structural anomalies.

Patients	Affiliation 1	Parcellation 1	Parcellation 2	Affiliation 2	Hemisphere	Relationship
21	Salience	R_a24pr	L_STSdp	Language	Bilateral	Intrahemispheric
	DMN	R_p24	R_24dd	Sensorimotor	Right	Intralobar
20	DMN	R_p24	R_p24pr	Salience	Right	Intralobar
	DMN	R_p24	R_33pr	DMN	Right	Intralobar
	DMN	R_33pr	R_24dd	Sensorimotor	Right	Intralobar
19	Basal ganglia	R_caudate	R_OFC	Orbitofrontal	Right	Corticobasal
	Basal ganglia	R_caudate	R_10v	DMN	Right	Corticobasal
	Orbitofrontal	R_OFC	R_putamen	Basal ganglia	Right	Corticobasal
17	Salience	R_a24pr	R_a24	DMN	Right	Intralobar
	DMN	R_7m	R_23d	DMN	Right	Intralobar
	Basal ganglia	R_pallidum	R_6a	Dorsal Premotor	Right	Corticobasal
	SPL	R_7Pm	R_23d	DMN	Right	Intralobar
16	Salience	R_p24pr	R_a24	DMN	Right	Intralobar
	Salience	R_p24pr	R_d32	DMN	Right	Intralobar
	DMN	R_23d	R_a24pr	Salience	Right	Intralobar
	Basal ganglia	R_pallidum	R_7PL	SPL	Right	Corticobasal
	DMN	R_10v	L_11l	Orbitofrontal	Bilateral	Intrahemispheric
	Basal ganglia	L_pallidum	R_8BL	DLPFC	Bilateral	Intrahemispheric
	Insula	L_52	L_PoI2	Insula	Left	Intralobar

TABLE 5 | Frequency of functional anomalies.

Patients	Affiliation 1	Parcellation 1	Parcellation 2	Affilliation 2	Hemisphere	Relationship
8	Sensorimotor	R_2	L_IFJa	DLPFC	Bilateral	Interhemispheric
	DMN	L_10v	L_ProS	Visual	Left	Long range
	Insula	L_Pir	L_AAIC	Insula	Left	Intralobal
7	DMN	L_10v	R_PFt	Parietal	Bilateral	Interhemispheric
	DMN	L_10v	R_9-46d	DLPFC	Bilateral	Interhemispheric
	DMN	L_10v	L_AAIC	Insula	Left	Long range
	Lateral parietal	R_PFt	R_8BL	DLPFC	Right	Long range
	Lateral parietal	R_PFt	L_s32	DMPFC	Bilateral	Interhemispheric
	DLPFC	L_IFJa	R_SFL	Sensorimotor	Bilateral	Interhemispheric
	DLPFC	L_IFJa	R_s32	DMPFC	Bilateral	Interhemispheric
	Limbic	R_hippocampus	L_3b	Sensorimotor	Bilateral	Interhemispheric
	Limbic	R_hippocampus	R_13l	Orbitofrontal	Right	Long range
	DMN	L_d32	L_A1	Auditory	Left	Long range
	DMN	L_d32	L_OFC	Orbitofrontal	Left	
	Visual	L_ProS	L_8BM	DMPFC	Left	Long range
	Visual	R_V7	R_VMV1	Visual	Right	
	DLPFC	R_IFJa	L_OP2-3	Lateral parietal	Bilateral	Interhemispheric
	Orbitofrontal	L_pOFC	L_9p	DLPFC	Left	Long range
	DLPFC	L_9-46d	L_V4t	Visual	Left	Long range
6	DMPFC	L_8BM	R_hippocampus		Bilateral	Interhemispheric
	DMPFC	L_8BM	R_2	Sensorimotor	Bilateral	Interhemispheric
	DMPFC	L_8BM	R_PFcm	Lateral parietal	Bilateral	Interhemispheric
	DMPFC	L_8BM	R_V7	Visual	Bilateral	Interhemispheric
	DMPFC	L_8BM	R_V1	Visual	Bilateral	Interhemispheric
	DMPFC	L_8BM	L_s32	DMPFC	Left	
	DMPFC	L_8BM	R_10v	DMN	Bilateral	Interhemispheric
	DMPFC	L_8BM	L_9-46d	DLPFC	Left	
	Lateral parietal	R_PFt	R_V3A	Visual	Right	Long range
	Lateral parietal	R_PFt	R_V7	Visual	Right	Long range
	Lateral parietal	R_PFt	L_ProS	Visual	Bilateral	Interhemispheric
	Lateral parietal	R_PFt	L_31pd	DMN	Bilateral	Interhemispheric

including AD, which shows a high level of activity during rest while deactivates its performance during cognitive tasks (44). These areas include the precuneus/posterior cingulate cortex, medial prefrontal cortex (MPFC), and medial, lateral, and inferior parietal cortex, and its activity holds potential as a non-invasive biomarker of incipient AD (45). Researchers have demonstrated the disconnection or decreased functional connectivity within/between DMN and other networks, which contribute to a cognition decline (46).

Regardless of the mechanism, functional data seems less consistent than structural data most in the DMN. There are good and bad points to using these data. This suggests that using machine learning–based on the variability of functional connectivity to classify or identify patients in early-stage disease, or to stage the extent of the disease, seems less promising than structural data as the anomalies seem to be more individual specific. However, the inherent variability of functional anomaly data in our patients suggests that it is highly promising at personalizing approaches to therapy, such as TMS (22). In this paradigm, an integrated understanding of the structural defects unique to that patient, as well as the functional consequences, can provide a unique approach to why certain symptoms occur in a specific patient. In other words, things that do not vary seldom provide variable outcomes.

The following are a few notes about our data modeling approach. First, parcellating the brain of structurally abnormal patients has long been a source of variability in the data, especially in the presence of brain atrophy. By using a machine learning approach based on structural connectivity patterns, we hold at least one variable (voxel identity in a parcellation) relatively constant, as the connectivity pattern should remain similar for a parcellation across brains (41, 47–49). Further, while the connectome has seemingly infinite interindividual variability, we hypothesize that clinically relevant phenotypes we are interested in at this early stage are less likely to result from the loss of rare individual variants in connectivity, and instead result from more constant interindividual connections. Thus, we eliminated many of the higher variance connectivity edges on the graphs to focus on similarities across individuals, and reduce the false discovery rate when scaling the results of machine learning models to individuals. In other words, we focused on brain connectivity, which we can more convincingly expect to be in a specific range.

As the potential treatment that non-invasively applying on cognitive decline, TMS may also begin to address etiological or syndrome's heterogeneity by targeting specific circuits to treat specific symptom clusters. However, it remains unknown whether the stimulation of different circuits is associated with improvement in different cognitive symptoms. In clinical practice, TMS targeting is usually based on scalp measurements and mostly without a flexible tracking device to fix the coil, resulting in different patients, or even the same patient during their series of sessions receiving stimulation of different sites in the prefrontal cortex.

Although there are important discoveries revealed by our study, there are also limitations. First, we included only 21 AD patients, which may lead to some potential bias for machine learning calculation-based results. Second, the way we eliminated one-third of parcellation pairs with the highest variance in the cohort of normal subjects, may have lost some original information, While, these areas were the smaller parcellations and is mainly aimed to reduce the problem of multiple comparisons (50). This should not be expected to introduce any subjective bias as it was based on the data. Finally, even after excluding one-third of the connectivity differences, the abnormities results we made have not been applied to selecting the individual target for rTMS treatment, although there may be a long way from being employed to the clinic, the outcome that we made may provide evidence for individualized and precise treatment for AD.

In conclusion, we demonstrated a machine learning–based approach to studying individual connectomes in a non-group averaged way. This critical exploratory work lays the groundwork for future larger-scale work in these patients. Our findings highlight the potential for a reproducible and generalizable functional brain imaging biomarker to aid the early diagnosis of AD and track its progression. This data-driven approach for identifying connectivity-specific targets may prove useful for other disorders and facilitate personalized neuromodulation therapy like rTMS. Collectively, our findings highlight the potential for mismatching between structural and functional brain imaging to provide a generalizable, and neuroscientifically interpretable imaging biomarker that may support clinicians in the non-invasive personalized treatment of AD. Further, our study may shed light on exploring new mechanisms and individualized stratagem based on the functional connectivity of brain networks in patients with dementia or even other neurodegenerative diseases.

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 Shenzhen People's Hospital Medical Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HR contributed to manuscript drafting and revising, data analysis and picture preparation. JZ and SC contributed to data acquisition and preparation. XS, SZ, XL, and L-YZ contributed to manuscript revising and data analysis. MS and YG contributed to study design, manuscript revising, data analysis, and study supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This research is funded by the Science and Technology planning project of Shenzhen Municipality (JCYJ20170818111012390), Sanming project of Medicine in Shenzhen (SYLY201905) and the Shenzhen health committee project (SZXJ2017034).

ACKNOWLEDGMENTS

We thank Peter Nicholas, Stephane Doyen from Omniscient neurotechnology company, for their technological assistance with the machine learning aspects of the project.

REFERENCES

1. Frisoni GB, Fox NC, Jack CR, Jr., Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol.* (2010) 6:67–77. doi: 10.1038/nrneurol.2009.215
2. Brier MR, Thomas JB, Ances BM. Network dysfunction in Alzheimer's disease: refining the disconnection hypothesis. *Brain Connect.* (2014) 4:299–311. doi: 10.1089/brain.2014.0236
3. Yan T, Wang W, Yang L, Chen K, Chen R, Han Y. Rich club disturbances of the human connectome from subjective cognitive decline to Alzheimer's disease. *Theranostics.* (2018) 8:3237–55. doi: 10.7150/thno.23772
4. Marcus C, Mena E, Subramaniam RM. Brain PET in the diagnosis of Alzheimer's disease. *Clin Nucl Med.* (2014) 39:e413–22; quiz e423–16. doi: 10.1097/RLU.0000000000000547
5. Peraza LR, Díaz-Parra A, Kennion O, Moratal D, Taylor JP, Kaiser M, et al. Structural connectivity centrality changes mark the path toward Alzheimer's disease. *Alzheimers Dement.* (2019) 11:98–107. doi: 10.1016/j.dadm.2018.12.004

6. Hojjati SH, Ebrahimzadeh A, Khazaee A, Babajani-Feremi A, Alzheimer's Disease Neuroimaging I. Predicting conversion from MCI to AD by integrating rs-fMRI and structural MRI. *Comput Biol Med.* (2018) 102:30–9. doi: 10.1016/j.compbiomed.2018.09.004
7. Wang P, Zhou B, Yao H, Xie S, Feng F, Zhang Z, et al. Aberrant hippocampal functional connectivity is associated with fornix white matter integrity in Alzheimer's disease and mild cognitive impairment. *J Alzheimers Dis.* (2020) 75:1153–68. doi: 10.3233/JAD-200066
8. Wang X, Huang W, Su L, Xing Y, Jessen F, Sun Y, et al. Neuroimaging advances regarding subjective cognitive decline in preclinical Alzheimer's disease. *Mol Neurodegener.* (2020) 15:55. doi: 10.1186/s13024-020-00395-3
9. Hojjati SH, Ebrahimzadeh A, Babajani-Feremi A. Identification of the early stage of Alzheimer's disease using structural MRI and resting-state fMRI. *Front Neurol.* (2019) 10:904. doi: 10.3389/fneur.2019.00904
10. Yang AC, Tsai SJ, Liu ME, Huang CC, Lin CP. The association of aging with white matter integrity and functional connectivity hubs. *Front Aging Neurosci.* (2016) 8:143. doi: 10.3389/fnagi.2016.00143
11. Gao H, Liu C, Li Y, Yang X. V2VR: reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability. In: *IEEE Transactions on Intelligent Transportation Systems.* (2020). p. 1–14.
12. Tucholka A, Grau-Rivera O, Falcon C, Rami L, Sánchez-Valle R, Lladó A, et al. Structural connectivity alterations along the Alzheimer's disease continuum: reproducibility across two independent samples and correlation with cerebrospinal fluid amyloid-β and Tau. *J Alzheimers Dis.* (2018) 61:1575–87. doi: 10.3233/JAD-170553
13. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science.* (2015) 349:255–60. doi: 10.1126/science.aaa8415
14. Ning Z, Dong P, Wang X, Hu X, Liu J, Guo L, et al. Partial computation offloading and adaptive task scheduling for 5G-enabled vehicular networks. In: *IEEE Transactions on Mobile Computing*. (2020). p. 1.
15. Castellazzi G, Cuzzoni MG, Cotta Ramusino M, Martinelli D, Denaro F, Ricciardi A, et al. A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by MRI selected features. *Front Neuroinform.* (2020) 14:25. doi: 10.3389/fninf.2020.00025
16. Jin D, Zhou B, Han Y, Ren J, Han T, Liu B, et al. Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. *Adv Sci.* (2020) 7:2000675. doi: 10.1002/advs.202000675
17. Li J, Jin D, Li A, Liu B, Song C, Wang P, et al. ASAF: altered spontaneous activity fingerprinting in Alzheimer's disease based on multisite fMRI. *Sci Bull.* (2019) 64:998–1010. doi: 10.1016/j.scib.2019.04.034
18. Wang X, Ning Z, Guo S, Wang L. Imitation learning enabled task scheduling for online vehicular edge computing. In: *IEEE Transactions on Mobile Computing*. (2020). p. 1.
19. Gao H, Kuang L, Yin Y, Guo B, Dou K. Mining consuming behaviors with temporal evolution for personalized recommendation in mobile marketing apps. *Mobile Netw Appl.* (2020) 25:1233–48. doi: 10.1007/s11036-020-01535-1
20. Jo T, Nho K, Saykin AJ. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci*. (2019) 11:220. doi: 10.3389/fnagi.2019.00220
21. Ning Z, Zhang K, Wang X, Guo L, Hu X, Huang J, et al. Intelligent edge computing in internet of vehicles: a joint computation offloading and caching solution. In: *IEEE Transactions on Intelligent Transportation Systems*. (2020). p. 1–14.
22. Weiler M, Stieger KC, Long JM, Rapp PR. Transcranial magnetic stimulation in Alzheimer's disease: are we ready? *eNeuro*. (2020) 7:1–11. doi: 10.1523/ENEURO.0235-19.2019
23. Tombaugh TN, McIntyre NJ. The mini-mental state examination: a comprehensive review. *J Am Geriatrics Soc.* (1992) 40:922–35. doi: 10.1111/j.1532-5415.1992.tb01992.x
24. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatrics Soc.* (2005) 53:695–9. doi: 10.1111/j.1532-5415.2005.53221.x
25. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology.* (1984) 34:939–44. doi: 10.1212/WNL.34.7.939
26. Afyouni S, Nichols TE. Insight and inference for DVARS. *Neuroimage.* (2018) 172:291–312. doi: 10.1016/j.neuroimage.2017.12.098
27. Behzadi Y, Restom K, Liau J, Liu TT. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage.* (2007) 37:90–101. doi: 10.1016/j.neuroimage.2007.04.042
28. Garyfallidis E, Brett M, Amirbekian B, Rokem A, van Der Walt S, Descoteaux M, et al. Dipy, a library for the analysis of diffusion MRI data. *Front Neuroinform.* (2014) 8:8. doi: 10.3389/fninf.2014.00008
29. Gao H, Xu Y, Yin Y, Zhang W, Li R, Wang X. Context-aware QoS prediction with neural collaborative filtering for internet-of-things services. *IEEE Internet Things J.* (2020) 7:4532–42. doi: 10.1109/JIOT.2019.2956827
30. Ning Z, Kwok RYK, Zhang K, Wang X, Obaidat MS, Guo L, et al. Joint computing and caching in 5G-envisioned internet of vehicles: a deep reinforcement learning-based traffic control system. In: *IEEE Transactions on Intelligent Transportation Systems*. (2020). p.1–12.
31. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. *Nature.* (2016) 536:171–8. doi: 10.1038/nature18933
32. Baker CM, Burks JD, Briggs RG, Conner AK, Glenn CA, Morgan JP, et al. A connectomic atlas of the human cerebrum-chapter 2: the lateral frontal lobe. *Oper Neurosurg.* (2018) 15:S10–74. doi: 10.1093/ons/opy254
33. Baker CM, Burks JD, Briggs RG, Smitherman AD, Glenn CA, Conner AK, et al. The crossed frontal aslant tract: a possible pathway involved in the recovery of supplementary motor area syndrome. *Brain Behav.* (2018) 8:e00926. doi: 10.1002/brb3.926
34. Briggs RG, Conner AK, Baker CM, Burks JD, Glenn CA, Sali G, et al. A connectomic atlas of the human cerebrum-chapter 18: the connectional anatomy of human brain. *Networks Oper Neurosurg.* (2018) 15:S470–80. doi: 10.1093/ons/opy272
35. Burks JD, Boettcher LB, Conner AK, Glenn CA, Bonney PA, Baker CM, et al. White matter connections of the inferior parietal lobule: a study of surgical anatomy. *Brain Behav.* (2017) 7:e00640. doi: 10.1002/brb3.640
36. Dadi K, Rahim M, Abraham A, Chyzhyk D, Milham M, Thirion B, et al. Benchmarking functional connectome-based predictive models for resting-state fMRI. *Neuroimage.* (2019) 192:115–34. doi: 10.1016/j.neuroimage.2019.02.062
37. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings–a practical guide. *Biol Rev.* (2017) 92:1941–68. doi: 10.1111/brv.12315
38. Allan PG, Briggs RG, Conner AK, O'Neal CM, Bonney PA, Maxwell BD, et al. Parcellation-based tractographic modeling of the dorsal attention network. *Brain Behav*. (2019) 9:e01365. doi: 10.1002/brb3.1365
39. Burks JD, Conner AK, Bonney PA, Glenn CA, Baker CM, Boettcher LB, et al. Anatomy and white matter connections of the orbitofrontal gyrus. *J Neurosurg.* (2018) 128:1865–1872. doi: 10.3171/2017.3.JNS162070
40. Conner AK, Briggs RG, Palejwala AH, Sali G, Sughrue ME. The safety of post-operative elevation of mean arterial blood pressure following brain tumor resection. *J Clin Neurosci.* (2018) 58:156–9. doi: 10.1016/j.jocn.2018.09.001
41. Thomas C, Sadeghi N, Nayak A, Trefler A, Sarlls J, Baker CI, et al. Impact of time-of-day on diffusivity measures of brain tissue derived from diffusion tensor imaging. *Neuroimage.* (2018) 173:25–34. doi: 10.1016/j.neuroimage.2018.02.026
42. Jessen F, Amariglio RE, Buckley RF, van der Flier WM, Han Y, Molinuevo JL, et al. The characterisation of subjective cognitive decline. *Lancet Neurol.* (2020) 19:271–8. doi: 10.1016/S1474-4422(19)30368-0
43. Wang Y, Yang Y, Guo X, Ye C, Gao N, Fang Y, et al. A novel multimodal MRI analysis for Alzheimer's disease based on convolutional neural network. *Conf Proc IEEE Eng Med Biol Soc.* (2018) 2018:754–7. doi: 10.1109/EMBC.2018.8512372
44. Li Y, Yao H, Lin P, Zheng L, Li C, Zhou B, et al. Frequency-dependent altered functional connections of default mode network in Alzheimer's disease. *Front Aging Neurosci.* (2017) 9:259. doi: 10.3389/fnagi.2017.00259
45. Greicius MD, Srivastava G, Reiss AL, Menon V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc Natl Acad Sci USA.* (2004) 101:4637–4642. doi: 10.1073/pnas.0308627101
46. Zhang L, Zuo XN, Ng KK, Chong JSX, Shim HY, Ong MQW, et al. Distinct BOLD variability changes in the default mode and

salience networks in Alzheimer's disease spectrum and associations with cognitive decline. *Sci Rep.* (2020) 10:6457. doi: 10.1038/s41598-020-63540-4

47. Allan PG, Briggs RG, Conner AK, O'Neal CM, Bonney PA, Maxwell BD, et al. Parcellation-based tractographic modeling of the ventral attention network. *J Neurol Sci.* (2020) 408:116548. doi: 10.1016/j.jns.2019.116548
48. Conner AK, Briggs RG, Rahimi M, Sali G, Baker CM, Burks JD, et al. A connectomic atlas of the human cerebrum-chapter 10: tractographic description of the superior longitudinal fasciculus. *Oper Neurosurg.* (2018) 15:S407–22. doi: 10.1093/ons/opy264
49. Conner AK, Briggs RG, Rahimi M, Sali G, Baker CM, Burks JD, et al. A connectomic atlas of the human cerebrum-chapter 12: tractographic description of the middle longitudinal fasciculus. *Oper Neurosurg.* (2018) 15:S429–35. doi: 10.1093/ons/opy266
50. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage.* (2002) 15:870–8. doi: 10.1006/nimg.2001.1037

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: 22 December 2020
doi: 10.3389/fgene.2020.598064

Parkinson's Disease in Teneurin Transmembrane Protein 4 (*TENM4*) Mutation Carriers

Jia-Li Pu†, Ting Gao†, Xiao-Li Si, Ran Zheng, Chong-Yao Jin, Yang Ruan, Yi Fang, Ying Chen, Zhe Song, Xin-Zhen Yin, Ya-Ping Yan, Jun Tian and Bao-Rong Zhang**

Department of Neurology, College of Medicine, Second Affiliated Hospital, Zhejiang University, Hangzhou, China

Introduction: Mutations in the teneurin transmembrane protein 4 (*TENM4*) gene, known to be involved in neuropsychiatric disorders, have been identified in three pedigree of essential tremor (ET) from Spain. ET has overlapping clinical manifestations and epidemiological symptoms with Parkinson's disease (PD), suggesting these two disorders may reflect common genetic risk factors. In this study, we investigated clinical and genetic manifestations in four unrelated pedigrees with both ET and PD in which *TENM4* variants were identified.

Methods: We subsequently explored whether *TENM4* variants contributed to the risk of developing PD. The frequency of *TENM4* variants was evaluated from four PD pedigrees and other 407 subjects.

Results: The results revealed 12 different novel heterozygous variants, all at low frequency. A clear general enrichment of *TENM4* variants was detected in early onset PD patients ($p < 0.001$, OR = 5.264, 95% CI = 1.957–14.158).

Conclusion: The results indicate that rare *TENM4* variants may be associated with an increased risk of PD.

Keywords: variant, genetic testing, Parkinson's disease, pedigree, *TENM4*

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Jun Wang,
Qingdao University, China
Marcelo Adrian Marti,
University of Buenos Aires, Argentina
Pingyi Xu,
First Affiliated Hospital of Guangzhou Medical University, China

****Correspondence:***
Bao-Rong Zhang
brzhang@zju.edu.cn
Jun Tian
2313061@zju.edu.cn;
juntian@zju.edu.cn

†These authors have contributed equally to this work

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

Received: *25 September 2020*
Accepted: *02 December 2020*
Published: *22 December 2020*

Citation:
Pu J-L, Gao T, Si X-L, Zheng R, Jin C-Y, Ruan Y, Fang Y, Chen Y, Song Z, Yin X-Z, Yan Y-P, Tian J and Zhang B-R (2020) Parkinson's Disease in Teneurin Transmembrane Protein 4 (TENM4) Mutation Carriers. Front. Genet. 11:598064. doi: 10.3389/fgene.2020.598064

INTRODUCTION

Parkinson's disease (PD), one of the most frequent neurodegenerative disorders, is mainly characterized by bradykinesia, resting tremor and rigidity (Lees et al., 2009). Interactions between environmental and genetic factors underlie the degeneration of nigral dopaminergic (DA) neurons and ensuing PD. Genetic factors account for ∼5−10% of PD cases (Deng et al., 2018). To date, 27 Mendelian genes have been reported to be linked with PD, and genome-wide association studies have succeeded in identifying many low-risk variants (Deng et al., 2018; Lunati et al., 2018).

Essential tremor (ET) is a common hyperkinetic movement disorder with an estimated prevalence of 5% among people over 65 years old (Deuschl et al., 2015). ET is characterized mainly by rhythmic, involuntary shaking of parts of the body, and occurs exclusively during voluntary movements or in positions against gravity. While the majority of PD cases are sporadic, ET has a strong genetic component, and more than half of affected individuals have a positive family history (Louis and Ottman, 2006). Although ET and PD are generally considered distinct entities, Spanaki and Plaitakis (2009) found ET occurred more frequently in relatives of PD patients, compared with that in controls. Furthermore, the risk of developing PD is up to fourfold greater in ET sufferers (Algarni and Fasano, 2018). The overlapping clinical manifestations and epidemiological symptoms suggest that PD and ET may underlie common genetic risk factors.

Mutations in the Teneurin Transmembrane Protein 4 (*TENM4*; MIM 610084) gene, known to be involved in neuropsychiatric disorders (Xue et al., 2018). have been identified recently in three pedigrees of ET from Spain (Hor et al., 2015). Additionally, *in vitro* and model organism analyses showed that mutations in *TENM4* gene result in protein mislocalization and axon guidance defects (Hor et al., 2015). However, further screening in a cohort of 269 Canadian ET cases and 288 matched controls revealed a negative association between *TENM4* and the Canadian population (Houle et al., 2017). In our previous study, Yan et al. (2020) found no evidence support that *TENM4* associated with ET. In addition, Chao et al. (2016) found that the c.4324 G > A mutation in *TENM4* originally identified by Hor and colleagues (Hor et al., 2015) was also present in the control group (379 ET cases and 398 healthy controls) in a Chinese population. Thus, similar studies have yielded inconsistent results.

Increasing evidence suggests that ET and PD may share genetic mutations, and a subset of patients may have a combination of long-standing ET with subsequent PD (ET-PD). Furthermore, one family with five affected individuals presented with either ET or PD, consistent with mutation of the of *PRKN* (*PARK2*) gene (Pellecchia et al., 2007). Unal Gulsuner et al. (2014) reported that High Temperature Requirement Protein A2 (*HTRA2*) is responsible for hereditary essential tremor and that homozygotes for this allele develop Parkinson disease. And Fused in sarcoma (*FUS*) mutations have been found in individuals with ALS/PD (Yan et al., 2010).

The aim of the present study was to further explore the associations between *TENM4* mutations and PD, and investigate whether *TENM4* variant carriers are at increased risk of developing PD. We first explored the clinical features and genetic features of four ET-PD pedigrees, then investigated whether *TENM4* variants might be associated with PD by comparing mutations in a cohort of sporadic PD cases and controls.

MATERIALS AND METHODS

Family Study

Pedigrees

Four pedigrees of ET-PD with *TENM4* mutations were included in this study. Four probands and their family members were examined by two neurologists and genetically tested for neurodegenerative disorders (PD, ET, Alzheimer's disease, etc.). Clinical and demographic features of the probands in four family pedigrees are described in "Results" section (**Table 1**). The study was approved by the Medical Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine in accordance with the Declaration of Helsinki. All subjects participated in this study completed informed consent before the evaluation and original sample collection.

Genetic Analysis

Blood samples (2 ml) were collected from all cases, and genomic DNA was extracted from peripheral blood leucocytes using standard procedures. Probands and family members were screened for *TENM4* (NCBI transcript NM_001098816.2), *HTRA2* (NCBI transcript NM_013247), and *FUS* (NCBI transcript NM_004960.3) mutations by standard bi-directional Sanger sequencing of all coding exons and exon-intron boundaries (primer sequences available on request). Dosage analysis for *TENM4* exonic deletions and duplications was performed by multiplex ligation-dependent probe amplification (MLPA, MRC) (Mencacci et al., 2014). The other known PD pathogenic genes (*SNCA*, *GBA*, *LRRK2*, *UCHL1*, *VPS35*, *PRKN*, *PINK1*, *DJ-1*, *ATP13A2*, *GIGYF2*, *PLA2G6*, *HtrA2*, *FBXO7*, *SYNJ1*, *DNAJC6*, *DNAJC13*, *CHCHD2*, *Rab39B*) were also analyzed in all participants.

Target Sequencing, Variant Filtering, Identification, and Analysis

Participants

The study included 207 unrelated patients with PD and 200 healthy control subjects from East China. Healthy controls were recruited from local communities. All patients were enrolled from outpatient neurology clinics of the Second Affiliated Hospital of Zhejiang University School of Medicine and local communities, and evaluated by two movement disorder specialists for diagnosis of PD according to criteria provided by the Movement Disorder Society (Postuma et al., 2015). The exclusion criteria were described in our previous study (Gao et al., 2019). To summarize, participants with secondary causes of parkinsonism such as vascular, drug-induced, and toxin-induced, and other neurodegenerative diseases such as progressive supranuclear palsy, multiple system atrophy, essential tremor, Wilson's disease and ET convert to PD were excluded. Additionally, other internal diseases which might also present tremor symptom such as hyperthyroidism were also excluded. The protocol was approved by the Medical Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine in accordance with the Declaration of Helsinki. Written informed consent was completed for every participant before the evaluation and sample collection.

DNA Preparation, Target Resequencing, Variant Filtering, Validation, and Analysis

TENM4 and two additional ET-related genes (*HTRA2* and *FUS*) were selected as targeted genes for capturing and sequence analyses. Molecular inversion probes (MIPs) were designed to capture all exons and intron-exon boundaries (5 bp flanking sequences) of target genes (Yang et al., 2019). Briefly, fragmented genomic DNA was captured by a customized array designed to target all exons, splicing sites, and flanking intronic sequences of the three genes (NimbleGen, Roche). Captured DNA fragments were sequenced on an Illumina HiSeq2000 Analyzer (Ahmed et al., 2003). Variants were filtered based on a read depth $\geq 4\times$, a genotype quality $\geq$ 20, and the proportions of reads with alternative alleles $\geq$ 0.3. Two publicly available resources were

TABLE 1 | Clinical manifestation of the probands in four family pedigrees.

Variants		Sex	AAO	IS	B	R	T	PI	Levodopa responsive	DK	DM
c.974G > A	p.R325Q	M	72	T	+	+	+	–	Responsive	–	–
c.2287C > T	p.R763C	F	52	B	+	+	–	–	Slightly good	–	–
c.6209G > A	p.R2070H	M	50	T	+	+	+	–	Not treated	–	–
c.5545A > G	p.T1849A	M	47	B	+	+	–	–	Responsive	–	–

AAO, age at onset; IS, initial symptoms; B, bradykinesia; R, rigidity; T, resting tremor; PI, postural instability; DK, dyskinesia; DM, dementia; +, positive; –, negative.

used to obtain variant frequency data; the 1000 Genomes Project and the Exome Aggregation Consortium (Dec 2019).

Criteria for Pathogenicity of Rare Variants

All non-synonymous variants were analyzed by a database of human non-synonymous SNVs and their functional predictions and annotations (dbNSFP, versions 3.3–3.5) (Liu et al., 2016). For interpretation of the validated variants, multiple prediction indices were adopted to clarify their pathogenicity, and variants were considered as likely pathogenic based on Sorting Intolerant From Tolerant (SIFT) score < 0.05 (Ng and Henikoff, 2003), Polyphen-2 score > 0.86 (Adzhubei et al., 2013) and Combined Annotation-Dependent Depletion (CADD) score > 12.35 (Kircher et al., 2014).

Statistical Analysis

Variants with a minor allele frequency < 0.1% (gnome AD or 1,000 G) were defined as rare variants and included in the gene-based burden test. The association between rare variants and PD was analyzed using Fisher's exact tests, odds ratio (OR) and 95% confidence intervals (CI). All statistical analyses were performed using IBM SPSS Statistics 23.0[1], and two-tailed $p < 0.05$ was considered statistically significant.

RESULTS

Family Study

Family A

The proband (Case II-3, **Figure 1A**) was a 75 year-old right-handed male of East Chinese origin with PD, with disease onset at age 72, and tremor of the left foot and arm. He now presents with anosmia, constipation, progressive loss of dexterity and slowness in the left foot. Examination showed an asymmetric rigid-akinetic parkinsonian syndrome with rest tremor and bradykinesia in the left foot and arm. Postural instability, dyskinesia and dementia were not observed. The efficacy of levodopa therapy was responsive and symptoms slightly relieved. His mother (Case I-2) passed away but was described with tremor in both hands. Whereas, further clinical information couldn't be acquired. Examination of the proband's brother and sister revealed kinetic tremor in both hands, without dyskinesia or hypertonia. They were clinically diagnosed of essential tremor and on no medications regards of mild symptoms.

Gene screening in this family revealed one variant, *TENM4* c.974G > A; p.R325Q (carried by Case II-3, II-1, and II-2), in the proband and his sister and brother. The R325Q (rs373911172) mutation has not been reported previously in ET or PD. Children of the proband and that of his sister were unaffected and non-carriers, as well as father of the proband. No rare variants of *HTRA2* or *FUS* were detected in any of the tested family members. *PRKN* and *LRRK2* gene mutations were found in the proband, while the pathogenicity analysis assigned the mutations as benign.

Family B

The proband (Case II-4, **Figure 1B**) was a 55 year-old female with bradykinesia, rest tremor in lower limbs and being first diagnosed with PD 6 years ago. She complained of bradykinesia and poor dexterity, and suffered tremor in both hands 1 year ago. The symptoms slightly ameliorated after taking levodopa. Her father (Case I-1) was diagnosed with PD in his sixties and died 5 years ago. Her sister (Case II-3) presented with head tremors when nervous or excited, while bradykinesia and rest tremor were not observed. Genetic tests revealed that the proband and her sister carried *TENM4* c.2287C > T; p.R763C. One of her brothers (Case II-5) was not a carrier. Rare variants of *HTRA2* and *FUS* were not identified. Genetic analysis data were not available for other family members. No other PD pathogenic gene mutations were found in the proband.

Family C

The proband (Case II-1, **Figure 1C**) was a 72 year-old male who presented with bilateral hand tremors at the age of 50. He recently attended the outpatient clinic for 2 years for stiffness of facial expression and slowness of the left hand and foot. No dopaminergic drugs had been prescribed. His mother (Case I-1) had a history of tremor in both hands, but did not experience dexterity or walking problems. No tremor or bradykinesia were observed in his brothers or sister.

The proband was heterozygous for *TENM4* c.6209G > A; p.R2070H. *LRRK2* mutation was found in the proband and allocated as benign by rare variant pathogenicity analysis. No rare variants were detected for *HTRA2* or *FUS*. His father and brother (Case II-3) are non-carriers for *TENM4* c.6209G > A. Unfortunately, genetic information for other members of the family was unavailable.

Family D

The proband (Case III-3, **Figure 1D**) was a 51 year-old right-handed male with bradykinesia been diagnosed as Parkinson's disease for 4 years. Physical examination showed a mask face and

[1] https://www.ibm.com/support/knowledgecenter/SSLVMB_23.0.0/spss/product_landing

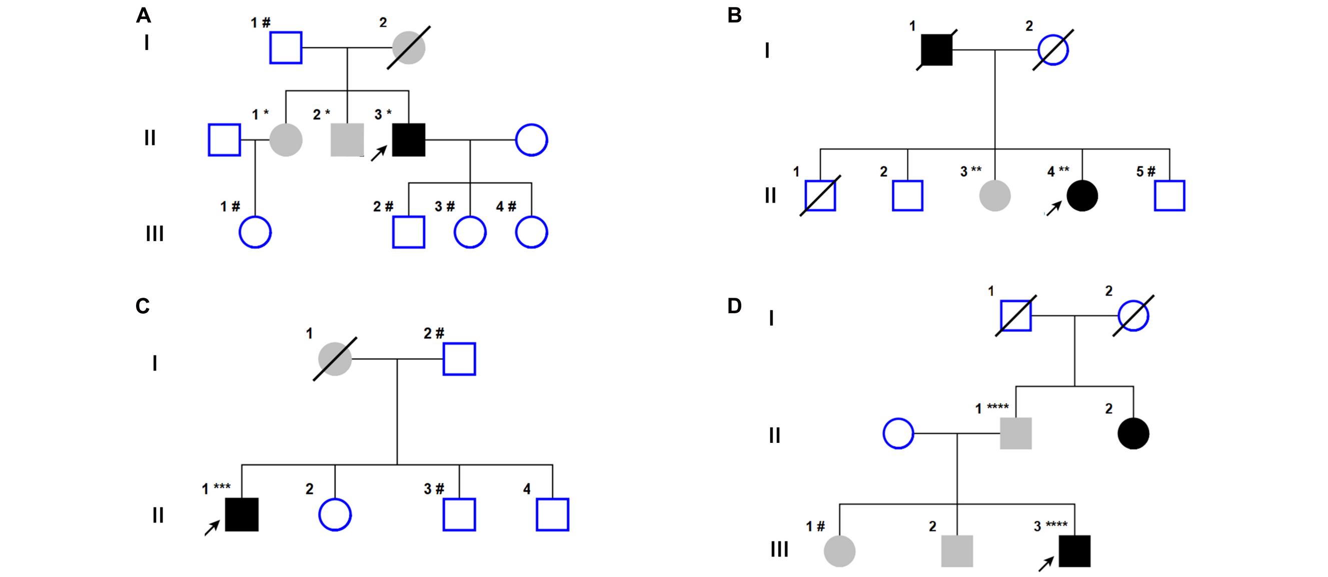


FIGURE 1 | Pedigrees of the four families with TENM4 mutations involved in this study. **(A–D)** Pedigress of four different families with TENM4 mutations. Notes: black symbols, individuals with ET and PD; blank symbols, unaffected; gray symbols, individuals who are reported to have ET by history but some are not examined; arrow, probands; diagonal lines, deceased individuals; circle, female; square, male. *: c.974 G > A/wild type; **: c.2287 C > T/wild type; ***: c.6209 G > A/wild type; ****: c.5545 A > G/wild type; #: wild type/wild type.

rigid-akinetic parkinsonian syndrome without tremor of the right lower extremities. Dopaminergic therapy resulted in tangible improvement in parkinsonian symptoms. 11C-labeled 2β-carbomethoxy-3β-(4-fluorophenyl) tropane positron emission tomography/computed tomography (11^{C}-CFT PET/CT) analysis revealed an asymmetric bilateral reduced tracer uptake, more marked in the left putamen. His sister (Case III-1), brother (Case III-2), and father (Case II-1) presented with tremor of both hands but without bradykinesia and rigidity. By contrast, his aunt had difficulty walking and poor dexterity and was diagnosed as Parkinson's disease in her sixties.

Gene sequencing of all available family members revealed that the proband and his father were heterozygous for the rare *TENM4* c.5545A > G; p.T1849A mutation, while his sister (Case III-1) was a non-carrier. *PRKN* and *LRRK2* mutations were found in the proband, however, rare variant pathogenicity analysis determined them as benign mutations.

Targeted Gene Panel Sequencing

We hypothesized that pathogenic variants in *TENM4* may also be found in subjects with PD without a family history. To investigate this, we examined targeted gene panel sequencing data for a large cohort of patients predominantly affected by PD, alongside controls. In total, 207 patients with sporadic PD (male/female = 112/95, age = 52.83 ± 10.56 years) and 200 healthy control participants (male/female = 85/115, age = 46.29 ± 11.05 years) were included and analyzed (**Table 2**). The percentage read depth of target genes was 98, 96, and 92% of bases covered by at least 4×, 10×, and 20×, respectively.

TABLE 2 | Summary of sporadic PD-control demographics.

Series	Number	Age	Male/Female
Total PD	207	52.83 ± 10.56	112/95
EOPD (AAO ≤ 50)	100	44.74 ± 7.34	53/47
LOPD (AAO > 50)	107	60.34 ± 6.86	59/48
Controls	200	46.29 ± 11.05	85/115

PD, Parkinson's disease; EOPD, early onset Parkinson's disease; LOPD, late onset Parkinson's disease; AAO, age at onset.

Overall, 12 rare non-synonymous-coding variants with a minor allele frequency < 0.1% were identified in the exon regions of the *TENM4* gene after applying quality filter (**Table 3**). None of these rare variants have been reported previously, and all were categorized as disease-causing based on SIFT, Polyphen-2 and CADD values, and remained conserved based on GERP + + prediction (Ioannidis et al., 2016). The structures and functions were predicted as altered structures and/or functions (**Supplementary Materials**). Unfortunately, due to technical issues, one of the variants (p.R763C) could not be sequenced in healthy controls. In addition, six of these variants (p.R1952H, p.T1849A, p.Y1760F, p.D632N, p.G222R, and p.Q2735E) were absent in our gender-matched healthy control cohort (**Table 3**), and their locations are depicted in **Figure 2**. However, no rare variants of *HTRA2* or *FUS* were detected in PD or healthy controls. None of the previously reported PD risk genes had been identified in all participants.

Gene-Based Burden Analysis

To determine whether these rare variants of *TENM4* contribute collectively to sporadic PD risk in our cohort, we performed a gene-based burden analysis using Fisher's exact test (Nicolae, 2016), and a clear general enrichment was detected for early

TABLE 3 | Summary of variants of TENM4 classified as likely pathogenic in the cohort.

Variants	Position	Freq. patient (%)	Freq. control (%)	Freq. 1,000 G	Freq. ExAC	dbSNP ID	SIFT score	Polyphen score	CADD score	GERP++ score
p.R763C	78498021	0.4651	NA	NA	1.03E-4	rs751467112	0	0.998	35	4.98
p.R325Q	78600940	0.9302	0.5000	3.99E-4	4.60E-5	rs373911172	0.009	0.999	35	4.81
p.R1952H	78381535	0.4651	0	3.99E-4	2.09E-4	rs140341040	0.205	1	24.8	4.93
p.T1849A	78383326	0.9302	0	NA	1.08E-4	rs772977333	0.003	0.994	24.6	5.65
p.Y1760F	78387414	0.9302	0	NA	8.72E-6	rs745395614	0.01	0.997	24.3	4.71
p.R2733Q	78369215	0.4651	2.0000	7.99E-4	8.34E-6	rs185503085	0.047	0.85	23.2	5.65
p.D632N	78523251	0.9302	0	3.99E-4	2.82E-4	rs370767956	0.221	0.923	23.2	4.94
p.L1937V	78381581	0.4651	1.5000	7.99E-4	NA	rs192931562	0.238	0.997	22.8	4.81
p.V2423M	78380123	0.4651	0.5000	NA	8.34E-6	rs759903805	0.024	0.999	26.5	5.67
p.A1165T	78437181	0.4651	1.0000	2.00E-4	4.14E-5	rs550968777	0.121	0.968	23.9	5.32
p.G222R	78614398	0.4651	0	3.99E-4	2.20E-4	rs541146378	0.076	0.989	23.5	5.64
p.Q2735E	78369210	0.4651	0	NA	8.28E-6	rs527857070	0.242	0.969	12.59	5.65

Threshold values for deleteriousness: SIFT-less than 0.05; Polyphen-2 greater than 0.86; CADD-greater than 12.35; GERP++ is a score for the conservation of the amino acid; scores > 3 can be considered as highly conserved. NA, not found; Freq, frequency; 1,000 G, 1,000 Genomes Project; ExAC, Exome Aggregation Consortium.

onset Parkinson's disease (EOPD, age at onset ≤ 50) patients ($p < 0.001$, OR = 5.264, 95% CI = 1.957–14.158). No significant differences were found for late onset Parkinson's disease (LOPD, age at onset > 50) or total PD (**Supplementary Table 1**).

DISCUSSION

Advances in next-generation sequencing have revealed a growing number of causal genes in Mendelian form PD patients (Deng et al., 2018). However, a large number of early onset cases still remain to be explained, which indicates that there are many genetic factors yet to be clarified. The connection between PD and ET has received much attention (Jankovic, 1989). Accumulating evidence supports an association between PD and ET, including overlapping clinical features, obviously increased prevalence of PD in patients with a family history of ET, and increased prevalence of ET in family members of PD patients (Tarakad and Jankovic, 2018).

Recent discoveries confirmed links between ET and *LINGO1*, *FUS,* and *TENM4* (Hor et al., 2015; Clark and Louis, 2018). In our current family study, we assessed four unrelated ET and PD pedigrees (family coexistence of ET and PD) in which *TENM4* variants were identified in individuals without evidence of mutations in *LINGO1* or *FUS* genes. Most cases presenting PD phenotypes were *TENM4* variant carriers. Thus, we speculated that *TENM4* may be linked to the risk of developing PD.

We subsequently identified 12 novel rare variants of *TENM4* in a Chinese cohort of sporadic PD patients that may be associated with PD developing, including five that were also present in controls. With other PD related genes tested in our sporadic PD patients, no significant risk genes were found, which therefore strengthen our hypothesis that mutations in TNEM4 gene may associated with PD. Burden analysis indicated no overrepresentation of variant alleles in sporadic PD cases, but did reveal an association between *TENM4* rare variants and disease in EOPD case-controls. It should be noted that the results of burden analysis can be impacted by the detection methods, read depth, and data from the GnomAD database (compared to using ethnically-, age-, and gender-matched controls). Thus, data from burden analysis should be interpreted cautiously. However, the results implied *de novo* variants or incomplete penetrance.

Despite dramatic advances in our understanding of the genetic basis of PD, a large number of early onset and sporadic cases still remain to be clarified. There is a possible functional link between Mendelian genes and sporadic PD, and previous studies suggest that rare and low-frequency variants of PD Mendelian genes may play a role in sporadic forms of the disease (Kun-Rodrigues et al., 2015; Spataro et al., 2015). In addition, previous research confirmed the positive contribution of rare coding GTP cyclohydrolase1 (*GCH1*), the causative gene in dopamine related dystonia (DRD) for which gene variants have been identified in a large cohort of sporadic PD cases (Mencacci et al., 2014). Our present study is the first to link ET with the *TENM4* gene in PD cases. Those PD patients with *TENM4* mutations mildly response to levodopa treatment in four pedigrees indicated an undefined mechanism of gene-related on dopaminergic therapy of PD.

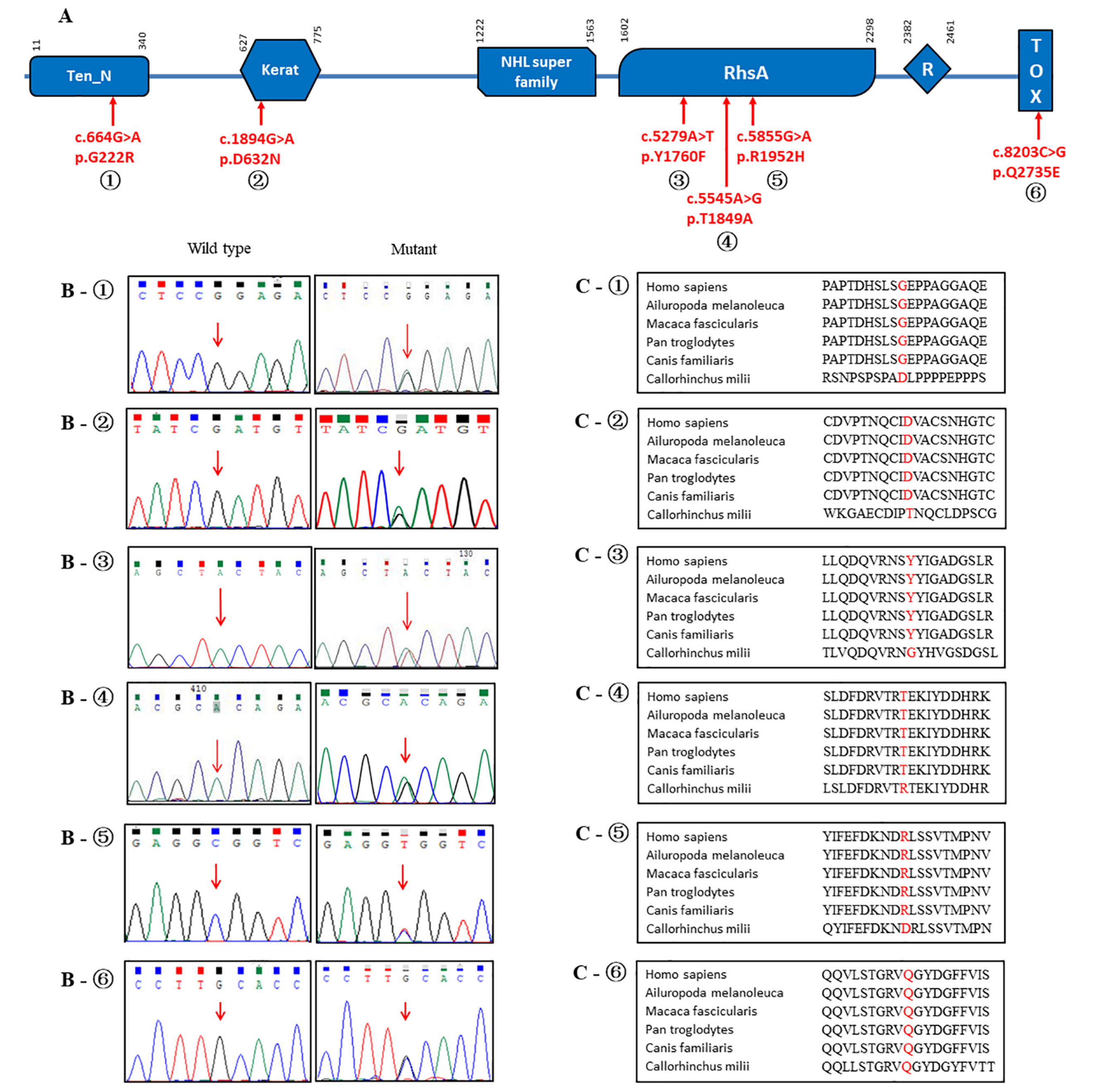

A
11
340
Ten_N
627
Kerat
775
1222
NHL super family
1563
1602
RhsA
2298
2382
R
2461
T O X
c.664G>A p.G222R ①
c.1894G>A p.D632N ②
c.5279A>T p.Y1760F ③
c.5545A>G p.T1849A ④
c.5855G>A p.R1952H ⑤
c.8203C>G p.Q2735E ⑥
Wild type
Mutant
B - ①
B - ②
B - ③
B - ④
B - ⑤
B - ⑥
C - ①
Homo sapiens PAPTDHSLSGEPPAGGAQE
Ailuropoda melanoleuca PAPTDHSLSGEPPAGGAQE
Macaca fascicularis PAPTDHSLSGEPPAGGAQE
Pan troglodytes PAPTDHSLSGEPPAGGAQE
Canis familiaris PAPTDHSLSGEPPAGGAQE
Callorhinchus milii RSNPSPSPADLPPPPEPPPS
C - ②
Homo sapiens CDVPTNQCIDVACSNHGTC
Ailuropoda melanoleuca CDVPTNQCIDVACSNHGTC
Macaca fascicularis CDVPTNQCIDVACSNHGTC
Pan troglodytes CDVPTNQCIDVACSNHGTC
Canis familiaris CDVPTNQCIDVACSNHGTC
Callorhinchus milii WKGAECDIPTNQCLDPSCG
C - ③
Homo sapiens LLQDQVRNSYYIGADGSLR
Ailuropoda melanoleuca LLQDQVRNSYYIGADGSLR
Macaca fascicularis LLQDQVRNSYYIGADGSLR
Pan troglodytes LLQDQVRNSYYIGADGSLR
Canis familiaris LLQDQVRNSYYIGADGSLR
Callorhinchus milii TLVQDQVRNGYHVGSDGSL
C - ④
Homo sapiens SLDFDRVTRTEKIYDDHRK
Ailuropoda melanoleuca SLDFDRVTRTEKIYDDHRK
Macaca fascicularis SLDFDRVTRTEKIYDDHRK
Pan troglodytes SLDFDRVTRTEKIYDDHRK
Canis familiaris SLDFDRVTRTEKIYDDHRK
Callorhinchus milii LSLDFDRVTRTEKIYDDHR
C - ⑤
Homo sapiens YIFEFDKNDRLSSVTMPNV
Ailuropoda melanoleuca YIFEFDKNDRLSSVTMPNV
Macaca fascicularis YIFEFDKNDRLSSVTMPNV
Pan troglodytes YIFEFDKNDRLSSVTMPNV
Canis familiaris YIFEFDKNDRLSSVTMPNV
Callorhinchus milii QYIFEFDKNDRLSSVTMPN
C - ⑥
Homo sapiens QQVLSTGRVQGYDGFFVIS
Ailuropoda melanoleuca QQVLSTGRVQGYDGFFVIS
Macaca fascicularis QQVLSTGRVQGYDGFFVIS
Pan troglodytes QQVLSTGRVQGYDGFFVIS
Canis familiaris QQVLSTGRVQGYDGFFVIS
Callorhinchus milii QQLLSTGRVQGYDGYFVTT

FIGURE 2 | Schematic representations of the *TENM4* protein with blue boxes indicating different domains. **(A)** Variants found in present study are indicated by arrows. **(B)** Electropherograms of the sequence of the variants. **(C)** The position and surroundings of those variants are highly conserved across different species.

The pathogenetic mechanism linking *TENM4* mutations with ET is uncertain. Biochemical evidence from *TENM4*-deficient mice revealed loss of embryonic mesoderm and differentiation in a cell-autonomous manner (Nakamura et al., 2013). Furthermore, functional studies are needed to elucidate the importance mutations in this gene.

The limited contribution of the *TENM4* gene to PD revealed by our study could be due to a lack of functional studies confirm pathogenic variants. Furthermore, we evaluated a cohort of sporadic PD cases in which *TENM4* variants may not reflect the frequency in familial cases. The relatively small sample size and absence of family co-segregation may be limitations of our study.

In conclusion, we provide evidence that rare *TENM4*-coding variants may be considered a risk factor for PD. However, determining how *TENM4* mutations known to cause ET may be related to risk alleles in PD requires further investigation. Due to racial heterogeneity and the limited sample size of our cohort, more robust independent studies are needed to further illuminate the relationship between PD and *TENM4* gene variants.

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 the Medical Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

J-LP and TG were involved with study concept and design, acquisition of data, analysis, and interpretation of data, drafting and revising the manuscript. TG, X-LS, RZ, C-YJ, YR, YF, and YC were involved with acquisition of data, analysis, and interpretation of data. B-RZ, J-LP, JT, ZS, Y-PY, and X-ZY were involved with PD patients' recruitment. B-RZ and JT were involved with revising the manuscript and were responsible for supervision of study. All authors listed meet the criteria for authorship.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 81520108010 and 81771216), the Key Research and Development Program of Zhejiang Province (No. 2020C03020), and the Natural Science Foundation of Zhejiang Province (No. LY18H090003).

ACKNOWLEDGMENTS

We express our gratitude to all the patients and their families who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.598064/full#supplementary-material

REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7:Unit720.

Ahmed, Z. M., Riazuddin, S., Riazuddin, S., and Wilcox, E. R. (2003). The molecular genetics of Usher syndrome. *Clin. Genet.* 63, 431–444. doi: 10.1034/j.1399-0004.2003.00109.x

Algarni, M., and Fasano, A. (2018). The overlap between essential tremor and Parkinson disease. *Parkinsonism Relat. Disord.* 46(Suppl. 1), S101–S104.

Chao, Y. X., Lin, Ng, E. Y., Tio, M., Kumar, P., Tan, L., et al. (2016). Essential tremor linked TENM4 mutation found in healthy Chinese individuals. *Parkinsonism Relat Disord* 31, 139–140. doi: 10.1016/j.parkreldis.2016.05.003

Clark, L. N., and Louis, E. D. (2018). Essential tremor. *Handb. Clin. Neurol.* 147, 229–239.

Deng, H., Wang, P., and Jankovic, J. (2018). The genetics of Parkinson disease. *Ageing Res. Rev.* 42, 72–85.

Deuschl, G., Petersen, I., Lorenz, D., and Christensen, K. (2015). Tremor in the elderly: essential and aging-related tremor. *Mov. Disord* 30, 1327–1334. doi: 10.1002/mds.26265

Gao, T., Wu, J., Zheng, R., Fang, Y., Jin, C. Y., Ruan, Y., et al. (2019). Assessment of three essential tremor genetic loci in sporadic Parkinson's disease in Eastern China. *CNS Neurosci. Ther.* 26, 448–452. doi: 10.1111/cns.13272

Hor, H., Francescatto, L., Bartesaghi, L., Ortega-Cubero, S., Kousi, M., Lorenzo-Betancor, O., et al. (2015). Missense mutations in TENM4, a regulator of axon guidance and central myelination, cause essential tremor. *Hum. Mol. Genet.* 24, 5677–5686. doi: 10.1093/hmg/ddv281

Houle, G., Schmouth, J. F., Leblond, C. S., Ambalavanan, A., Spiegelman, D., Laurent, S. B., et al. (2017). Teneurin transmembrane protein 4 is not a cause for essential tremor in a Canadian population. *Mov. Disord* 32, 292–295. doi: 10.1002/mds.26753

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., Mcdonnell, S. K., Baheti, S., et al. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, 877–885.

Jankovic, J. (1989). Essential tremor and Parkinson's disease. *Ann. Neurol.* 25, 211–212.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

Kun-Rodrigues, C., Ganos, C., Guerreiro, R., Schneider, S. A., Schulte, C., Lesage, S., et al. (2015). A systematic screening to identify de novo mutations causing sporadic early-onset Parkinson's disease. *Hum. Mol. Genet.* 24, 6711–6720. doi: 10.1093/hmg/ddv376

Lees, A. J., Hardy, J., and Revesz, T. (2009). Parkinson's disease. *Lancet* 373, 2055–2066.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Louis, E. D., and Ottman, R. (2006). Study of possible factors associated with age of onset in essential tremor. *Mov. Disord* 21, 1980–1986. doi: 10.1002/mds.21102

Lunati, A., Lesage, S., and Brice, A. (2018). The genetic landscape of Parkinson's disease. *Rev. Neurol. (Paris)* 174, 628–643.

Mencacci, N. E., Isaias, I. U., Reich, M. M., Ganos, C., Plagnol, V., Polke, J. M., et al. (2014). Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. *Brain* 137, 2480–2492.

Nakamura, H., Cook, R. N., and Justice, M. J. (2013). Mouse Tenm4 is required for mesoderm induction. *BMC Dev. Biol.* 13:9. doi: 10.1186/1471-213X-13-9

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Nicolae, D. L. (2016). Association tests for rare variants. *Annu. Rev. Genom. Hum. Genet.* 17, 117–130.

Pellecchia, M. T., Varrone, A., Annesi, G., Amboni, M., Cicarelli, G., Sansone, V., et al. (2007). Parkinsonism and essential tremor in a family with pseudo-dominant inheritance of PARK2: an FP-CIT SPECT study. *Mov. Disord* 22, 559–563. doi: 10.1002/mds.21262

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., et al. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord* 30, 1591–1601. doi: 10.1002/mds.26424

Spanaki, C., and Plaitakis, A. (2009). Essential tremor in Parkinson's disease kindreds from a population of similar genetic background. *Mov. Disord* 24, 1662–1668. doi: 10.1002/mds.22655

Spataro, N., Calafell, F., Cervera-Carles, L., Casals, F., Pagonabarraga, J., Pascual-Sedano, B., et al. (2015). Mendelian genes for Parkinson's disease contribute to the sporadic forms of the disease. *Hum. Mol. Genet.* 24, 2023–2034. doi: 10.1093/hmg/ddu616

Tarakad, A., and Jankovic, J. (2018). Essential tremor and Parkinson's disease: exploring the relationship. *Tremor Other Hyperkinet Mov (N Y)* 8:589. doi: 10.5334/tohm.441

Unal Gulsuner, H., Gulsuner, S., Mercan, F. N., Onat, O. E., Walsh, T., Shahin, H., et al. (2014). Mitochondrial serine protease HTRA2 p.G399S in a kindred with essential tremor and Parkinson disease. *Proc. Natl. Acad. Sci. U S A.* 111, 18285–18290. doi: 10.1073/pnas.1419581111

Xue, C. B., Xu, Z. H., Zhu, J., Wu, Y., Zhuang, X. H., Chen, Q. L., et al. (2018). Exome sequencing identifies TENM4 as a novel candidate gene for schizophrenia in the SCZD2 locus at 11q14-21. *Front. Genet.* 9:725. doi: 10.3389/fgene.2018.00725

Yan, J., Deng, H. X., Siddique, N., Fecto, F., Chen, W., Yang, Y., et al. (2010). Frameshift and novel mutations in FUS in familial amyotrophic lateral sclerosis and ALS/dementia. *Neurology* 75, 807–814. doi: 10.1212/wnl.0b013e3181f07e0c

Yan, Y. P., Xu, C. Y., Gu, L. Y., Zhang, B., Shen, T., Gao, T., et al. (2020). Genetic testing of FUS, HTRA2, and TENM4 genes in Chinese patients with essential tremor. *CNS Neurosci. Ther.* 26, 837–841. doi: 10.1111/cns.13305

Yang, N., Zhao, Y., Liu, Z., Zhang, R., He, Y., Zhou, Y., et al. (2019). Systematically analyzing rare variants of autosomal-dominant genes for sporadic Parkinson's disease in a Chinese cohort. *Neurobiol. Aging* 76, 215.e1–215.e7.

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

MINI REVIEW
published: 10 February 2021
doi: 10.3389/fgene.2021.624820

Deep Learning in Head and Neck Tumor Multiomics Diagnosis and Analysis: Review of the Literature

Xi Wang[1,2] *and Bin-bin Li*[1,2]*

[1] Department of Oral Pathology, Peking University School and Hospital of Stomatology & National Clinical Research Center for Oral Diseases & National Engineering Laboratory for Digital and Material Technology of Stomatology & Beijing Key Laboratory of Digital Stomatology, Beijing, China, [2] Research Unit of Precision Pathologic Diagnosis in Tumors of the Oral and Maxillofacial Regions, Chinese Academy of Medical Sciences, Beijing, China

Head and neck tumors are the sixth most common neoplasms. Multiomics integrates multiple dimensions of clinical, pathologic, radiological, and biological data and has the potential for tumor diagnosis and analysis. Deep learning (DL), a type of artificial intelligence (AI), is applied in medical image analysis. Among the DL techniques, the convolution neural network (CNN) is used for image segmentation, detection, and classification and in computer-aided diagnosis. Here, we reviewed multiomics image analysis of head and neck tumors using CNN and other DL neural networks. We also evaluated its application in early tumor detection, classification, prognosis/metastasis prediction, and the signing out of the reports. Finally, we highlighted the challenges and potential of these techniques.

Keywords: artificial intelligence, deep learning, head and neck tumors, diagnosis, multi-omics

OPEN ACCESS

Edited by:
Ying Li,
Zhejiang University, China

Reviewed by:
Yuqi Zhao,
Jackson Laboratory, United States
Khanh N. Q. Le,
Taipei Medical University, Taiwan

****Correspondence:***
Bin-bin Li
kqlibinbin@bjmu.edu.cn

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

Received: *01 November 2020*
Accepted: *07 January 2021*
Published: *10 February 2021*

Citation:
Wang X and Li B-b (2021) Deep Learning in Head and Neck Tumor Multiomics Diagnosis and Analysis: Review of the Literature. Front. Genet. 12:624820. doi: 10.3389/fgene.2021.624820

INTRODUCTION

Head and neck tumors are the sixth most common neoplasms (529,000 new cases annually) and cause 350,000 cancer-related deaths each year (Ferlay et al., 2015; Fidler et al., 2017). Accurate diagnosis and analysis, especially histologic, radiologic, and biological findings, are crucial for therapeutic efficacy and prognosis prediction in precision medicine. A histologic section typically contains 10^6–10^7 cells and provides information on cell numbers and the tumor microenvironment (Koelzer et al., 2017). Radiological images contain 50–5,000 quantitative features (Limkin et al., 2017). Therefore, pathologists and radiologists must spend considerable time and effort on the qualitative and quantitative analyses of cell subsets and biomarker expression in a series of images. Also, inter- and intraobserver variations caused by subjective evaluation are inevitable in clinical practice.

Artificial intelligence (AI) was developed in the 1950s (Bini, 2018). The term *big data* was first proposed by the National Aeronautics and Space Administration in 1997 because a dataset is too large to be easily manipulated and managed. Big data refers to extra huge amounts of data integration, storage, analysis, and reuse of various forms of data, such as audio, video, and images. Big data is aimed at generating a large amount of information to assist decision-making and estimate outcomes, at a lower cost in time and labor (Conway et al., 2018). Computer algorithms and well-integrated data are critical for decoding medical big data. Because radiologic images are digitalized, no additional processing is required. Hung et al. (2020) used clinical big data from the SEER database to predict the survival time of patients of oral tumor by machine learning algorithms in 2020. For pathologic diagnosis, the first major step in adopting deep learning (DL) is to use digital whole-slide imaging (WSI) in routine practice (Jeyaraj and Samuel Nadar, 2019). WSI is non-inferior to traditional microscopy for clinical diagnosis (Halicek et al., 2019).

Machine learning (ML), a type of AI, refers to a computer software performing a task by being exposed to the manually crafted features of representative data (Niel and Bastard, 2019). Head and neck tumors are diverse in histology, in the pattern of underlying genetic alterations, and in metabolic signatures, which need a new method to reveal the sophisticated features. An evolution of ML—DL (Helm et al., 2020)—was first applied to the analysis of pathologic images of the head and neck in 2017 (Lu et al., 2017). Several new theories and methods have arisen to facilitate the application of DL in precision medicine, such as backpropagation and multiple layers in the convolution network. The main beauty of DL is to get rid of the handcrafted features and the end-to-end learning procedure. In the same year, DL was applied to radiomics image segmentation of head and neck tumors (Ibragimov and Xing, 2017). As a result of the improvements of computer algorithms and computational pathology, DL now facilitates the identification of benign and malignant tumors, grading of malignant tumors, and prognosis prediction.

Here, we outlined the application of DL algorithms in multiomics to diagnose and analyze head and neck tumors. Because pathological diagnosis of tumors is the gold standard, the application of DL in pathomics is emphasized in the diagnosis section and radiomics in the prognosis section. Finally, we review the challenges and prospects of DL in multiomics diagnosis and analysis.

APPLICATION OF DL IN TUMOR DIAGNOSIS AND MULTIOMICS ANALYSIS

The term "multiomics" in medicine refers to the combination of multiple sources of information (genomics, transcriptomics, proteomics, metabolomics, radiomics, and pathomics) to provide a deeper understanding of the tumor pathogenesis and lesion nature (Mars et al., 2020; Ye Y. et al., 2020; **Figure 1**). A schematic representation of a synergetic integration of multiomics data is shown in **Figure 1**. DL techniques have already been applied for multiomics analysis in various tumors. The identification of tumor origin and essential gene is critical for molecular targeted therapies and accurate treatment and lays the foundation to reveal changes in oncogenic mutation by liquid biopsy. Actually, multiomics is heterogeneous data which is difficult to be comprehensively analyzed. However, the DL network takes this challenge into an opportunity. DL-based multiomics analysis has allowed to classify groups of patients based on a more individual scale in the era of precision medicine. A timeline demonstrating the researches of DL in tumor diagnosis and multiomics analysis is shown in **Supplementary Figure 1**. Identifying robust survival subgroups of head and neck squamous cell carcinoma (HNSCC) will significantly improve patient outcome. Zhao et al. (2020) established a DL-based disease progression model on 86 HNSCC patients' data using methylation data, RNA sequencing (RNA-Seq), and miRNA sequencing (miRNA-Seq) from The Cancer Genome Atlas (TCGA). The results of the autoencoder DL model demonstrated that patients were classified into two subgroups with a significant difference in progression-free survival (PFS). The predictability of this model was validated using three independent cohorts. The different biological origin of the tumor tissue has distinct clinical behavior. In practical clinical situations, it is difficult to distinguish between poorly differentiated carcinoma and metastatic carcinomas. Jiao et al. (2020) constructed a multiclass deep learning/neural network (DNN) model to integrate the whole genome sequence and pathomics data to shed light on a comprehensive view of the histological origin of the tumor cells. They evaluated three features, namely mutation distribution, mutation type, and driver gene/pathway. The classifier achieved predictive accuracies of 91% in 24 types of tumors.

Artificial neural network models have been used to investigate the relationship between the symptoms of oral cancer and its prognosis (Tseng et al., 2015). Phillips et al. (2019) used DL models to detect pigmented dermoscopic images, thus improving the accuracy of early melanoma diagnosis. Clinically, it is difficult to differentiate ameloblastomas from keratocystic odontogenic tumors depending only on X-ray. CNN can assist in the diagnosis of ameloblastoma and keratocystic odontogenic tumors based on transfer learning. The sensitivity, specificity, and accuracy were 81.8, 83.3, and 83.0%, respectively. Interestingly, the model performed consistently well, just like skilled experts (Poedjiastoeti and Suebnukarn, 2018). In addition to X-ray research, Fu et al. (2020) used clinical photographic images to predict the early occurrence of oral tumor through a cascaded CNN model. After training by 1,469 samples, the sensitivity and specificity reached 94.9 and 88.7%, separately. This study also provided a non-invasive and highly efficient perspective on oral tumor detection. It was also possible to start providing early treatment immediately. In IDH1 wild-type glioblastomas, methylation modification had a great influence on chemotherapy response and prognosis. Le et al. (2020) used a radiomics-based eXtreme Gradient Boosting (XGBoost) model to predict the IDH1 wild-type patients with O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation status. Nine robust radiomics features were selected based on the *F* score to improve the diagnosis of MGMT methylation status in IDH1 wild-type glioblastomas and predict patient prognosis. Sulfation of the protein S site is an important posttranscriptional modification, which plays a vital role in signal transduction, transcriptional regulation, and cell apoptosis. However, traditional experiments for its biological functions were not timely due to its rapid degradation. Do et al. (2020) used a DL network to predict protein phosphorylation S site based on the proteomics data. The DL network was also used to predict the function of fertility-related proteins in infertility patients and paved the way for a better understanding of the function of fertility proteins (Le, 2019). Therefore, the integration of DL, image analysis, and big data enables the evaluation of tumor

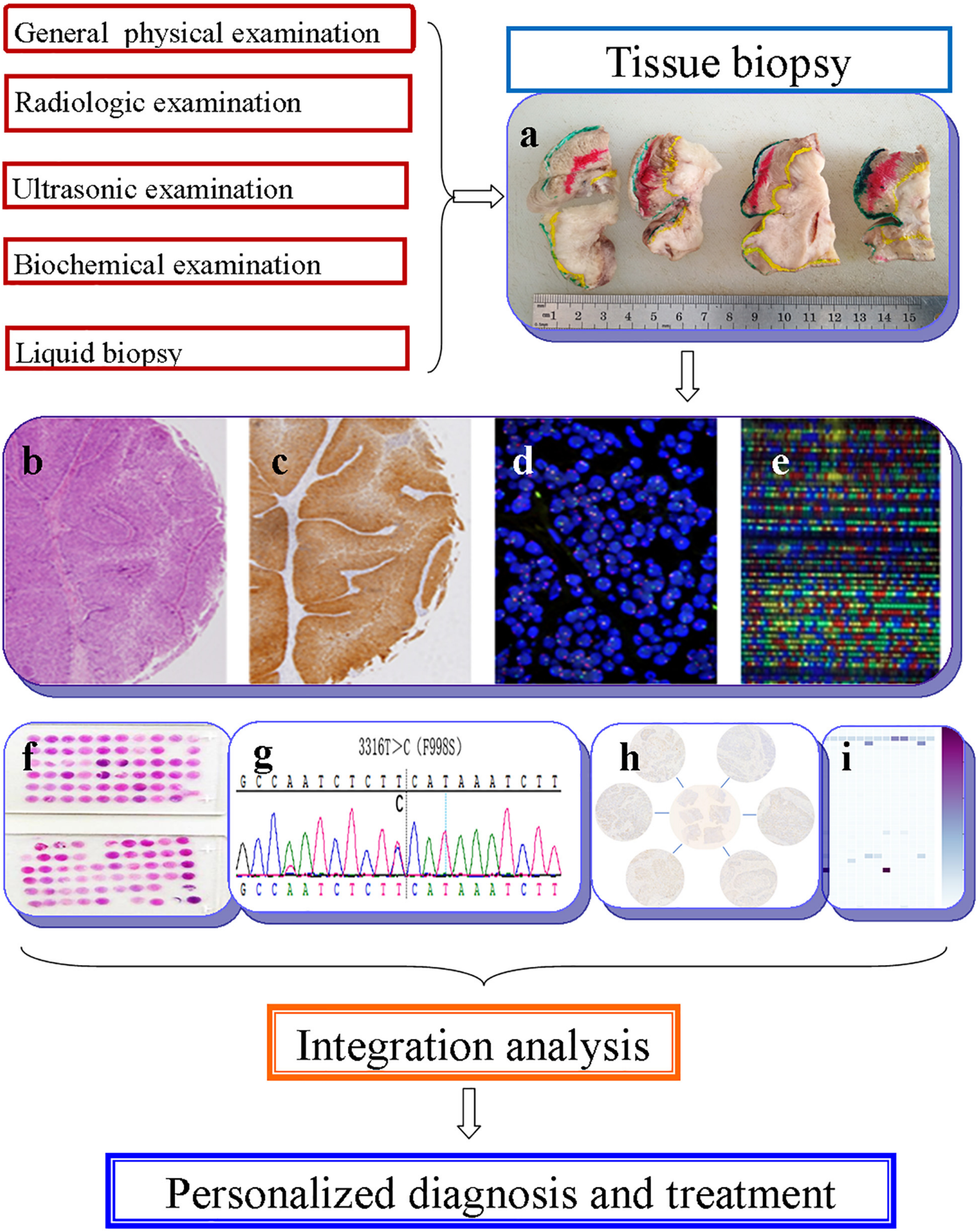

General physical examination
Radiologic examination
Ultrasonic examination
Biochemical examination
Liquid biopsy
Tissue biopsy
a
b
c
d
e
f
g
3316T>C (F998S)
G C C A A T C T C T T C A T A A A T C T T
C
G C C A A T C T C T T C A T A A A T C T T
h
i
Integration analysis
Personalized diagnosis and treatment

FIGURE 1 | Schematic representation of a synergetic integration of multiomics data. **(a)** Biopsy specimens from head and neck tumor tissue. **(b)** Pathological data. **(c)** Tumor protein data. **(d)** Chromosomal data. **(e)** Gene microarray data. **(f)** Tissue microarray data. **(g)** Gene mutation data. **(h)** mRNA expression data. **(i)** Methylation data.

biological behavior and, hence, facilitates diagnosis, personalized treatment, and survival prediction.

HEAD AND NECK TUMOR MULTIOMICS ANALYSIS

Multiomics Analysis in Early Detection of Tumors

The global incidence of head and neck cancer is 1.3 million annually. The risk factors for head and neck tumors are chewing tobacco, local irritation, smoking, alcohol abuse, human papillomavirus infection, etc. It is necessary to monitor the occurrence of oral cancer in high-risk groups. Early diagnosis could reduce the mortality rate to 70% at present (Erickson et al., 2018). Also, DL could enable regular follow-up of high-risk groups. Moreover, DL methods can be applied not only to low-level tasks (e.g., recognition, detection, and segmentation) but also to more advanced tasks (e.g., selection of the optimal treatment and prediction of prognosis).

As we know, routine tissue biopsies are invasive. Although it is safe, some risks may be brought in rare cases and non-invasive biopsy comes into being. In non-invasive modalities, a large number of images appeared combined with training of DL networks based on oral clinical examinations and histological findings, which would assist in the evaluation of precancerous and cancerous lesions. The human eyes and cameras capture three color channels—red, green, and blue. Hyperspectral imaging involves multiple wavelengths, enabling the identification of cancerous and normal tissue by optical biopsy. Halicek et al. (2017) trained a CNN to identify hyperspectral images of squamous cell carcinoma (SCC). The reported accuracy, sensitivity, and specificity of the training set were 81, 81, and 80%, respectively. The hypercube contained 91 spectral bands, ranging from 450 to 900 nm with a 5-nm spectral sampling interval. Similarly, confocal laser endomicroscopy (CLE) allows real-time visualization of epithelium *in vivo* and enables early diagnosis of oral cancer and prediction of the prognosis. In 2007, Soo et al. reported the application of CLE for the diagnosis of oral SCC (OSCC) (Thong et al., 2007). Subsequently, Nathan et al. applied CLE to detect head and neck precancerous lesions; the sensitivity and specificity for the diagnosis of oral epithelial dysplasia were 85.7 and 80.0%, respectively (Moore et al., 2016). Aubreville et al. (2017) proposed an automatic framework for the application of CLE to detect cancerous lesions by CNN. In the proteomics research of head and neck tumors, Ni et al. (2015) used artificial neural networks to screen out proteins which were related to lymph node metastasis using the proteins extracted from the saliva of OSCC patients.

Radiomics is also used as one of the non-invasive clinical examinations. Ren et al. (2018) used the least absolute shrinkage and selection operator (LASSO) logistic regression to extract features from magnetic resonance images (MRI) of head and neck SCC to predict the histological grade before surgery. Subsequently, the same method was applied in floor-of-the-mouth and tongue SCC by Ren et al. (2020). Computed tomography (CT) can also be used to predict the histological classification before surgery by kernel principal component analysis (KPCA) and the random forest classifier (Wu et al., 2019). Mukherjee et al. (2020) performed principal component analysis and regularized regression to predict tumor grade, extracapsular spread, perineural invasion, lymphovascular invasion, and human papillomavirus infection status. The accuracy, sensitivity, and specificity of the model were 0.72, 0.83, and 0.48, respectively. DL is also applied in radiomics. Ye J. et al. (2020) used a CNN model for histological classification of head and neck tumors; the accuracy, sensitivity, and specificity were 0.79, 0.71, and 0.85, respectively. The utility of AI for the analysis of head and neck pathologic sections and radiologic images is summarized in **Tables 1**, **2**.

Multiomics Analysis in Tumor Detection, Segmentation, and Classification

Deep learning is suitable for digital pathology (DP)-related image analysis tasks, such as detection (e.g., lymphocyte), segmentation (e.g., nuclei and epithelium), and classification (e.g., the tumor subclass). **Figures 2**, **3** demonstrate an example of epithelial segmentations on WSI images and an example of segmentation of nuclei in a cell layer on WSI images. Different from ML, which classifies handcrafted features (Das et al., 2018), DL takes an agnostic approach by combining feature extraction and the interest region analysis.

In head and neck tumor diagnosis, the morphology of heterogeneous cell types needs to be evaluated. This can be formulated as a pixel-wise detection task. The detection tasks frequently align with the classification tasks, and the algorithms learn the weighted parameters of the feature map. The algorithms map clusters of similar features to the output labels. The workflow for DL approaches in digital pathology is shown in **Figure 4**. In traditional ML, the workflow is comprised of two steps: detection and classification. For instance, Lewis et al. (2014) developed an approach to quantify automatically the morphologic features used for the classification of aggressive or indolent p16-positive oropharyngeal SCC. A cluster cell graph was generated to evaluate the spatial distribution of mitotic cells, and a random forest (RF) decision tree and SVM were used to classify features. The accuracy of the model was 87.5% (140 patients). However, it may not be applicable to other situations because of the small training dataset and the overfitting problem. Moreover, the accuracy of DL is unsatisfactory. Several proposed DL models for detecting head and neck tumors overcome the abovementioned shortcomings. Aubreville et al. (2017) trained a DL model to detect an image patch from doubtful OSCC cases. Overall image recognition had an area under the curve (AUC) of 0.96 and a mean accuracy of 88.3% (sensitivity 86.6% and specificity 90%).

Halicek et al. trained a deep CNN to identify surgical margins accurately in hyperspectral images. Additionally, an end-to-end DL network can simultaneously detect and enumerate mitotic cells (Jimenez and Racoceanu, 2019). In the above reports, DL was only used for dimension reduction or feature extraction. It also may be a classifier to perform classification (Boldrini et al., 2019). Usually, the end-to-end DL approach

TABLE 1 | Summary of deep learning models for H&N tumor Pathomics analysis.

Topic	H&N tumor subtype	Task	Model	References
H&N tumor detection and classification	H&N squamous cell carcinoma (HNSCC)& thyroid carcinoma	Malignant vs. non-malignant classification	CNN	Witjes et al., 2018
	OSCC	According to the keratin pearl to classify the high-grade or low-grade OSCC	CNN	Das et al., 2018
	OSCC	Malignant vs. non-malignant classification in CLE image	CNN	Aubreville et al., 2017
	Oral tumor	Malignant vs. benign vs. precancerous classification	CNN	Jeyaraj and Samuel Nadar, 2019
H&N tumor segmentation	OSCC	Tumor margin detection and segmentation	CNN	Halicek et al., 2018
	OSCC	Segmentation the boundary of tumor and normal tissue	CNN	Halicek et al., 2017
	TSCC	Tumor margin detection and segmentation	CNN	Yu et al., 2019
	OSCC	Quantity nuclear morphology to stratify patients of high or low risk	CNN	Lu et al., 2017
	OSCC	Based on clinic-hiotopathology features to predict patient's outcome	DL	Kim et al., 2019
	OSCC	Quantity tumor infiltrating lymphocytes to predict the patients' outcome and treatment response	CNN	Shaban et al., 2019

H&N, head and neck; OSCC, oral squamous cell carcinoma; CNN, convolution neural network; CLE, confocal laser endomicroscopy; DL, deep learning.

TABLE 2 | Summary of machine learning and deep learning models for H&N tumor Radiomics analysis.

Topic	H&N tumor subtype	Task	Model	References
H&N tumor prognosis	H&N squamous cell carcinoma (HNSCC)	Loco-regional control (LRC)	PCA	Bogowicz et al., 2019b
	head and neck cancer (HNC)		Z-Rad radiomics software	Bogowicz et al., 2019a
	Locally advanced head and neck cancer		Free LifeX software package	Cozzi et al., 2019
	HNSCC	Overall survival (OS)	RadiomiX Discovery Toolbox.	Keek et al., 2020
			In-house built Accurate tool	Martens et al., 2020
			LASSO	Yuan et al., 2019
			PCA	Mes et al., 2020
	H&N tumor		Random survival forests (RSF) and random forest (RF)	Leger et al., 2019
			Velocity AI v3.0.1 software and Imaging Biomarker Explorer and k-medians	Tosado et al., 2020
			Matlab R2018b	Lv et al., 2020
			Z-Rad software and Hierarchical Clustering	Bogowicz et al., 2020
			IBEX, an open-source radiomics tool	Ger et al., 2019
	Aryngeal squamous cell carcinoma		LASSO	Chen L. et al., 2020
Biologic markers prediction	Oropharyngeal squamous cell carcinoma	HPV status prediction	In-house developed software, using Matlab 2014a	Leijenaar et al., 2018
	Oropharyngeal cancers	HPV status prediction	PCA	Bagher-Ebadian et al., 2020
	HNSCC	HPV status and T-cell infiltration prediction	Unsupervised consensus clustering and PCA	Katsoulakis et al., 2020
H&N tumor recurrence and metastasis	HPV-related Oropharyngeal Carcinoma	Distant metastasis	Unpublished MATLAB code	Kwan et al., 2018
	H&N tumor	Metastatic lymph nodes	Naive Bayes, and k-nearest neighbor classifiers	Tran et al., 2019
	Locally advanced head and neck cancer	Recurrence	Random forest	Beaumont et al., 2019
	H&N tumor	Lymph node metastasis	3-dimensional CNN	Zhou et al., 2018; Chen et al., 2019
	Papillary thyroid carcinoma		SVM	Liu et al., 2019
	H&N tumor		Matlab	Zhai et al., 2020

H&N, head and neck; PCA, principal component analysis; LASSO, the least absolute shrinkage and selection operator; SVM, support vector machine.

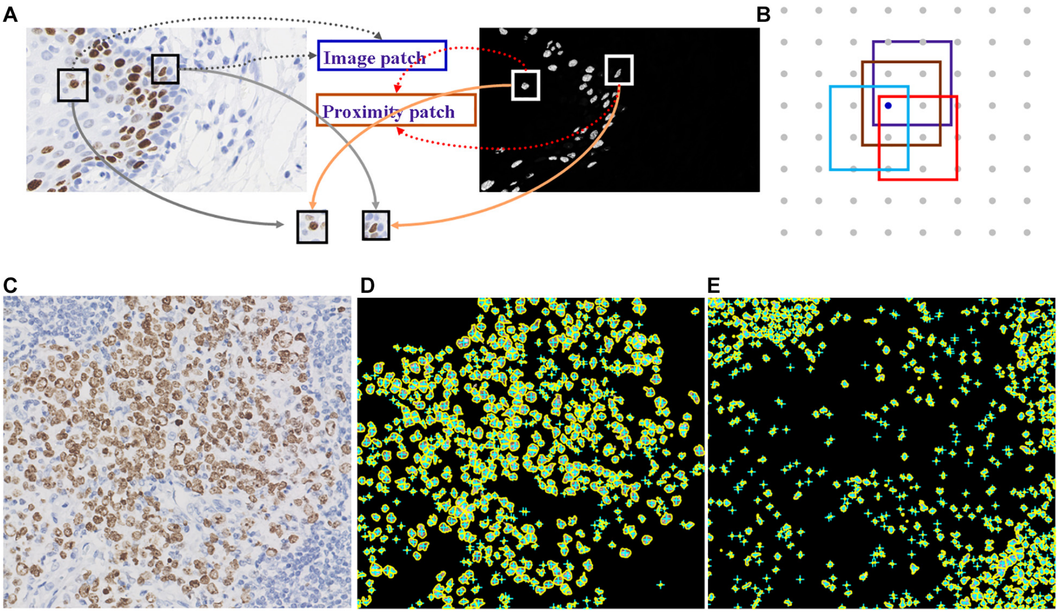


FIGURE 2 | Example of segmentation of nuclei in a cell layer on WSI images. **(A,B)** Illustrations of nucleus segmentation based on the DL model. **(C)** The original image of immunohistochemical staining. **(D)** Nucleus segmentation for immunohistochemistry-positive cells. **(E)** Nucleus segmentation for immunohistochemistry-negative cells.

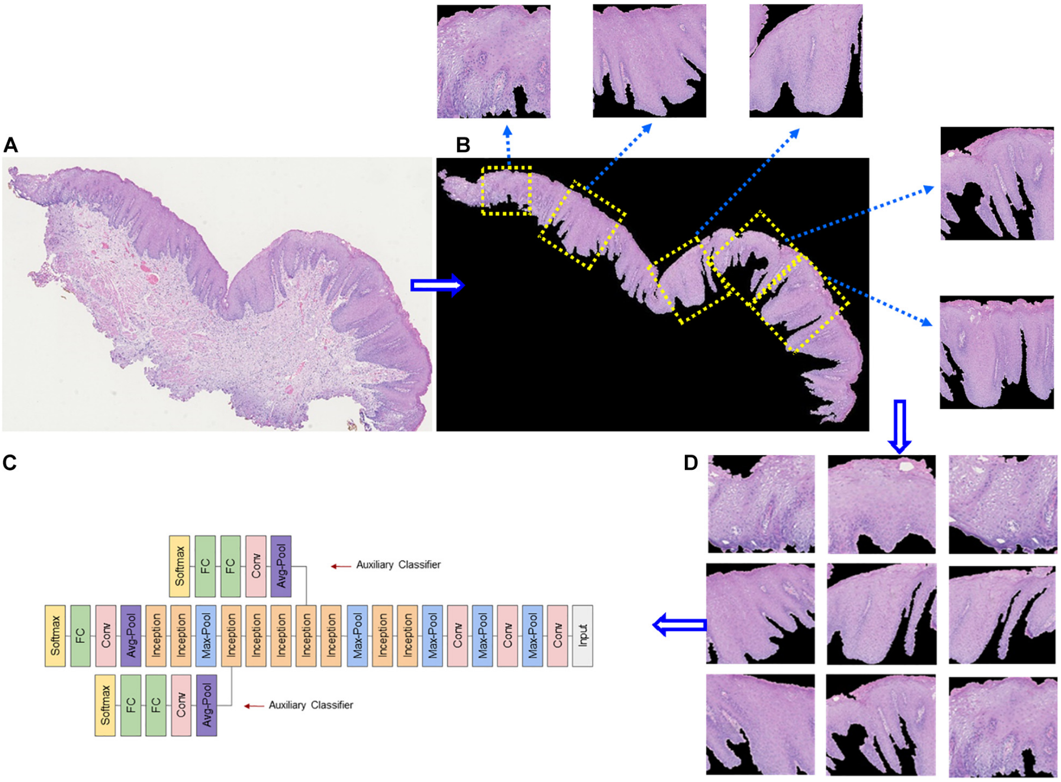


FIGURE 3 | Example of epithelial segmentations on WSI images. **(A)** The original image. **(B)** Black curves indicate the segmented boundary of the epithelium. **(C)** Input patches for training the DL network. **(D)** Schematic of a deep learning framework.

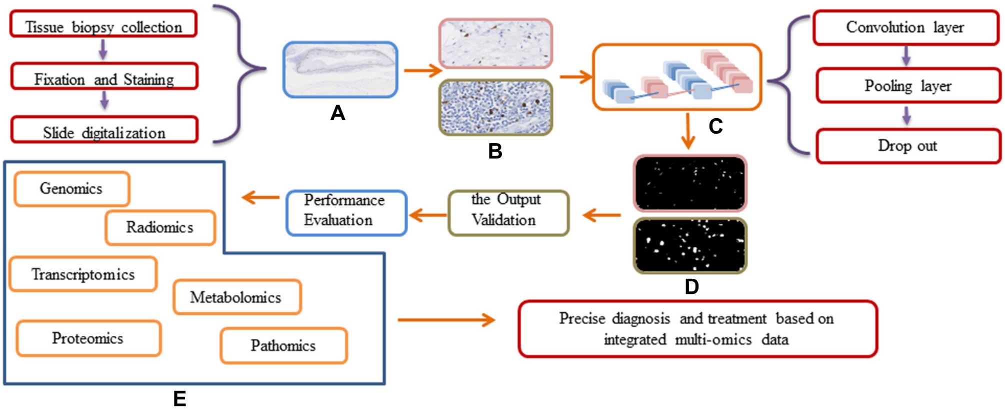


FIGURE 4 | Workflow and general framework for DL approaches in digital pathology. **(A)** The original WSI image. **(B)** Different resolution of WSI images before input the DL network. **(C)** The Schematic of a deep learning framework. **(D)** The output of different resolution of WSI images. **(E)** The integration of multi-omics data.

performed better than none end-to-end learning. However, as for pathomics, one end-to-end DL model cannot perform multiple tasks simultaneously. Lei et al. (2019) trained a convolutional neural network by DL to extract mitosis features automatically and proposed a network to determine the location of all mitotic cells. This approach showed an unexpectedly high accuracy in the International Conference on Pattern Recognition (ICPR2012) mitosis detection test dataset. The remaining challenges include accurate identification and enumeration of mitotic cells in two-dimensional (2D) digital histology images. The imaging of three-dimensional (3D) tissues in 2D results in loss of spatial information. Radiomics enables enhanced 3D assessment of tumor growth by quantifying changes in tumor cellularity and angiogenesis. Radiomics analysis also shows potential for the accurate quantification of heterogeneity and outcome prediction.

Microscopic cell structure recognition is emphasized in pathomics. A common strategy for detecting cells or nuclei is to train a CNN classifier as a pixel classifier, in which a patch centered on the object of interest is used to train the network under supervised conditions. Trained CNN models typically comprise two classifiers (yes or no) and can be applied to WSI in a sliding window to detect all histological components of interest and output a probability map, where each pixel is transferred to a probable value. Therefore, in principle, the target objects can be located by finding a local maximum in the generated probability map. Fully convolutional networks can share calculations on sliding windows. After completing nuclear or mitotic detection tasks, it begins counting or extracting quantitative indicators in WSI. The algorithm is built on mapping an input image patch to a density map, which is used to estimate the number of cells in the original image.

Deep learning also plays an important role in the analysis of tumor microenvironment characteristics (TMC). The crucial step in TMC analysis is segmenting different types of tissue and cell structures in pathology images. Tumor cells can be classified into parenchymal and stromal cells. Niranjan and Sarathy (2018) reported the ratio of tumor to stroma (TSR) as a reliable histologic predictor of overall survival and outcome in OSCC. In a cohort of 60 OSCC patients, the 3-year overall survival (OS) and disease-free survival (DFS) rates of patients with >50% intratumor stroma had been shown to be better than the patients with <50% intratumor stroma.

The segmentation task is more difficult than mitosis detection because parenchyma segmentation can be labeled by experts at lower magnifications. However, stroma (e.g., lymphocytes, macrophages, fibroblasts, etc.) must be analyzed at high magnification. Indeed, × 40 magnification performed better than × 20 magnification for nucleus segmentation. By contrast, epithelium segmentation is typically more precise by experts at × 20 than at 40 × magnification, as indicated by a higher accuracy and *F* score (Janowczyk and Madabhushi, 2016). To remedy this drawback, the fully convolutional network (FCN) and UNet were designed to accept discretional size as an input and product proportionate-sized outputs by removing all fully connected layers and introducing unsampled layers to offset the shortcomings of downsampling in CNN (Zhou et al., 2019). Considering that head and neck tumors are heterogeneous and complex, segmentation may involve varied anonymous anaplastic cells and then can be achieved by data augmentation. Halicek et al. (2018) trained a CNN to segment the tumor and normal tissue of OSCC with 81% accuracy, 84% sensitivity, and 77% specificity. The sensitivity and specificity of FCN for cervical tumor segmentation on 3D FDG-PET images were 88 and 98%, which were markedly superior to CNN. Unfortunately, FCN has not been used for segmentation of pathologic images of head and neck tumors. Moreover, tumor segmentation accuracy is associated with loss function. Now, the well-known loss function is cross-entropy loss. A new loss function, class-wise DSC loss, for training the segmentation network of colonoscopy pathology images was presented by Feng et al. (2020).

Multiomics Analysis in Tumor Prognosis and Metastasis

The high heterogeneity and complexity of head and neck tumor pathology images hamper the prediction of outcomes only by

TNM stage. In recent years, more and more scholars have been interested in the potential of DL networks for predicting postoperative outcomes. The applications of radiomics to predict overall survival, biomarker status, recurrence, distant or local metastasis, and lymph node metastasis are summarized in **Table 2**. Tixier et al. used the Genomica software to analyze PET and transcriptomics data of 45 patients with locally advanced head and neck cancer. They applied a fuzzy locally adaptive Bayesian (FLAB) algorithm to assess the associations between radiomics features (a total of 28 image biomarker standardization initiative-compliant radiomic features) (Zwanenburg et al., 2020) and alterations of biological pathways (e.g., extracellular matrix organization, cell cycle, signal transduction, cell cycle, etc.). The results demonstrated that FDG-PET radiomic features were associated with cell cycle, DNA repair, extracellular matrix organization, immune system, metabolism, and signal transduction pathways, providing a thorough understanding of genetic mutations and minimizing the costs (Tixier et al., 2020). Zhu et al. (2019) integrated the genome-wide multiomics data of 126 patients with head and neck SCC with CT imaging data and found the significant association between genomic characteristics and CT features. The use of DL together with sophisticated biomarkers can significantly improve prognostic and predictive accuracy. Subsequently, the DL-extracted imaging features of morphology structure on digitized H&E-stained tissue sections have been used for risk stratification of head and neck tumor patients. Patients with p16-positive human papillomavirus-related oropharyngeal SCC have a more favorable prognosis than those negative for P16 (Ali et al., 2013). Lewis et al. (2014) used a typical ML approach (the random forest decision tree) to extract nuclear morphologic features and predict progression. Before the advent of DL, improvement of prognosis was evaluated by multifactor analysis, conventional logistic regression, and Cox analysis in traditional ML models. However, the absence of a decision rule and linear combinations of covariates hampered the prediction of outcomes. DL-based survival prediction has improved predictive accuracy and, together with nonlinear algorithms, will facilitate precision medicine. Therefore, it is suitable for predicting the survival of inpatients (Tan et al., 2016). Tseng et al. constructed a DNN to predict the survival of patients with oral tumors using clinical variables and histopathological features. It was suggested that the DNN model established by data mining was superior to logistic regression in terms of both training accuracy and cross-validation accuracy. Brennan et al. (2017) used an unsupervised cluster analysis method to interpret the genomics and epigenetics data of morphologically atypical head and neck SCC and found CpG island methyl groups in atypical SCC. Therefore, novel prognostic factors, such as genetic mutations and molecular markers, combined with clinicopathologic and radiologic features and a multi-nonlinear DL network would yield optimal results.

Proteomics and transcriptome have also been used to study lymph node and distant metastasis and recurrence of SCC patients. Onken et al. (2014) used an unsupervised clustering algorithm to extract transcriptome signature predicting distant metastasis in oral tumor over four SCC datasets. Xu et al. (2014) applied a ML approach called maximum relevance minimum redundancy algorithm to a set of transcriptome data generated from papillary carcinoma and anaplastic carcinoma for differential diagnoses. The lung is the most common site of distant metastasis of OSCC. Primary SCC can also occur in the lung. Through supervised learning and analysis of proteomic data, Bohnenberger et al. (2018) found the vital difference of protein characteristics between lung metastatic head and neck SCC and primary lung SCC. Their data provided reference information for the origin of lung SCC. Carnielli et al. (2018) used histological morphology-oriented proteomics analysis of the protein expression in tumor islands and stroma to forecast the possibility of tumor recurrence and lymph node metastasis. Six ML approaches were used by Kaddi and Wang (2017) to analyze proteomics and transcriptome data, including KNN, SVM, naive Bayes, DT, AdaBoost, and RF. It was shown that the prognostic model based on both transcriptome and proteomics data had better predictive performance than transcriptomics or proteomics alone.

Diagnostic Reports: Automatic Extraction of Tumor Information

Zhang et al. (2017) developed MDNet, which generated pathological reports by directly mapping pathology images with simultaneous retrieval of pathology images according to symptom descriptions. MDNet added a language network to the original image model. Integration of a language model with the multiscale features proposed by the image model allowed the identification of critical image features and enabled the direct mapping from words to pixels. Changes in the size or density of nuclei and epithelial thickness may indicate neoplastic invasion. However, these discriminant imaging features were not directly supported to generate a diagnostic report. MDNet allowed direct multimodal mapping from medical images and diagnostic reports. Mimicking diagnosis by pathologists, long short-term memory (LSTM) networks were used to generate semantic information as a language model. The LSTM was a representative gated RNN that controlled the forget gate and input gate to emphasize or forget some weights. It could reduce the problem of multiple layers and vanished gradient multilayers from input to output.

Radiogenomics Analysis for Radiotherapy Patients

Radiogenomics is a computational nomenclature which identifies correlations between radiomics imaging features and genomics or proteomics data. These imaging feature correlations can be used to predict a tumor's molecular profile in clinical radiomics data (West and Rosenstein, 2010). Radiogenomics has two goals: i) discover the patients who are more likely to develop radiotherapy complications based on molecular data and (ii) analyze the targeted molecular pathway responsible for radiotoxicity in radiation-induced normal tissue (Kerns et al., 2014). Postoperative radiotherapy is an effective treatment for head and neck tumors. The existence of radiosensitivity and radioresistance may be related to genetic factors partially. The remaining differences between individuals were caused by differences in treatment (radiation dose), physical habits, and random factors (Rattay and Talbot, 2014). Werbrouck et al. (2009) reported that the DNA repair genes *XRCC3* and *Ku70* were connected to the intensity of dysphagia after radiotherapy in H&N tumor in 2009. For the study of postradiotherapy mucositis, Yang et al. (2020) sequenced and located the gene expression in 1,497 patients with postoperative radiotherapy. They found that 64 target genes were enriched in the process of telomerase regulation, which confirmed the importance of telomere function in the development of radiation-induced adverse reactions. The combination of PET-based spatial radiation features and sequencing data provided a new perspective for further revealing the spatial heterogeneity of tumors (Clasen et al., 2020). Furthermore, the predictive analysis of gene expression and cellular and molecular expression can be provided from a non-invasive point of view, based on the radiological characteristics and gene differential expression data of head and neck tumors obtained from the TCGA and TCIA databases (Katsoulakis et al., 2020).

DIFFICULTIES AND EXPECTATION

AI is highly dependent on a robust and large database, but the database of pathological slides of head and neck tumors has not been established yet. Apart from the hardware needed to set up the database, setting up an autoprocessed image database is also needed. When there were images captured from clinical cases, the database could have the images with their properties at the same time, which would help in further analysis. As time goes on, the database could grow by itself (Ibrahim et al., 2020). The low-quality images are also a problem for DL analysis. According to a jointed framework proposed by Chen J. et al. (2020), a novel transfer learning strategy called channel fusion transfer learning and a deep super-resolution framework called SRFBN+ were dedicated to generating higher-resolution slice images with lower-resolution ones as input. The most successful application of DL in medical image analysis has been in supervised learning. On the other hand, the rarity of pathologists added the extra difficulties in data cleaning and labeling, while the high heterogeneity of head and neck tumors means that many rare tumors need to be accurately labeled.

A crucial step is to avoid subjective and sample biases in the training sets as the quality of the output depends on the quality of the input data (Oakden-Rayner, 2020). So, establishing a unified standard to normalize the image input in the network by multi-institution datasets can not only reduce the bias from the samples and the bias caused by inconsistent diagnostic from the physicians but also fully fit and train the model to reduce overfitting and reduce to a maximum the highly opaque nature of medical image (Martorell-Marugán et al., 2019). However, current DL algorithms are mainly trained on a small dataset from a single center (Jiang et al., 2020). The limited availability of well-characterized and adequately stored clinical tumor and non-tumor samples is a major challenge in proteomics and genomics researches (Matta et al., 2010).

For the algorithms themselves, the tendency has been to propose new algorithms rather than optimize those already used, leading to the conclusion that there is no improvement of some subdomain algorithms. In addition, due to the limitations of, for instance, data and computational power, the improvement of algorithms must take into account various trade-offs. Additionally, some studies used a non-open-source code or a non-open-source model, such as an in-house developed model, hampering model verification in other types of tumors (Parmar et al., 2015; Leijenaar et al., 2018). A flowchart demonstrating the relationship for the subsection of difficulties and expectation of DL in tumor diagnosis and multiomics analysis is shown in **Supplementary Figure 2**.

Difficulties Related to Unified Evaluation Standards

The lack of unified innovation evaluation standards in AI has led to some exaggeration of the improvements achieved. This can be overcome by a variety of methods, e.g., an open-source or source model. Unifying evaluation standards is difficult but is possible for some mature domains. The relevant data management domains are as follows: (i) administrative standards, (ii) patient privacy protection standards, and (iii) intellectual property protection standards. The establishment of data management standards would allow access to diverse anonymized imaging datasets. Technical standardization cannot resolve all of the issues described in this review. The use of different image normalization or style conversion methods (e.g., rotating, cropping, zooming, and image histogram-based modifications) for preprocessing could overcome the technical obstacles.

Difficulties in Image Analysis

The architectures of CNNs have been especially powerful for computer vision, particularly in image interpretation and procession. WSI combined with DL algorithms for tumor detection, classification, and prognosis prediction has played an ever-increasing part in supporting pathologists in clinical assessments. The main components of CNN are convolutional layers and pooling layers. Although CNN has advantages in the processing of object detection, it has notable drawbacks: (1) both the training and the detection process is considerably time-consuming and (2) the normalization method would lead to lose some discriminative details. FCN is suitable for image segmentation at the pixel level. It consists of convolution and deconvolution layers, which can accept input images of any size and retain the spatial information of the original input lines. The major disadvantages of FCN may be that it is noisy and contains redundant information, requiring a huge number of reliable samples. To overcome the issues mentioned above, more novel architectures (e.g., UNet++, SegNet, and ENet) based on FCN or CNN have been proposed for image segmentation. Pan et al. (2020) proposed a DL model based on the architectures of FCN to automatically recognize lymph node metastasis of esophageal SCC. Compared with previous studies focused on the isolated tasks in the analysis of pathology and radiology images, the integration of independent DL models into a general model would be beneficial (Wang et al., 2019). It was also anticipated that biological pathways and gene regulation networks would be incorporated into prediction models, improving their performance and interpretability. For multimodal learning, collecting data from the required modalities simultaneously could be problematic. A slight disturbance to the inputs of multimodal can influence the stability of CNN. Lin et al. (2020) trained a multiscale activity transition network to provide an activity state pyramid consisting of multiscale recurrent neural networks to capture the accurate feature of input. Transfer learning is frequently used and is an effective pretraining strategy. The fusion of different modal representations is the key point of a multimodal task. Specific fusion operations are based on an attention mechanism or bilinear pooling. In practice, fusion operations are often diverse and complicated (Mormont et al., 2020).

Integration of Multiomics Data and Precision Medicine

Now, DL algorithms still have several difficulties of integrating multiomics data or various sources of information such as pathology images and electronic medical records. The use of DL to accomplish simple tasks can yield useful results. Furthermore, complex datasets, abundant neural network architecture, and adequate DL methods are anticipated to provide useful information for precision medicine. Pathomics and radiomics are crucial components of multiomics, which also include genomics, transcriptomics, proteomics, and metabolomics information. Although there are still some limitations that restricted the direct clinical usage of multiomics analysis, there is still an increasing effort in solving the drawbacks to provide promising applications. The increasing number of omics datasets is fuelling the quantitative analysis of biological specimens at the gene, cell, and tissue levels. It will generate novel hypotheses on the molecular mechanisms of tumor development and progression for guiding precise diagnosis and treatment.

AUTHOR CONTRIBUTIONS

BL: conceptualization, writing—review and editing, supervision, and funding acquisition. XW: formal analysis, data curation, and writing—original draft. Both authors contributed to the article and approved the submitted version.

FUNDING

This study received funding from the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-038).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624820/full#supplementary-material

Supplementary Figure 1 | A timeline demonstrating the DL researches in tumor diagnosis and multi-omics analysis.

Supplementary Figure 2 | A flow chart demonstrating the difficulties and expectation of DL in tumor diagnosis and multi-omics analysis.

REFERENCES

Ali, S., Lewis, J., and Madabhushi, A. (2013). "Spatially aware cell cluster(spACCl) graphs: predicting outcome in oropharyngeal p16+ tumors," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science*, Vol. 8149, eds K. Mori, I. Sakuma, Y. Sato, C. Barillot and N. Navab (Berlin: Springer), 412–419. doi: 10.1007/978-3-642-40811-3_52

Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., et al. (2017). Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Sci. Rep.* 7:11979. doi: 10.1038/s41598-017-12320-8

Bagher-Ebadian, H., Lu, M., Siddiqui, F., Ghanem, A. I., Wen, N., Wu, Q., et al. (2020). Application of radiomics for the prediction of HPV status for patients with head and neck cancers. *Med. Phys.* 47, 563–575. doi: 10.1002/mp.13977

Beaumont, J., Acosta, O., Devillers, A., Palard-Novello, X., Chajon, E., de Crevoisier, R., et al. (2019). Voxel-based identification of local recurrence sub-regions from pre-treatment PET/CT for locally advanced head and neck cancers. *EJNMMI Res.* 9:90. doi: 10.1186/s13550-019-0556-z

Bini, S. A. (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care? *J. Arthroplasty* 33, 2358–2361. doi: 10.1016/j.arth.2018.02.067

Bogowicz, M., Jochems, A., Deist, T. M., Tanadini-Lang, S., Huang, S. H., Chan, B., et al. (2020). Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. *Sci. Rep.* 10:4542. doi: 10.1038/s41598-020-61297-4

Bogowicz, M., Tanadini-Lang, S., Guckenberger, M., and Riesterer, O. (2019a). Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Sci. Rep.* 9:15198. doi: 10.1038/s41598-019-51599-7

Bogowicz, M., Tanadini-Lang, S., Veit-Haibach, P., Pruschy, M., Bender, S., Sharma, A., et al. (2019b). Perfusion CT radiomics as potential prognostic biomarker in head and neck squamous cell carcinoma. *Acta Oncol.* 58, 1514–1518. doi: 10.1080/0284186x.2019.1629013

Bohnenberger, H., Kaderali, L., Ströbel, P., Yepes, D., Plessmann, U., Dharia, N. V., et al. (2018). Comparative proteomics reveals a diagnostic signature for pulmonary head-and-neck cancer metastasis. *EMBO Mol. Med.* 10:e8428. doi: 10.15252/emmm.201708428

Boldrini, L., Bibault, J. E., Masciocchi, C., Shen, Y., and Bittner, M. I. (2019). Deep learning: a review for the radiation oncologist. *Front. Oncol.* 9:977. doi: 10.3389/fonc.2019.00977

Brennan, K., Koenig, J. L., Gentles, A. J., Sunwoo, J. B., and Gevaert, O. (2017). Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the CpG island methylator phenotype. *EBioMedicine* 17, 223–236. doi: 10.1016/j.ebiom.2017.02.025

Carnielli, C. M., Macedo, C. C. S., De Rossi, T., Granato, D. C., Rivera, C., Domingues, R. R., et al. (2018). Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. *Nat. Commun.* 9:3598. doi: 10.1038/s41467-018-05696-2

Chen, J., Ying, H., Liu, X., Gu, J., Feng, R., Chen, T., et al. (2020). A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/tcbb.2020.2991173 [Epub ahead of print].

Chen, L., Wang, H., Zeng, H., Zhang, Y., and Ma, X. (2020). Evaluation of CT-based radiomics signature and nomogram as prognostic markers in patients with laryngeal squamous cell carcinoma. *Cancer Imaging* 20:28. doi: 10.1186/s40644-020-00310-5

Chen, L., Zhou, Z., Sher, D., Zhang, Q., Shah, J., Pham, N. L., et al. (2019). 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.* 64:075011. doi: 10.1088/1361-6560/ab083a

Clasen, K., Leibfarth, S., Hilke, F. J., Admard, J., Winter, R. M., Welz, S., et al. (2020). PET/MRI and genetic intrapatient heterogeneity in head and neck cancers. *Strahlenther. Onkol.* 196, 542–551. doi: 10.1007/s00066-020-01606-y

Conway, D. I., Purkayastha, M., and Chestnutt, I. G. (2018). The changing epidemiology of oral cancer: definitions, trends, and risk factors. *Br. Dent. J.* 225, 867–873. doi: 10.1038/sj.bdj.2018.922

Cozzi, L., Franzese, C., Fogliata, A., Franceschini, D., Navarria, P., Tomatis, S., et al. (2019). Predicting survival and local control after radiochemotherapy in locally advanced head and neck cancer by means of computed tomography based radiomics. *Strahlenther. Onkol.* 195, 805–818. doi: 10.1007/s00066-019-01483-0

Das, D. K., Bose, S., Maiti, A. K., Mitra, B., Mukherjee, G., and Dutta, P. K. (2018). Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis. *Tissue Cell* 53, 111–119. doi: 10.1016/j.tice.2018.06.004

Do, D., Le, T., Le, T. Q. T., and Le, N. Q. K. (2020). Using deep neural networks and biological subwords to detect protein S-sulfenylation sites. *Brief. Bioinformatics*. doi: 10.1093/bib/bbaa128 [Epub ahead of print].

Erickson, B. J., Korfiatis, P., Kline, T. L., Akkus, Z., Philbrick, K., and Weston, A. D. (2018). Deep learning in radiology: Does one size fit all? *J. Am. Coll. Radiol.* 15, 521–526. doi: 10.1016/j.jacr.2017.12.027

Feng, R., Liu, X., Chen, J., Chen, D. Z., Gao, H., and Wu, J. (2020). A deep learning approach for colonoscopy pathology WSI analysis: accurate segmentation and classification. *IEEE J. Biomed. Health Inform.* doi: 10.1109/jbhi.2020.3040269 [Epub ahead of print].

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

Fidler, M. M., Bray, F., Vaccarella, S., and Soerjomataram, I. (2017). Assessing global transitions in human development and colorectal cancer incidence. *Int. J. Cancer* 140, 2709–2715. doi: 10.1002/ijc.30686

Fu, Q., Chen, Y., Li, Z., Jing, Q., Hu, C., Liu, H., et al. (2020). A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. *EClinicalMedicine* 27:100558. doi: 10.1016/j.eclinm.2020.100558

Ger, R. B., Zhou, S., Elgohari, B., Elhalawani, H., Mackin, D. M., Meier, J. G., et al. (2019). Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- and PET-imaged head and neck cancer patients. *PLoS One* 14:e0222509. doi: 10.1371/journal.pone.0222509

Halicek, M., Little, J. V., Wang, X., Chen, A. Y., and Fei, B. (2019). Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. *J. Biomed. Opt.* 24, 1–9. doi: 10.1117/1.Jbo.24.3.036007

Halicek, M., Little, J. V., Wang, X., Patel, M., Griffith, C. C., Chen, A. Y., et al. (2018). Tumor margin classification of head and neck cancer using hyperspectral imaging and convolutional neural networks. *Proc. SPIE Int. Soc. Opt. Eng.* 10576:1057605. doi: 10.1117/12.2293167

Halicek, M., Lu, G., Little, J. V., Wang, X., Patel, M., Griffith, C. C., et al. (2017). Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *J. Biomed. Opt.* 22:60503. doi: 10.1117/1.Jbo.22.6.060503

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., et al. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr. Rev. Musculoskelet. Med.* 13, 69–76. doi: 10.1007/s12178-020-09600-8

Hung, M., Park, J., Hon, E. S., Bounsanga, J., Moazzami, S., Ruiz-Negrón, B., et al. (2020). Artificial intelligence in dentistry: harnessing big data to predict oral cancer survival. *World J. Clin. Oncol.* 11, 918–934. doi: 10.5306/wjco.v11.i11.918

Ibragimov, B., and Xing, L. (2017). Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med. Phys.* 44, 547–557. doi: 10.1002/mp.12045

Ibrahim, A., Gamble, P., Jaroensri, R., Abdelsamea, M. M., Mermel, C. H., Chen, P. C., et al. (2020). Artificial intelligence in digital breast pathology: techniques and applications. *Breast* 49, 267–273. doi: 10.1016/j.breast.2019.12.007

Janowczyk, A., and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* 7:29. doi: 10.4103/2153-3539.186902

Jeyaraj, P. R., and Samuel Nadar, E. R. (2019). Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J. Cancer Res. Clin. Oncol.* 145, 829–837. doi: 10.1007/s00432-018-02834-7

Jiang, Y., Yang, M., Wang, S., Li, X., and Sun, Y. (2020). Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun.* 40, 154–166. doi: 10.1002/cac2.12012

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., et al. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* 11:728. doi: 10.1038/s41467-019-13825-8

Jimenez, G., and Racoceanu, D. (2019). Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. *Front. Bioeng. Biotechnol.* 7:145. doi: 10.3389/fbioe.2019.00145

Kaddi, C. D., and Wang, M. D. (2017). Models for predicting stage in head and neck squamous cell carcinoma using proteomic and transcriptomic data. *IEEE J. Biomed. Health Inform.* 21, 246–253. doi: 10.1109/jbhi.2015.2489158

Katsoulakis, E., Yu, Y., Apte, A. P., Leeman, J. E., Katabi, N., Morris, L., et al. (2020). Radiomic analysis identifies tumor subtypes associated with distinct molecular and microenvironmental factors in head and neck squamous cell carcinoma. *Oral Oncol.* 110:104877. doi: 10.1016/j.oraloncology.2020.104877

Keek, S., Sanduleanu, S., Wesseling, F., de Roest, R., van den Brekel, M., van der Heijden, M., et al. (2020). Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemo-radiotherapy. *PLoS One* 15:e0232639. doi: 10.1371/journal.pone.0232639

Kerns, S. L., Ostrer, H., and Rosenstein, B. S. (2014). Radiogenomics: using genetics to identify cancer patients at risk for development of adverse effects following radiotherapy. *Cancer Discov.* 4, 155–165. doi: 10.1158/2159-8290.cd-13-0197

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I. H., and Kim, H. J. (2019). Deep learning-based survival prediction of oral cancer patients. *Sci. Rep.* 9:6994. doi: 10.1038/s41598-019-43372-7

Koelzer, V. H., Sokol, L., Zahnd, S., Christe, L., Dawson, H., Berger, M. D., et al. (2017). Digital analysis and epigenetic regulation of the signature of rejection in colorectal cancer. *Oncoimmunology* 6:e1288330. doi: 10.1080/2162402x.2017.1288330

Kwan, J. Y. Y., Su, J., Huang, S. H., Ghoraie, L. S., Xu, W., Chan, B., et al. (2018). Radiomic biomarkers to refine risk models for distant metastasis in HPV-related oropharyngeal carcinoma. *Int. J. Radiat. Oncol. Biol. Phys.* 102, 1107–1116. doi: 10.1016/j.ijrobp.2018.01.057

Le, N. (2019). Fertility-GRU: identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles. *J. Proteome Res.* 18, 3503–3511. doi: 10.1021/acs.jproteome.9b00411

Le, N., Do, D., Chiu, F., Yapp, E., Yeh, H., and Chen, C. (2020). XGBoost improves classification of MGMT promoter methylation status in IDH1 wildtype glioblastoma. *J. Pers. Med.* 10:128. doi: 10.3390/jpm10030128

Leger, S., Zwanenburg, A., Pilz, K., Zschaeck, S., Zöphel, K., Kotzerke, J., et al. (2019). CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer. *Radiother. Oncol.* 130, 10–17. doi: 10.1016/j.radonc.2018.07.020

Lei, H., Liu, S., Xie, H., Kuo, J. Y., and Lei, B. (2019). "An improved object detection method for mitosis detection," in *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, 130–133. doi: 10.1109/embc.2019.8857343

Leijenaar, R. T., Bogowicz, M., Jochems, A., Hoebers, F. J., Wesseling, F. W., Huang, S. H., et al. (2018). Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *Br. J. Radiol.* 91:20170498. doi: 10.1259/bjr.20170498

Lewis, J. S. Jr., Ali, S., Luo, J., Thorstad, W. L., and Madabhushi, A. (2014). A quantitative histomorphometric classifier (QuHbIC) identifies aggressive versus indolent p16-positive oropharyngeal squamous cell carcinoma. *Am. J. Surg. Pathol.* 38, 128–137. doi: 10.1097/pas.0000000000000086

Limkin, E. J., Sun, R., Dercle, L., Zacharaki, E. I., Robert, C., Reuzé, S., et al. (2017). Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann. Oncol.* 28, 1191–1206. doi: 10.1093/annonc/mdx034

Lin, B., Deng, S., Gao, H., and Yin, J. (2020). A Multi-scale activity transition network for data translation in EEG signals decoding. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/tcbb.2020.3024228 [Epub ahead of print].

Liu, T., Zhou, S., Yu, J., Guo, Y., Wang, Y., Zhou, J., et al. (2019). Prediction of lymph node metastasis in patients with papillary thyroid carcinoma: a radiomics method based on preoperative ultrasound images. *Technol. Cancer Res. Treat.* 18:1533033819831713. doi: 10.1177/1533033819831713

Lu, C., Lewis, J. S., Dupont, W. D., Plummer, W. D., Janowczyk, A., and Madabhushi, A. (2017). An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. *Mod. Pathol.* 30, 1655–1665. doi: 10.1038/modpathol.2017.98

Lv, W., Ashrafinia, S., Ma, J., Lu, L., and Rahmim, A. (2020). Multi-level multi-modality fusion radiomics: application to PET and CT imaging for prognostication of head and neck cancer. *IEEE J. Biomed. Health Inform.* 24, 2268–2277. doi: 10.1109/jbhi.2019.2956354

Mars, R. A. T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., et al. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* 182,1460–1473.e17. doi: 10.1016/j.cell.2020.08.007

Martens, R. M., Koopman, T., Noij, D. P., Pfaehler, E., Übelhör, C., Sharma, S., et al. (2020). Predictive value of quantitative (18)F-FDG-PET radiomics analysis in patients with head and neck squamous cell carcinoma. *EJNMMI Res.* 10:102. doi: 10.1186/s13550-020-00686-2

Martorell-Marugán, J., Tabik, S., Benhammou, Y., del Val, C., Zwir, I., Herrera, F., et al. (2019). "Deep learning in Omics data analysis and precision medicine," in *Computational Biology*, ed. H. Husi (Brisbane: Codon Publications).

Matta, A., Ralhan, R., DeSouza, L. V., and Siu, K. W. (2010). Mass spectrometry-based clinical proteomics: head-and-neck cancer biomarkers and drug-targets discovery. *Mass Spectrom. Rev.* 29, 945–961. doi: 10.1002/mas.20296

Mes, S. W., van Velden, F. H. P., Peltenburg, B., Peeters, C. F. W., Te Beest, D. E., van de Wiel, M. A., et al. (2020). Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures. *Eur. Radiol.* 30, 6311–6321. doi: 10.1007/s00330-020-06962-y

Moore, C., Mehta, V., Ma, X., Chaudhery, S., Shi, R., Moore-Medlin, T., et al. (2016). Interobserver agreement of confocal laser endomicroscopy for detection of head and neck neoplasia. *Laryngoscope* 126, 632–637. doi: 10.1002/lary.25646

Mormont, R., Geurts, P., and Maree, R. (2020). Multi-task pre-training of deep neural networks for digital pathology. *IEEE J. Biomed. Health Inform.* doi: 10.1109/jbhi.2020.2992878 [Epub ahead of print].

Mukherjee, P., Cintra, M., Huang, C., Zhou, M., Zhu, S., Colevas, A. D., et al. (2020). CT-based radiomic signatures for predicting histopathologic features in head and neck squamous cell carcinoma. *Radiol. Imaging Cancer* 2:e190039. doi: 10.1148/rycan.2020190039

Ni, Y. H., Ding, L., Hu, Q. G., and Hua, Z. C. (2015). Potential biomarkers for oral squamous cell carcinoma: proteomics discovery and clinical validation. *Proteomics Clin. Appl.* 9, 86–97. doi: 10.1002/prca.201400091

Niel, O., and Bastard, P. (2019). Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *Am. J. Kidney Dis.* 74, 803–810. doi: 10.1053/j.ajkd.2019.05.020

Niranjan, K. C., and Sarathy, N. A. (2018). Prognostic impact of tumor-stroma ratio in oral squamous cell carcinoma - A pilot study. *Ann. Diagn. Pathol.* 35, 56–61. doi: 10.1016/j.anndiagpath.2018.05.005

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Acad. Radiol.* 27, 106–112. doi: 10.1016/j.acra.2019.10.006

Onken, M. D., Winkler, A. E., Kanchi, K. L., Chalivendra, V., Law, J. H., Rickert, C. G., et al. (2014). A surprising cross-species conservation in the genomic landscape of mouse and human oral cancer identifies a transcriptional signature predicting metastatic disease. *Clin. Cancer Res.* 20, 2873–2884. doi: 10.1158/1078-0432.Ccr-14-0205

Pan, Y., Sun, Z., Wang, W., Yang, Z., Jia, J., Feng, X., et al. (2020). Automatic detection of squamous cell carcinoma metastasis in esophageal lymph nodes using semantic segmentation. *Clin. Transl. Med.* 10:e129. doi: 10.1002/ctm2.129

Parmar, C., Leijenaar, R. T., Grossmann, P., Rios Velazquez, E., Bussink, J., Rietveld, D., et al. (2015). Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. *Sci. Rep.* 5:11044. doi: 10.1038/srep11044

Phillips, M., Marsden, H., Jaffe, W., Matin, R. N., Wali, G. N., Greenhalgh, J., et al. (2019). Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw. Open* 2:e1913436. doi: 10.1001/jamanetworkopen.2019.13436

Poedjiastoeti, W., and Suebnukarn, S. (2018). Application of convolutional neural network in the diagnosis of jaw tumors. *Healthc. Inform. Res.* 24, 236–241. doi: 10.4258/hir.2018.24.3.236

Rattay, T., and Talbot, C. J. (2014). Finding the genetic determinants of adverse reactions to radiotherapy. *Clin. Oncol.* 26, 301–308. doi: 10.1016/j.clon.2014.02.001

Ren, J., Qi, M., Yuan, Y., and Tao, X. (2020). Radiomics of apparent diffusion coefficient maps to predict histologic grade in squamous cell carcinoma of the oral tongue and floor of mouth: a preliminary study. *Acta Radiol.* doi: 10.1177/0284185120931683 [Epub ahead of print].

Ren, J., Tian, J., Yuan, Y., Dong, D., Li, X., Shi, Y., et al. (2018). Magnetic resonance imaging based radiomics signature for the preoperative discrimination of stage I-II and III-IV head and neck squamous cell carcinoma. *Eur. J. Radiol.* 106, 1–6. doi: 10.1016/j.ejrad.2018.07.002

Shaban, M., Khurram, S. A., Fraz, M. M., Alsubaie, N., Masood, I., Mushtaq, S., et al. (2019). A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Sci. Rep.* 9:13341. doi: 10.1038/s41598-019-49710-z

Tan, M. S., Tan, J. W., Chang, S.-W., Yap, H. J., Abdul Kareem, S., and Zain, R. B. (2016). A genetic programming approach to oral cancer prognosis. *PeerJ* 4:e2482. doi: 10.7717/peerj.2482

Thong, P. S., Olivo, M., Kho, K. W., Zheng, W., Mancer, K., Harris, M., et al. (2007). Laser confocal endomicroscopy as a novel technique for fluorescence diagnostic imaging of the oral cavity. *J. Biomed. Opt.* 12:014007. doi: 10.1117/1.2710193

Tixier, F., Cheze-le-Rest, C., Schick, U., Simon, B., Dufour, X., Key, S., et al. (2020). Transcriptomics in cancer revealed by positron emission tomography radiomics. *Sci. Rep.* 10:5660. doi: 10.1038/s41598-020-62414-z

Tosado, J., Zdilar, L., Elhalawani, H., Elgohari, B., Vock, D. M., Marai, G. E., et al. (2020). Clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction. *Sci. Rep.* 10:3811. doi: 10.1038/s41598-020-60140-0

Tran, W. T., Suraweera, H., Quaioit, K., Cardenas, D., Leong, K. X., Karam, I., et al. (2019). Predictive quantitative ultrasound radiomic markers associated with treatment response in head and neck cancer. *Future Sci OA* 6:FSO433. doi: 10.2144/fsoa-2019-0048

Tseng, W. T., Chiang, W. F., Liu, S. Y., Roan, J., and Lin, C. N. (2015). The application of data mining techniques to oral cancer prognosis. *J. Med. Syst.* 39:59. doi: 10.1007/s10916-015-0241-3

Wang, S., Yang, D. M., Rong, R., Zhan, X., Fujimoto, J., Liu, H., et al. (2019). Artificial intelligence in lung cancer pathology image analysis. *Cancers* 11, 1673–1689. doi: 10.3390/cancers11111673

Werbrouck, J., De Ruyck, K., Duprez, F., Veldeman, L., Claes, K., Van Eijkeren, M., et al. (2009). Acute normal tissue reactions in head-and-neck cancer patients treated with IMRT: influence of dose and association with genetic polymorphisms in DNA DSB repair genes. *Int. J. Radiat. Oncol. Biol. Phys.* 73, 1187–1195. doi: 10.1016/j.ijrobp.2008.08.073

West, C., and Rosenstein, B. S. (2010). Establishment of a radiogenomics consortium. *Radiother. Oncol.* 94, 117–118. doi: 10.1016/j.radonc.2009.12.007

Witjes, M. J., Ilgner, J. F., Wong, B. J. F., El-Deiry, M. W., Chen, A. Y., Griffith, C. C., et al. (2018). "Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks," in *Proceedings of the Optical Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology*, San Francisco, CA.

Wu, W., Ye, J., Wang, Q., Luo, J., and Xu, S. (2019). CT-based radiomics signature for the preoperative discrimination between head and neck squamous cell carcinoma grades. *Front. Oncol.* 9:821. doi: 10.3389/fonc.2019.00821

Xu, Y., Deng, Y., Ji, Z., Liu, H., Liu, Y., Peng, H., et al. (2014). Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. *PLoS One* 9:e94022. doi: 10.1371/journal.pone.0094022

Yang, D. W., Wang, T. M., Zhang, J. B., Li, X. Z., He, Y. Q., Xiao, R., et al. (2020). Genome-wide association study identifies genetic susceptibility loci and pathways of radiation-induced acute oral mucositis. *J. Transl. Med.* 18:224.

Ye, J., Luo, J., Xu, S., and Wu, W. (2020). One-slice CT image based kernelized radiomics model for the prediction of low/mid-grade and high-grade HNSCC. *Comput. Med. Imaging Graph* 80:101675. doi: 10.1016/j.compmedimag.2019.101675

Ye, Y., Zhang, Z., Liu, Y., Diao, L., and Han, L. (2020). A multi-omics perspective of quantitative trait loci in precision medicine. *Trends Genet.* 36, 318–336. doi: 10.1016/j.tig.2020.01.009

Yu, M., Yan, H., Xia, J., Zhu, L., Zhang, T., Zhu, Z., et al. (2019). Deep convolutional neural networks for tongue squamous cell carcinoma classification using Raman spectroscopy. *Photodiagnosis Photodyn. Ther.* 26, 430–435. doi: 10.1016/j.pdpdt.2019.05.008

Yuan, Y., Ren, J., Shi, Y., and Tao, X. (2019). MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. *Eur. J. Radiol.* 117, 193–198. doi: 10.1016/j.ejrad.2019.06.019

Zhai, T. T., Langendijk, J. A., van Dijk, L. V., van der Schaaf, A., Sommers, L., Vemer-van den Hoek, J. G. M., et al. (2020). Pre-treatment radiomic features predict individual lymph node failure for head and neck cancer patients. *Radiother. Oncol.* 146, 58–65. doi: 10.1016/j.radonc.2020.02.005

Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. (2017). "MDNet: a semantically and visually interpretable medical image diagnosis network," in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*, Honolulu, HI, 3549–3557.

Zhao, Z., Li, Y., Wu, Y., and Chen, R. (2020). Deep learning-based model for predicting progression in patients with head and neck squamous cell carcinoma. *Cancer Biomark.* 27, 19–28. doi: 10.3233/cbm-190380

Zhou, Z., Chen, L., Sher, D., Zhang, Q., Shah, J., Pham, N. L., et al. (2018). Predicting lymph node metastasis in head and neck cancer by combining many-objective radiomics and 3-dimensioal convolutional neural network through evidential reasoning. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2018, 1–4. doi: 10.1109/embc.2018.8513070

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2019). UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging.* 39, 1856–1867. doi: 10.1109/tmi.2019.2959609

Zhu, Y., Mohamed, A. S. R., Lai, S. Y., Yang, S., Kanwar, A., Wei, L., et al. (2019). Imaging-genomic study of head and neck squamous cell carcinoma: associations between radiomic phenotypes and genomic mechanisms via integration of the cancer genome atlas and the cancer imaging archive. *JCO Clin. Cancer Inform.* 3, 1–9. doi: 10.1200/cci.18.00073

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H., Andrearczyk, V., Apte, A., et al. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295, 328–338. doi: 10.1148/radiol.2020191145

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership