MACHINE LEARNING FOR NON/ LESS-INVASIVE METHODS IN HEALTH INFORMATICS

EDITED BY: Kun Qian, Liang Zhang, Kezhi Li and Juan Liu
PUBLISHED IN: Frontiers in Digital Health, Frontiers in Physiology and Frontiers in Medicine

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-708-8
DOI 10.3389/978-2-88971-708-8

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 FOR NON/LESS-INVASIVE METHODS IN HEALTH INFORMATICS

Topic Editors:
Kun Qian, Beijing Institute of Technology, China
Liang Zhang, Xidian University, China
Kezhi Li, University College London, United Kingdom
Juan Liu, Huazhong University of Science and Technology, China

Citation: Qian, K., Zhang, L., Li, K., Liu, J., eds. (2021). Machine Learning for Non/Less-Invasive Methods in Health Informatics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-708-8

Table of Contents

frontiers in Digital Health

EDITORIAL
published: 06 October 2021
doi: 10.3389/fdgth.2021.763109

Editorial: Machine Learning for Non/Less-Invasive Methods in Health Informatics

Kun Qian[1]*, *Liang Zhang*[2,3], *Kezhi Li*[4] and *Juan Liu*[5]

[1] School of Medical Technology, Beijing Institute of Technology, Beijing, China, [2] School of Computer Science and Technology, Xidian University, Xi'an, China, [3] Xi'an Key Laboratory of Intelligent Software Engineering, Xidian University, Xi'an, China, [4] Institute of Health Informatics (IHI), University College London (UCL), London, United Kingdom, [5] Department of Plastic Surgery, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Keywords: digital health, medicine 4.0, intelligent medicine, machine learning, deep learning, artificial intelligence, non/less-invasive methods

Editorial on the Research Topic

Machine Learning for Non/Less-Invasive Methods in Health Informatics

1. INTRODUCTION

At the time of writing this editorial, COVID-19, as an unprecedented pandemic, has caused more than 4.4 million people left us foreverdeaths worldwide (with more than 210 million confirmed cases) in the world[1]. As researchers, this fact urges us to think about how to leverage the power of advanced technologies in improving the life quality of human beings and fighting against the ongoing and/or future pandemic. In particular, the core technology of artificial intelligence (AI), i. e., machine learning (ML) (1), has been playing an increasingly important role in leading the frontiers of Medicine improving the field of medicine 4.0.

In recent years, non/less-invasive methods are have been fast developing in clinical practice, which can considerably reduce the pains and burdens to patients physiologically and psychological pain of patiently. On the one hand, benefited from the breakthroughs in big data, the internet of things (IoT), 5G, cloud computing, high performance computing (HPC), and wearable sensors, and other AI-enabled methods have been successfully applied to tremendous scenarios such as diagnose, treatment, and management of diseases, assisted living, and rehabilitation training. On the other hand, there are existing challenges and technical and ethical issues that need to be addressed. To this end, we organized a research topic entitled "*Machine Learning for Non/Less-Invasive Methods in Health Informatics*" to build an open forum for scientists, engineers, and clinicians to exchange their studies, insights, and perspectives via a multidisciplinary point of view. The collection work lasted for 1 year (from February 2020 to February 2021), and finally it leadsled to 16 articles accepted and published after the peer-reviewed process. There are 127 authors involved in this research topic which has attracted more than 22 000 views (to as of September 2021).

In the following parts of this editorial, we will make a brief description of the published research articles within this research topic. After that we give our perspectives toward future work.

[1]https://coronavirus.jhu.edu/map.html.

OPEN ACCESS

Edited and reviewed by:
Uwe Aickelin,
The University of Melbourne, Australia

***Correspondence:**
Kun Qian
qian@bit.edu.cn

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

Received: *23 August 2021*
Accepted: *03 September 2021*
Published: *06 October 2021*

Citation:
Qian K, Zhang L, Li K and Liu J (2021) Editorial: Machine Learning for Non/Less-Invasive Methods in Health Informatics. Front. Digit. Health 3:763109. doi: 10.3389/fdgth.2021.763109

2. DATA MODALITIES

Figure 1 shows the proportion of articles that used one kind of data modality in our collected contributions. We can find that medical imaging dominated in the application, which is related to computer vision (CV).

2.1. Image

Aging population has Aging populations have become an inevitable challenge for both developing and developed countries, which and is continuously attracting efforts from the community of AI and IoT (2). The early diagnosis of brain diseases, e. g., Alzheimer's disease (AD) (3), can be very much important essential for benefiting guaranteeing a safe, easy, and independent life for the elderly, particularly for those who are living alone. Song et al. proposed a multimodal image fusion method that combines the representations learnt from the magnetic resonance imaging (MRI) (4) and the positron emission tomography (PET) (5). In their method, both the contour and the metabolic characteristics of the subject's brain tissue are retained.

Diagnosis of cancer via imaging has always been regarded as a crucial computer-aided form of medical technology. Li et al. built a dataset of pulmonary lesions with multiple-level attributes and fine contours. Wang et al. contributed two articles in their recent studies on tumor segmentation: One used octave convolutions to learn multiple-spatial-frequency features from the computed tomography (CT) (6) images for liver tumor segmentation. The other one proposed a framework of multi-modalities interactive feature learning for brain tumor segmentation.

A hierarchical deep learning (DL) (7) network was proposed by Hong et al. in their work for diagnosing multiple visual impairment diseases. A family of multi-task and multi-label learning classifiers was employed to represent different levels of eye diseases. Forte et al. proposed a DL method for identification of acute illness and facial cues of illness. Interestingly, their experiments demonstrated that the synthetically generated data can be used to develop algorithms for health conditions.

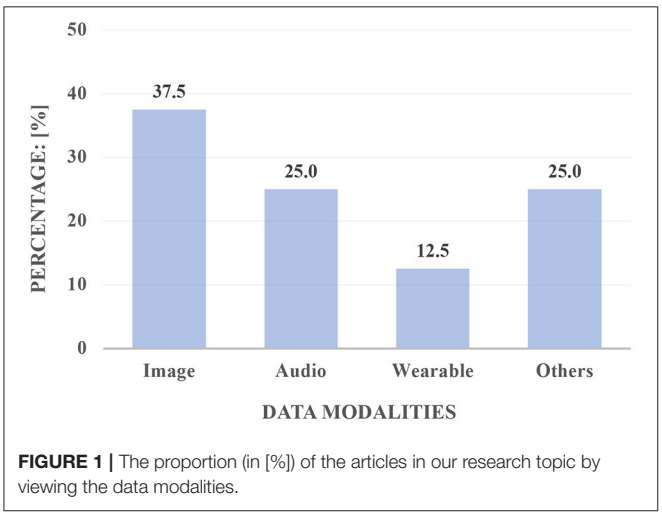

FIGURE 1 | The proportion (in [%]) of the articles in our research topic by viewing the data modalities.

2.2. Audio

Compared to its counterpart, CV, computer audition (CA) has been underestimated for a long time in the field of digital health. Nevertheless, audio as a novel digital phenotype, is attracting more attention in recent decades than ever before (8). Specifically, the analysis of cough sound has been found to be efficient in for an early-diagnosis of COVID-19 (9). Hou et al. proposed a novel feature set based on non-linear acoustic characteristics extracted from the snore sound. Their method can be used for estimating the severity levels of the obstructive sleep apnoeaapnea (OSA) (10). Li and Tian proposed an unsupervised learning method based on variational auto-encoders (VAEs) for detection of abnormal heart sounds. Yang et al. shared their clinical opinions of CA- based methods for bowel sound analysis and its potential in diagnosis of intestinal obstruction. Besides the aforementioned physiological diseases, audio can also be applied to the diagnosis of psychiatric diseases. For instance, Zhang et al. proposed a speech emotion recognition framework based on pre-trained attentive convolutional neural network, which may be adopted for developing a speech-driven method for detection of depression.

2.3. Wearable

Li et al. studied the ML-based models for estimating the associations between the body accelerations and the large-scale objective sleep data. Their study contributed to an objective evaluation of sleep quality by considering the seasonal changes in meteorological factors (e. g., ambient temperature, humidity, and sunlight). Ishaque et al. showed us a review on analyzing the heart rate variability (HRV) data and its associations in to morbidity, pain, drowsiness, stress, and exercise via signal processing (SP) and ML methods.

2.4. Others

Guo et al. used ML and DL models to predict the proximity to catastrophic decompensation from the synthetic electronic health record (EHR) data. This method can improve the timing of high-risk heart failure (HF) (11) surgical intervention. Elgendi et al. showed that unsupervised learning models can be used to reveal the novel correlates of chronic pelvic pain (CPP) (12) in women. Zhu et al. implemented ML models for predictingto predict the central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma (13). Sang et al. introduced a model using blood markers and logistic regression for diagnosis of fibrosis in southeast Asian patients suffering from the non-alcoholic fatty liver disease (NAFLD) (14).

3. PERSPECTIVES

It is encouraging to see the state-of-the-art ML models are being successfully applied to the field of non/less-invasive methods in health informatics. Nevertheless, we understand that there still exist several challenges: First, the data scarcity is restraining the reproducibility and sustainability of the relevant studies. Taking bowel sound analysis work as an example, the publicly accessible database is extremely limited. There is an urgent demand for

future collaborations between experts in AI and medicine to build open access databases. Second, breaking the walls between disciplines can never be an easy work. When reading the articles written by authors from different backgrounds, we may find there are limitations and drawbacks caused by knowledge frontiers. For instance, computer scientists can be more professional than clinicians in conducting a good ML/DL experiment whereas the latter may be clearer than the former about the motivation and the significance of the proposed research. Basic knowledge and skills training is a prerequisite for future training of experts in digital health. Third, multi-modal learning has already shown its superior performance to models trained by mono-modal. In future work, one should take image, audio, wearable, and other possible modalities into account when studying the complex associations between diseases and subjects' health data. Last but not least, ethical issues were not fully discussed in this research topic collection. We cannot ignore this important factor when working toward a human-centredcentered medical AI. Experts from social and humanity sciences are very welcome to be on board with usto collaborate with us.

AUTHOR CONTRIBUTIONS

KQ drafted the first version of this editorial and the other co-authors contributed to the proofreading work. All the co-authors contributed to this work.

FUNDING

This work was supported in part by the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China, in part by the National Natural Science Foundation of China under grant nos. 62073258 and 62072352, and in part by the National Key R&D Program of China under grant no. 2020YFF0304900, UK Rosetrees Trust UCL-IHE-2020-102 and GOSH NHS Foundation Trust DRE.

REFERENCES

1. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer (2006).
2. Qian K, Zhang Z, Yamamoto Y, Schuller BW. Artificial intelligence internet of things for the elderly: from assisted living to health-care monitoring. *IEEE Signal Proc Mag.* (2021) 38:78–88. doi: 10.1109/MSP.2021.3057298
3. Khachaturian ZS. Diagnosis of Alzheimer's disease. *Arch Neurol* (1985) 42:1097–105. doi: 10.1001/archneur.1985.04060100083029
4. Geva T. Magnetic resonance imaging: historical perspective. *J Cardiovasc Magn Reson.* (2006) 8:573–80. doi: 10.1080/10976640600755302
5. Ollinger JM, Fessler JA. Positron-emission tomography. *IEEE Signal Proc Mag.* (1997) 14:43–55.
6. Fleischmann D, Boas FE. Computed tomography–old ideas and new technology. *Eur Radiol.* (2011) 21:510–7. doi: 10.1007/s00330-011-2056-z
7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539
8. Qian K, Li X, Li H, Li S, Li W, Ning Z, et al. Computer audition for healthcare: opportunities and challenges. *Front Digit Health.* (2020) 2:5. doi: 10.3389/fdgth.2020.00005
9. Qian K, Schuller BW, Yamamoto Y. Recent advances in computer audition for diagnosing COVID-19: an overview. In: *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* (2021) Nara. p. 185–6.
10. Strollo Jr PJ, Rogers RM. Obstructive sleep apnea. *N Engl J Med.* (1996) 334:99–104.
11. Kemp CD, Conte JV. The pathophysiology of heart failure. *Cardiovasc Pathol.* (2012) 21:365–71. doi: 10.1016/j.carpath.2011.11.007
12. Howard FM. Chronic pelvic pain. *Obstet Gynecol.* (2003) 101:594–611. doi: 10.1016/s0029-7844(02)02723-0
13. Sherma SI. Thyroid carcinoma. *Lancet.* (2003) 361:501–11. doi: 10.1016/s0140-6736(03)12488-9
14. Bellentani S, Scaglioni F, Marino M, Bedogni G. Epidemiology of non-alcoholic fatty liver disease. *Dig Dis.* (2010) 28:155–61. doi: 10.1159/000282080

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 Digital Health

ORIGINAL RESEARCH
published: 07 December 2020
doi: 10.3389/fdgth.2020.576945

The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation

Aixia Guo[1], Randi E. Foraker[1,2], Robert M. MacGregor[3], Faraz M. Masood[3], Brian P. Cupps[3] and Michael K. Pasque[3]*

[1] Institute for Informatics (I[2]), Washington University School of Medicine, St. Louis, MO, United States, [2] Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, United States, [3] Department of Surgery, Washington University School of Medicine, St. Louis, MO, United States

OPEN ACCESS

Edited by:
Juan Liu,
Huazhong University of Science and Technology, China

Reviewed by:
Liang Zhang,
Xidian University, China
Zhibo Wang,
University of Central Florida, United States
Kongtao Chen,
University of Pennsylvania, United States

***Correspondence:**
Aixia Guo
aixia.guo@wustl.edu

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

Received: *27 June 2020*
Accepted: *13 November 2020*
Published: *07 December 2020*

Citation:
Guo A, Foraker RE, MacGregor RM, Masood FM, Cupps BP and Pasque MK (2020) The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. Front. Digit. Health 2:576945. doi: 10.3389/fdgth.2020.576945

Objective: Although many clinical metrics are associated with proximity to decompensation in heart failure (HF), none are individually accurate enough to risk-stratify HF patients on a patient-by-patient basis. The dire consequences of this inaccuracy in risk stratification have profoundly lowered the clinical threshold for application of high-risk surgical intervention, such as ventricular assist device placement. Machine learning can detect non-intuitive classifier patterns that allow for innovative combination of patient feature predictive capability. A machine learning-based clinical tool to identify proximity to catastrophic HF deterioration on a patient-specific basis would enable more efficient direction of high-risk surgical intervention to those patients who have the most to gain from it, while sparing others. *Synthetic* electronic health record (EHR) data are statistically indistinguishable from the original protected health information, and can be analyzed as if they were original data but without any privacy concerns. We demonstrate that *synthetic* EHR data can be easily accessed and analyzed and are amenable to machine learning analyses.

Methods: We developed *synthetic* data from EHR data of 26,575 HF patients admitted to a single institution during the decade ending on 12/31/2018. Twenty-seven clinically-relevant features were synthesized and utilized in supervised deep learning and machine learning algorithms (i.e., deep neural networks [DNN], random forest [RF], and logistic regression [LR]) to explore their ability to predict 1-year mortality by five-fold cross validation methods. We conducted analyses leveraging features from prior to/at and after/at the time of HF diagnosis.

Results: The area under the receiver operating curve (AUC) was used to evaluate the performance of the three models: the mean AUC was 0.80 for DNN, 0.72 for RF, and 0.74 for LR. Age, creatinine, body mass index, and blood pressure levels were especially important features in predicting death within 1-year among HF patients.

Conclusions: Machine learning models have considerable potential to improve accuracy in mortality prediction, such that high-risk surgical intervention can be applied only in those patients who stand to benefit from it. Access to EHR-based synthetic data derivatives eliminates risk of exposure of EHR data, speeds time-to-insight, and facilitates data sharing. As more clinical, imaging, and contractile features with proven predictive capability are added to these models, the development of a clinical tool to assist in timing of intervention in surgical candidates may be possible.

Keywords: electronic health record (EHR), machine/deep learning, heart failure, synthetic data, surgical intervention

INTRODUCTION

Heart failure (HF) patients comprise the largest, most rapidly growing, and most expensive subset of patients with cardiovascular disease[1]. In the early stages of new-onset HF, the clinical prediction of each patient's potential for a favorable response to medical therapy is critical since it determines initial management and sets the stage for their ultimate clinical course. This prediction is confounded by the fact that these patients commonly present in profound clinical HF with severely impaired left ventricular (LV) function (ejection fraction <20%) (1), only to subsequently demonstrate a very favorable response to medical therapy. Despite the gravity of their initial presentation, they are essentially cured by medical therapy alone. Conversely, many patients with an identical clinical presentation do in fact suffer precipitous deterioration (2).

Unfortunately, the poor prognostic performance of the qualitative metrics (echocardiographic, functional, metabolic, and others) that currently drive HF therapeutic clinical algorithms leaves little hope of accurate one-on-one individual patient risk-stratification (3). In fact, because of the lack of metrics that can accurately and reliably predict catastrophic hemodynamic deterioration, many HF programs have adopted a very low threshold for early and highly invasive surgical intervention (4). Thus, upon initial presentation with profound LV impairment, congestive symptoms, and borderline hemodynamics, new-onset HF patients are often rushed off to invasive surgery for intra-aortic balloon pump, extracorporeal membrane oxygenator (ECMO) support, or ventricular assist device (VAD) placement with immediate listing for cardiac transplantation (2). It is tragic to subject patients to the significant risks of surgical intervention if they can be managed on medical therapy alone. Similarly, however, over-compensating toward medical therapy in these critically ill patients also has a major downside: we are equally unable to determine which of these patients will suddenly deteriorate while on medical therapy. This deterioration is often so rapid and unheralded that sudden death or severe end-organ failure preclude any further efforts (5). All too often, we are left with patients whose "windows of opportunity" have passed under our watch.

Thus, our inability to accurately and consistently differentiate these two patient subsets at the time of presentation results in high-risk surgery being unnecessarily applied to some patients, while being denied to others who have the most to gain from it. Improving the accuracy of the metrics utilized to predict response to guideline-directed medical therapy has obvious potential to more accurately direct the clinical use of highly invasive, risky, and expensive HF surgical intervention. We seek to more accurately identify HF medical therapy *non-responders* on a one-by-one basis. This would enable their targeting for intense surveillance with an appropriately lowered threshold for early evaluation for high-risk therapy—while simultaneously sparing those who will ultimately respond to lower-risk medical therapy.

Machine learning can detect non-intuitive classifier patterns that allow for innovative combination of patient feature predictive capability (6). Recently, deep learning algorithms have been successfully used in electronic health record (EHR) data from healthcare fields. Deep learning algorithms can effectively capture the informative and useful features and patterns from the rich healthcare information in EHR data (7). For example, a very recent study showed that deep-learning-based model achieved significantly higher accuracy to predict mortality among acute heart failure patients than the existing score models and several machine learning models by using EHR data (8–13).

One of the problems with deep learning applications in heart failure is the management of large volumes of incomplete EHR information. The specter of public exposure of protected individual patient health information is also an important consideration when accessing the often-massive datasets commonly used in deep learning analysis of healthcare information (14). In regard to these concerns, *synthetic* electronic health record (EHR) data are statistically indistinguishable from that of original protected health information, and can be analyzed as if they were original data but without any privacy concerns (15).

In this investigation, we utilize an entirely synthetic dataset derived from a large cohort of HF patients seen at a single institution to test several machine learning methodologies regarding their prediction of HF outcomes. Using entirely synthetic data, we developed and compared a deep learning model—deep neural networks (DNN) (16)—with two machine learning models—random forest (RF) (17) and logistic regression (LR) (18)—to predict 1 year mortality among heart failure patients. Feature importance determinations by a tree-based classifier (19) were utilized to optimize comparison of model performance.

[1]Available online at: www.americanheart.org

TABLE 1 | Included 27 features and examples of feature values.

Feature names	Feature description and value examples
Gender	Gender (e.g., Female, Male)
Primary race	Race (e.g., White, Black, Asian, Other)
Age at event	Age of patients when the first time diagnosed with HF
Visit group	Visit types (e.g., Inpatient visit, Outpatient visit, Emergency room visit, Observation Same day Visit, Ancillary, Pre-visit, Series)
Source diagnosis	Diagnosis types (e.g., Cardiomyopathy, unspecified, Dilated cardiomyopathy, Other cardiomyopathies, Secondary cardiomyopathy, unspecified', Cardiomyopathy due to drug and external agent', Cardiomyopathy in other diseases classified elsewhere, Alcoholic cardiomyopathy, Cardiomyopathy in diseases classified elsewhere Nutritional and metabolic cardiomyopathy)
Diagnosis type	Diagnosis types (e.g., Final Diagnosis, Admitting, Reason for visit, Interim)
Facility	Facility (e.g., BJC/Washington University)
Present on admission	If HF present on admission (e.g., Yes, No, Ns)
Principal problem	If HF is the principal problem (e.g., True, False)
Problem class	Problem class (e.g., Chronic, Temporary)
Severity	Severity (e.g., High)
BMI-Age at measurement	Age of patients at the measure of BMI
BMI-Average calculated bmi	The numeric value of BMI
BP-Age at Measurement	Age of patients at the measure of BP
BP-Diastolic	The numeric value of BP Diastolic
BP-Systolic	The numeric value of BP Systolic
Steroids-Age at medication order	Age of patients at the order date of Steroids
VHD-Condition	Valvular heart disease (VHD) (e.g., Endocarditis, valve unspecified, unspecified cause, Endocarditis, valve unspecified)
Echo-Surgery code	Echocardiogram (Echo) (e.g., 1070001163)
kidD-Age at event	Kidney disease (KidD) (e.g., Chronic kidney disease, stage 3 (moderate), Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified, Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease, Chronic kidney disease, unspecified, End stage renal disease, Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease, Chronic kidney disease, Stage III (moderate), Chronic kidney disease, stage 2 (mild), Cystic kidney disease, unspecified)
creatinine-Age at event	Age of patients at the measure of creatinine
creatinine-Result value numeric	The numeric value of creatinine
SMK-Smoking tobacco status	Smoking (SMK) status (e.g., Former Smoker, Never Assessed, Never Smoker, Current Every Day Smoker, Unknown If Ever Smoked, Heavy Tobacco Smoker, Smoker, Current Status Unknown)
SMK-Age at event	Age of patients at the smoking
AF-Condition	Atrial fibrillation (AF) (e.g., Atrial fibrillation, Paroxysmal atrial fibrillation, Unspecified atrial fibrillation, Chronic atrial fibrillation, Persistent atrial fibrillation)
AF-Age at event	Age of patients at the diagnosis of AF
diab	• Diabetes (diab)—Identify diabetes presented based on if one of the following presented. • Fasting gluecose • Hemoglobin A1c • Diagnosis (e.g., Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Type 2 diabetes mellitus without complications)

METHODS

Data Source and Study Design

In this study, the electronic health records (EHR) data was from a single hospital, Barnes-Jewish Hospital from a large academic medical center, Washington University in St Louis. These data were synthesized by MDClone platform, which can create synthetic electronic health data that is statistically equivalent to original data, but contains no actual patient information[2]. The synthetic data generation platform creates a computationally derived data set which is statistically identical to that of the original patients. The computationally-derived variables and their pairwise correlations had the same or very similar distributions as the relationships among variables in the original data (20). We included a Spearman's correlation comparison between the variables in the original compared to the variables derived from the MDClone synthetic data platform (**Supplementary Figure 1**). The original patient cohort, from which the synthetic data was derived, were admitted for treatment at Barnes-Jewish Hospital with an admitting diagnosis of heart failure during the decade ending on 12/31/2018. Our goal was to predict their proximity to catastrophic HF decompensation by predicting 1-year mortality based upon features contained in their EHR after/at or prior to/at the earliest diagnoses of heart failure. We studied 26,575 (26,600) patients if using features prior to/at (if after/at) heart failure diagnoses.

For the feature extraction, we discarded features whose missing values rate exceeded 70%, as we expected that they may cause a substantial difference between features available prior to/at and after/at the time of HF diagnosis. For example, the feature "CABG—Procedure code" was included in the case of after/at HF diagnosis, but was excluded from the case of prior to/at HF diagnosis as it had a missing value rate more than 70%. For all others, we imputed any missing values as the mean value for the continuous variables and the mode value for the categorical variables. Under the criteria, there were 27 features and one outcome (death) were included in our study. The

[2]https://www.mdclone.com/

included features and possible value examples for each feature were listed in **Table 1**.

We classified the heart failure patients into two groups based upon their mortality dates: a positive class (patients who died within 365 days of initial HF presentation) and a negative class (patients who did not die or died later than 365 days after HF presentation). There were 1,768 (1,735) positive patients and 24,807 (24,865) negative patients if using features prior to/at (after/at) the first heart failure diagnosis dates.

Statistical Analysis

We then applied machine learning and deep learning models to predict the all-cause mortality within 1 year by using features either prior to/at or after/at heart failure diagnoses. The three models employed were deep neural networks (DNN), random forest (RF) and logistic regression (LR). For each model of each prediction, we utilized five-fold cross validation by dividing the dataset into five-folds, with each fold serving as a test dataset and the remaining four-folds comprising a training dataset. There was a significant imbalance between the positive and negative classes. We utilized Synthetic Minority Over-sampling Technique (SMOTE) (21) to deal with the imbalanced issue by oversampling positive patients to the same amount of negative patients in each cross validation, i.e., the four-folds training datasets was oversampled by SMOTE while the remaining one-fold which served as testing dataset kept as original without using SMOTE to oversample.

Our DNN was comprised of an input layer (with 27 dimensions), 5 hidden layers (with 256, 256, 128, 64, and 32 dimensions, respectively) and a scalar output layer. We used the Sigmoid function (22) at the output layer and ReLu function (23) at each hidden layer. Binary cross-entropy was used as loss function and Adam optimizer (24) was used to optimize the models with a mini-batch size of 64 samples. The hyperparameter of network depth was searched from 2 to 8 hidden layers. To avoid overfitting, an early stopping technique was used which would stop training when the monitored loss metric stopped improving after 5 epochs. We set the maximum epochs at 50. The LR and RF models were configured by the default options in package of Scikit-learn in Python 3. We performed a grid

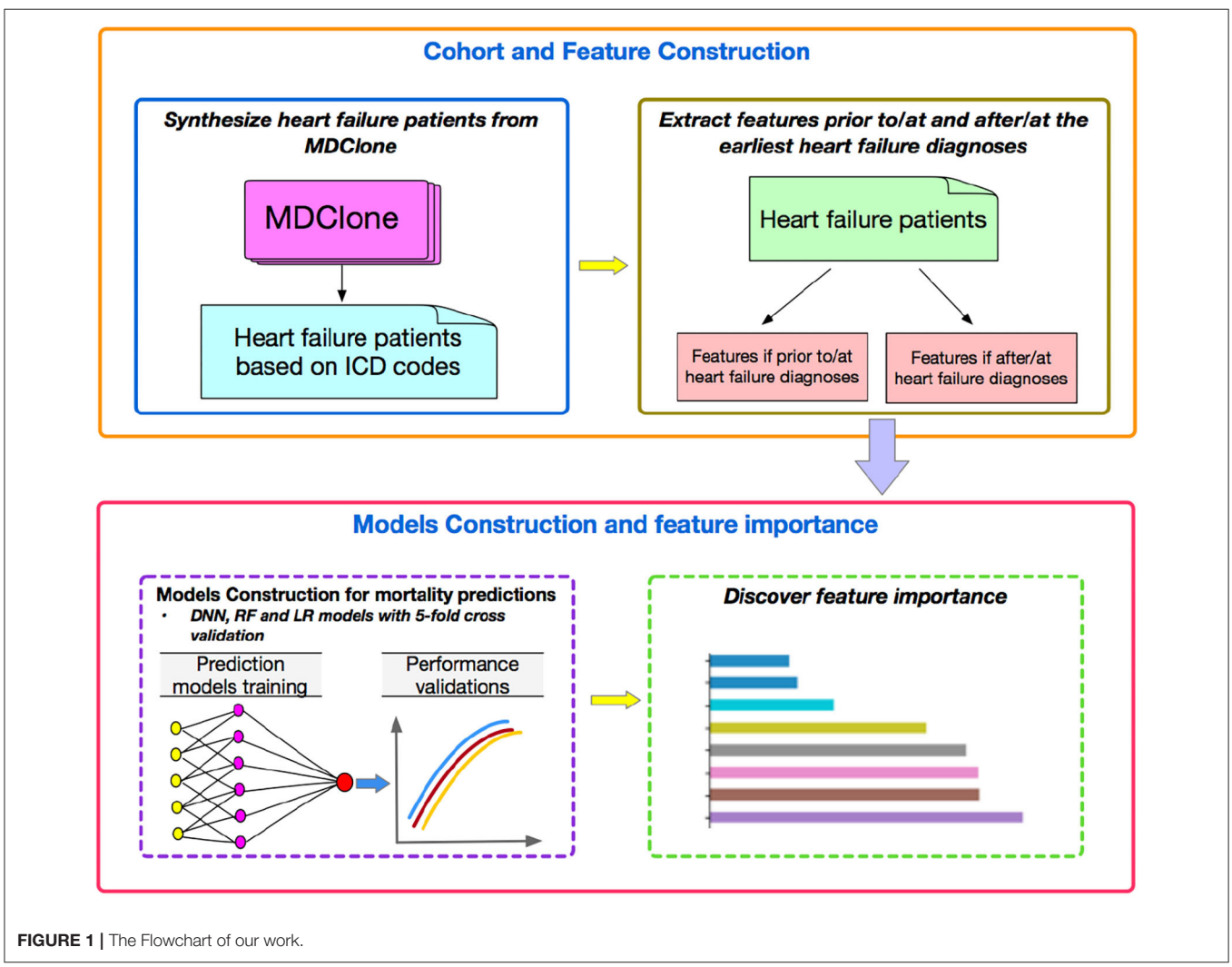

FIGURE 1 | The Flowchart of our work.

search of hyperparameters for the RF model by five-fold cross validation. We searched the number of trees in the forest for 100, 200, 500, and 700, and we considered the number of features for the best split according to auto, sqrt, and log2. We also did a grid search of hyperparameter tuning for LR models by five-fold cross validation. In penalization, we searched the norm for L1 and L2 norm, and the inverse value of regularization strength for 10 different numbers spaced evenly on a log scale of [0, 4]. We achieved the best hyperparameters on the default configurations for both RF and LR models.

Finally, we investigated the feature importance to better understand which features played more important roles compared to others by tree-based classifiers. We quantified the importance of features by ordering them in an ascending order. The prediction performances were then validated by using different numbers of top features in the three machine learning models. **Figure 1** represents a flowchart of our data analysis. Analyses were conducted by using the libraries of Scikit-learn, Keras, Scipy, Matplotlib with Python, version 3.6.5 (2019).

TABLE 2 | Characteristics [mean (SD) or *n* (%)] of the two study populations.

Demographics	After/at heart failure (*n* = 26,600)	Prior to/at heart failure (*n* = 26,575)
Age	63 (17)	63 (17)
Gender		
Female	11,116 (41.8)	11,103 (41.8)
Male	15,484 (58.2)	15,418 (58.0)
Race		
White	15,218 (57.2)	15,420 (58.0)
Black	4,738 (17.8)	5,015 (18.9)
Other/unknown	6,644 (25.0)	6,140 (23.1)
BMI	29.6 (6.3)	29.8 (6.2)
Diastolic blood pressure (DBP, mmHg)	73 (15)	75 (15)
Systolic blood pressure (SBP, mmHg)	127 (23)	131 (23)
Valvular heart disease (VHD) present	327 (1.2)	388 (1.5)
Echocardiogram (ECHO) present	38 (0.1)	5 (0.0)
Creatinine level	1.63 (1.01)	1.41 (0.88)
Current smoker	703 (2.6)	191 (0.7)
Diabetes present	3,809 (14.3)	5,174 (19.5)

RESULTS

The average age for the two study populations was 63 (**Table 2**). Approximately 58% of patients in both groups were male and white. There were 327 patients (388 for prior to/at heart failure) who also had a diagnosis of valvular heart disease (VHD) and 14% (19% for prior to/at heart failure) of the patients had diabetes. The average creatinine level was 1.63 (1.41 for prior to/at heart failure) for patients. Approximately 7% of the patients of both study populations died within 1 year from the earliest diagnosis of heart failure.

Figure 2 shows the prediction performance for 1-year mortality by using after/at and prior to/at first diagnosis heart failure start date. All the 27 features are used for these predictions. In the two study groups, DNN models outperformed the other two models of RF and LR and achieved the highest AUC values: the mean AUC value of DNN was 0.82 (0.80) compare to RF and LR with 0.74 (0.72) and 0.74 (0.74) in the five-fold cross validation models.

Figure 3 shows the feature importance by the tree-based classifier method for both cases. In the first case of after/at heart failure diagnosis, it shows that the most important features included blood pressure, creatinine levels, body mass index (BMI) etc. In the case of prior to/at heart failure diagnosis, the

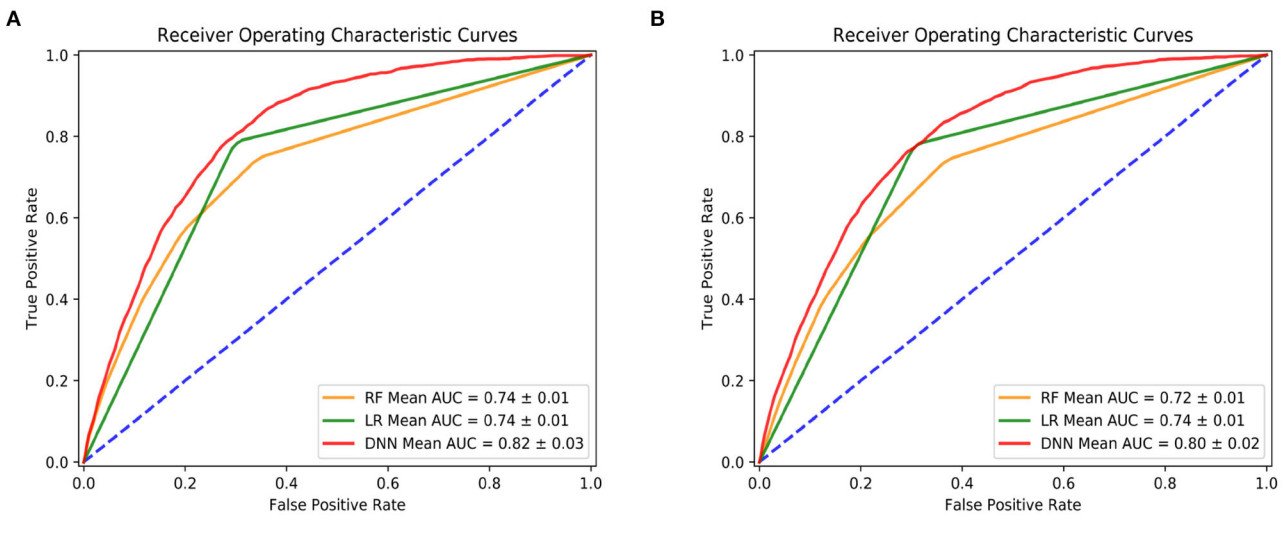

FIGURE 2 | Prediction performance by deep neural network (DNN), random forest (RF) and logistic regression (LR). **(A)** Is using features after and at heart failure diagnoses date; **(B)** is using features prior to and at heart failure diagnoses date.

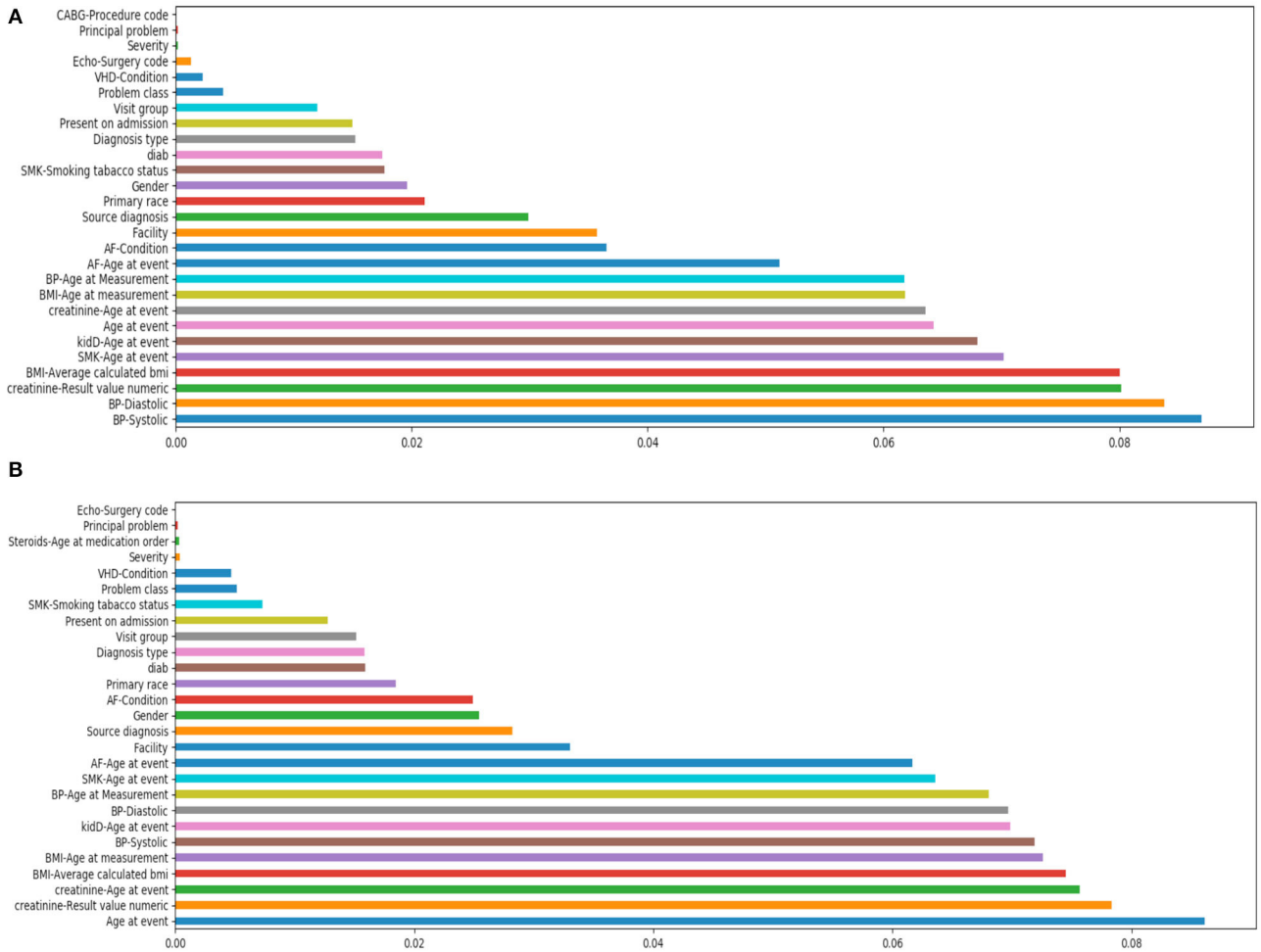

FIGURE 3 | Feature importance study by tree-based classifier. **(A)** Is using features after/at heart failure diagnoses date; **(B)** is using features prior to/at heart failure diagnoses date.

most important features were age at the first diagnosis of heart failure, creatinine, and BMI.

Figure 4 shows the prediction performance by using all different numbers of top features for 3 models of DNN, RF and LR. For example, if # of top feature = 12, it means the models used only the top 12 important features listed in **Figure 3** in each case. In all cases, DNN models outperformed RF and LR models. The AUC values were markedly reduced in both study groups when the features dropped from 12 to 11, for all three machine learning models.

DISCUSSION

In this study, we utilized 10-year synthetic EHR data by MDClone platform to identify heart failure patients to predict the mortality of patients within 1 year from the first diagnoses of heart failure by machine learning and deep learning models. We also investigated the top important features by tree-based classifier and tested all different possible numbers of top features as the inputs for all the three models in both two cases.

Our results indicated that the deep learning model DNN can effectively predict the mortality within 1 year of patients by using features such as measurements and diagnoses from either after/at or prior to/at the first diagnoses of heart failure. Our results also indicated that features such as blood pressure, BMI and creatinine levels are the most informative ones, and in all cases DNN models outperformed RF and LR models. Three models consistently indicated that there was a significant reduction in accuracy of model prediction, as represented by AUC values, when the number of most important features utilized in the model were reduced from 12 to 11, suggesting that 12 features would be a potential threshold if a reduction in features is necessary.

The case of using features from prior to/at HF diagnosis was to provide insights into the 1-year mortality prediction at the time of HF diagnosis, in which a mortality prediction risk score was calculated for patients at the time of HF diagnosis. The case of using features from after/at HF diagnosis to enhance the 1-year mortality prediction following HF diagnosis. At each follow-up time point, a predicted 1-year mortality risk score could be

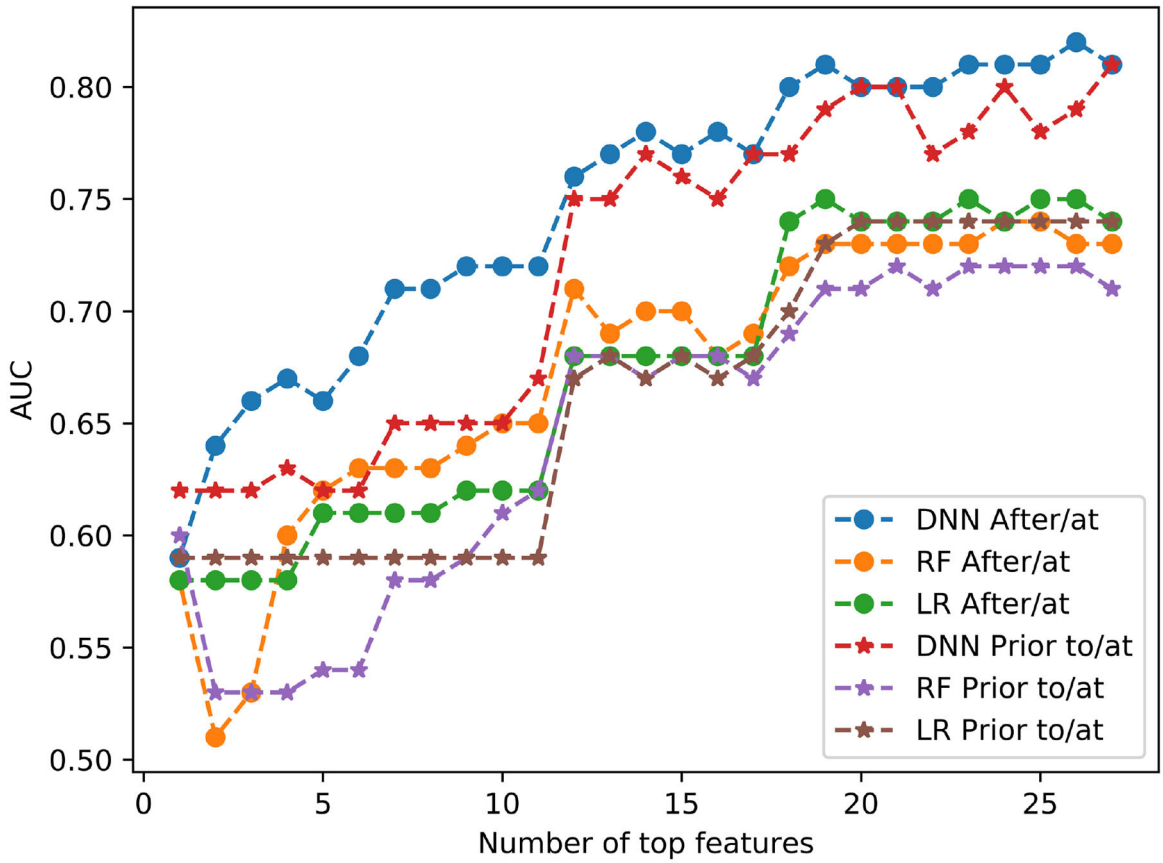

FIGURE 4 | Model performance with different numbers of top features for DNN, RF and LR.

calculated for patients. Based on these scores, providers may make particular treatment decisions to optimize prevention and more effectively manage these patients.

The use of synthetic EHR data in deep learning models to predict 1-year mortality among heart failure patients is unique to this investigation, which also emphasized the use of feature importance to guide mechanistic hypotheses in this HF patient population. This use of synthetic EHR data containing no protected health information uniquely allows a broader application of our results by enabling the sharing of data without risk of exposure of individual patient EHR information. In future work, we plan to pursue additional statistical analyses such as permutation tests and statistical comparisons to investigate the impact of feature importance. We acknowledge that our current DNN model had a relatively simple structure with 5 hidden layers. In future work, we will investigate more complicated structures of DNN models with more hidden layers (e.g., from 2 to 32) and evaluate other novel deep learning models.

LIMITATIONS

This study is limited by the small number of health-related features included in our machine learning applications. Many features were not used in our models because of a high proportion of missing values. As the EHR continues to expand health data inclusion and improve in the accuracy, consistency, and completeness of the data included, model performance will almost assuredly improve by the inclusion of clinical variables with proven predictive capability.

CONCLUSIONS

Machine learning models have obvious and considerable potential to improve accuracy in the risk stratification of HF patients. The ability to use EHR variables to identify HF patient proximity to HF decompensation and death would allow the more accurate and timely application of high-risk surgical intervention. Access to synthetic data derivatives speeds time-to-insight using EHR data, and allows the sharing of massive datasets—while simultaneously reducing privacy concerns by eliminating the risk of personal data exposure. As the EHR becomes more complete, the inclusion of advanced clinical, imaging, and contractile features—with proven predictive capability—in predictive machine learning models can be expected to improve their accuracy. As the accuracy of machine learning, and especially deep learning, models improves, the development of a clinical tool capable of assisting clinicians in the timing of intervention in surgical candidates may be

possible. Further, our ability to quantify individual EHR feature impact on mortality prediction may allow the generation of non-intuitive mechanistic hypotheses leading to potential preventative clinical intervention.

DATA AVAILABILITY STATEMENT

The datasets for the current study are available from the corresponding author on reasonable request. Requests to access these datasets should be directed to aixia.guo@wustl.edu.

ETHICS STATEMENT

Ethical approval was not provided for this study on human participants because synthetic electronic health data that contains no actual patient information was used. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

RF contributed to the study design. AG conducted the analysis and wrote the manuscript. RM, FM, BC, and MP provided insightful discussions, reviewed the results and revised the manuscript. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2020.576945/full#supplementary-material

REFERENCES

1. Niebauer J, Clark AL, Anker SD, Coats AJS. Three year mortality in heart failure patients with very low left ventricular ejection fractions. *Int J Cardiol.* (1999) 70:245–7. doi: 10.1016/S0167-5273(99)00088-1
2. Inamdar A, Inamdar A. Heart failure: diagnosis, management and utilization. *J Clin Med.* (2016) 5:62. doi: 10.3390/jcm5070062
3. Rastogi A, Novak E, Platts AE, Mann DL. Epidemiology, pathophysiology and clinical outcomes for heart failure patients with a mid-range ejection fraction. *Eur J Heart Fail.* (2017) 19:1597–605. doi: 10.1002/ejhf.879
4. Ministeri M, Alonso-Gonzalez R, Swan L, Dimopoulos K. Common long-term complications of adult congenital heart disease: avoid falling in a H.E.A.P. *Expert Rev Cardiovasc Ther.* (2016) 14:445–62. doi: 10.1586/14779072.2016.1133294
5. Tomaselli GF, Zipes DP. What causes sudden death in heart failure? *Circ Res.* (2004) 95:754–63. doi: 10.1161/01.RES.0000145047.14691.db
6. Liu H, Fu Z, Yang K, Xu X, Bauchy M. Machine learning for glass science and engineering: a review. *J Non Cryst Solids.* (2019) 119419. doi: 10.1016/j.nocx.2019.100036
7. Goodfellow I, Bengio Y, Courville A. *Deep Learning.* Montreal, QC: MIT Press (2016).
8. Kwon J, Kim K-H, Ki-Hyun J, Lee SE, Lee HY, Cho. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLoS ONE.* (2019) 14:e0219302. doi: 10.1371/journal.pone.0219302
9. Bello GA, Dawes TJW, Duan J, Biffi C, de Marvao A, Howard LSGE, et al. Deep learning cardiac motion analysis for human survival prediction. *Nat Mach Intell.* (2019) 1:95–104. doi: 10.1038/s42256-019-0019-2
10. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Hear Fail.* (2019) 6:428–35. doi: 10.1002/ehf2.12419
11. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail.* (2020) 22:139–47. doi: 10.1002/ejhf.1628
12. Guo A, Pasque M, Loh F, Mann DL, Payne PRO. Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models. *Curr Epidemiol Reports.* (2020). doi: 10.1007/s40471-020-00259-w
13. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Hear Fail.* (2020) 8:12–21. doi: 10.1016/j.jchf.2019.06.013
14. Nass SJ, Levit LA, Gostin LO. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research.* Washington, DC: National Academies Press (2009). doi: 10.17226/12458
15. Foraker RE, Mann DL, Payne PRO. Are synthetic data derivatives the future of translational medicine? *JACC BASIC TO Transl Sci.* (2018) 3:716–8. doi: 10.1016/j.jacbts.2018.08.007
16. Bengio Y. Learning deep architectures for AI. Found Trends Mach Learn. (2009) 2:1–127. doi: 10.1561/2200000006
17. Ho TK. Random decision forests. In: *Proceedings of the International Conference on Document Analysis and Recognition.* ICDAR. (1995). p. 1:278–82.
18. Hosmer D, Lemeshow S, Sturdivant RX. Model-building strategies and methods for logistic regression. *In: Applied Logistic Regression*. (2013). doi: 10.1002/9781118548387
19. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. New York, NY: Routledge(2017). doi: 10.1201/9781315139470
20. Foraker R, Yu S, Michelson A, et al. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA OPEN*. (2020).
21. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* (2002) 16:321–57. doi: 10.1613/jair.953
22. Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*. Heidelberg: Springer. (1995). p. 930. doi: 10.1007/3-540-59497-3_175
23. Nair V, Hinton GE. Rectified linear units improve Restricted Boltzmann machines. In: *ICML Proceedings, 27th International Conference on Machine Learning* (Madison, WI: Omnipress). (2010).
24. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *CoRR International Conference on Learning Representations.* (2014). *arXiv:1412.6980 [cs.LG]*

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 Digital Health

METHODS
published: 18 December 2020
doi: 10.3389/fdgth.2020.600604

Machine Learning Revealed New Correlates of Chronic Pelvic Pain in Women

Mohamed Elgendi[1,2,3]*, *Catherine Allaire*[2,3], *Christina Williams*[2,3], *Mohamed A. Bedaiwy*[2,3] *and Paul J. Yong*[2,3]

[1] School of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada, [2] Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada, [3] British Columbia (BC) Children's & Women's Hospital, Vancouver, BC, Canada

Chronic pelvic pain affects one in seven women worldwide, and there is an urgent need to reduce its associated significant costs and to improve women's health. There are many correlated factors associated with chronic pelvic pain (CPP), and analyzing them simultaneously can be complex and involves many challenges. A newly developed interaction ensemble, referred to as INTENSE, was implemented to investigate this research gap. When applied, INTENSE aggregates three machine learning (ML) methods, which are unsupervised, as follows: interaction principal component analysis (IPCA), hierarchical cluster analysis (HCA), and centroid-based clustering (CBC). For our proposed research, we used INTENSE to uncover novel knowledge, which revealed new interactions in a sample of 656 patients among 25 factors: age, parity, ethnicity, body mass index, endometriosis, irritable bowel syndrome, painful bladder syndrome, pelvic floor tenderness, abdominal wall pain, depression score, anxiety score, Pain Catastrophizing Scale, family history of chronic pain, new or re-referral, age when first experienced pain, pain duration, surgery helpful for pain, infertility, smoking, alcohol use, trauma, dysmenorrhea, deep dyspareunia, CPP, and the Endometriosis Health Profile for functional quality of life. INTENSE indicates that CPP and the Endometriosis Health Profile are correlated with depression score, anxiety score, and the Pain Catastrophizing Scale. Other insights derived from these ML methods include the finding that higher body mass index was clustered with smoking and a history of life trauma. As well, sexual pain (deep dyspareunia) was found to be associated with musculoskeletal pain contributors (abdominal wall pain and pelvic floor tenderness). Therefore, INTENSE provided expert-like reasoning without training any model or prior knowledge of CPP. ML has the potential to identify novel relationships in the etiology of CPP, and thus can drive innovative future research.

Keywords: obstetrics, gynecology, chronic pelvic pain in women, endometriosis, quality of life, infertility, data science, artificial intelligence

OPEN ACCESS

Edited by:
Kezhi Li,
University College London, United Kingdom

Reviewed by:
Chenggang Lai,
Walmart Labs, United States
Omero Benedicto Poli Neto,
University of São Paulo, Brazil

***Correspondence:**
Mohamed Elgendi
moe.elgendi@gmail.com

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

Received: *30 August 2020*
Accepted: *08 October 2020*
Published: *18 December 2020*

Citation:
Elgendi M, Allaire C, Williams C, Bedaiwy MA and Yong PJ (2020) Machine Learning Revealed New Correlates of Chronic Pelvic Pain in Women.
Front. Digit. Health 2:600604.
doi: 10.3389/fdgth.2020.600604

1. INTRODUCTION

Chronic pelvic pain affects nearly 15% of women, with major impact on quality of life and health care costs (1, 2). The etiology of chronic pelvic pain (CPP) is very complex, involving many interrelated and correlated factors over the course of one's life, including the presence or absence of endometriosis. Recently, Yosef et al. (3) performed an exploratory analysis of multifactorial variables independently associated with the severity of CPP in women. Among the findings, they found that abdominal wall pain (i.e., pain related to the abdominal wall musculature), tenderness of the pelvic floor musculature, and Pain Catastrophizing Scale were independently associated with the severity of CPP with significance of $p \leq 0.05$, but surprisingly, no association with endometriosis. However, the authors used multiple linear regression and thus did not investigate the simultaneous dynamics between factors. In complex clinical conditions, such as CPP, straightforward regression analyses may provide an incomplete view of the impact of each factor in relation to other factors.

It is our understanding that, as it currently stands, minimal effort has been made to examine different factors simultaneously using artificial intelligence (AI), with a focus on the network dynamics between potential factors, in this area of medicine. Thus, in this study, we utilize AI-informed machine learning (ML) methods to uncover the hidden interactions among all factors and explore the importance of each factor for CPP in women.

2. MATERIALS AND METHODS

2.1. Pelvic Pain and Endometriosis Dataset

This study is a re-analysis of cross-sectional data from Yosef et al. (3) ($N = 656$ subjects), which are taken from a prospective database from a tertiary referral center for pelvic pain and endometriosis using the REDCap system (4, 5). Participants from December 2013 to September 2015 were included, who completed an online questionnaire and underwent a complete history/examination. Exclusion criteria included age $>$ 50 or menopausal (6). The sample characteristics have been published previously (3), with a mean ($\pm$ 1 standard deviation) age of 34.5 ($\pm$7.6) years and body mass index (BMI) of 25.3 ($\pm$5.7) kg/m^2, with 49% of the sample nulligravid and 74% of the sample Caucasian, who had underlying diagnoses of endometriosis (57%), irritable bowel syndrome (53%), painful bladder syndrome (43%), and abdominal wall trigger points (27%). We chose 25 factors of clinical importance in this cohort based on the initial analysis of Yosef et al. (3), as shown in **Table 1**.

TABLE 1 | The 25 factors of clinical importance to chronic pelvic pain used in this study.

Variable name	Description
Age	Years
Parity	Nulliparous (no childbirth) vs. Parous (at least one childbirth)
Ethnicity	Other vs. Caucasian
Body mass index (BMI)	kg/m^2
Endometriosis (3)	Absent/not suspected (prior negative laparoscopy or not clinically suspected) vs. clinically suspected (based on history and/or tenderness on examination (6)) vs. confirmed present (prior surgical diagnosis or current nodule or endometrioma on exam/imaging)
Irritable bowel syndrome (IBS) (7)	Present vs. Absent
Painful bladder syndrome (PBS) (8)	Present vs. Absent
Pelvic floor tenderness	Tenderness of the levator ani pelvic floor musculature on examination, as a sign of myofascial pelvic pain syndrome: Present vs. Absent
Abdominal wall pain	Abdominal wall pain diagnosed by the Carnett test (1), with abdominal tenderness not changing or worsening with tensing of the abdominal wall musculature, often secondary to myofascial trigger points: Present vs. Absent
Depression	Patient Health Questionnaire-9 questionnaire (9)
Anxiety	Generalized Anxiety Disorder-7 questionnaire (10)
Pain catastrophizing	Pain Catastrophizing Scale (11) (measurement of magnification or rumination on symptoms, as well as feelings of helplessness)
Family history of chronic pain	Yes vs. No vs. Do not know
Referral type	New or re-referral
Age when first experienced pain	Years
Pain duration	Years
Patient report that prior surgery was helpful for pain	Yes vs. No vs. No prior surgery
Infertility	Yes vs. No vs. Never tried for pregnancy
Smoking	Yes vs. No
Alcohol use	Drinks/week
Trauma	Based on 7 questions about childhood or adult sexual, physical, or emotional abuse3 (scored from 0–7)
Dysmenorrhea	Menstrual cramps (rated 0–10)
Deep dyspareunia	Pain with deep penetration during sexual activity (rated 0–10)
Chronic pelvic pain	Chronic pain in the pelvic (rated 0–10)
Endometriosis Health Profile-30 (12)	indicating worse quality-of-life)

2.2. Pre-processing Step

To standardize the values of each factor, we applied the Z-score normalization. It is implemented by subtracting the mean from each factor, then divide the result by the standard deviation of each factor as follows: $F = (X - \bar{X})/(\sigma)$,

where F is the normalized factor vector, X is the raw factor vector, ($\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$) is the mean of the factor vector, ($\sigma = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2$) is the standard deviation of the factor vector, N is the number of subjects, which equals 656 in this work.

2.3. INTENSE Algorithm

INTENSE, a newly developed interaction ensemble method that utilizes various clustering models (13) was used. Multiple models for clustering are used in existing literature; however, each has

its own set of rules for defining factors with "mathematical similarity." When implementing the INTENSE method, results are aggregated from three different interaction methods, with a different mathematical view for each:

1. Interaction based on principal component analysis examines correlations between all factors based on eigen values.
2. Interaction based on connectivity clustering examines distance-wise similarity between each two factors.
3. Interaction based on centroid clustering examines distance-wise similarity between all factors and the proposed number of centroids.
4. INTENSE aggregates findings from the above interactions for a consensus regarding how all the factors in the dataset interact with each other.

Since each method for interactions has limitations, such as the initial value used, fixed thresholds, and so on, INTENSE was created. When combining results from various interaction models that utilized different geometrical concepts, the output will be an aggregate of agreed upon results, thus creating a more robust conclusion.

2.3.1. Interaction Principal Component Analysis

A correlation-based machine learning method was used in this study, referred to as the IPCA, proposed in a previous study (13). As an unsupervised ML technique, within set of observations of attributes that are potentially correlated, it identifies linearly uncorrelated attributes (in this instance, factors). A decorrelation process is first used that does not need any initial conditions for the processed attributes. Next, the Pearson's correlation is applied. In the absence of any training of labeling, IPCA can automatically reveal hidden interactions between factors, and provide a true level of learning where new behaviors among the factors examined are uncovered. Algorithm 1 shows the pseudocode of IPCA.

Algorithm 1: Interaction Principal Component Analysis (IPCA).

// Assume that there is a set of M factors

$$\{\Gamma_1, \Gamma_2, ..., \Gamma_M, \}$$

// Shape the size for each factor to be $N \times 1$, where $N =$ number of subjects

$$\underset{N\times 1}{\Gamma}$$

// Normalize factors

$$\Psi_j = \frac{1}{N}\sum_{i=1}^{N}\Gamma_{i,j} \quad i = 1, 2, ..., N; j = 1, 2, ..., M$$

$$\sigma_j = \frac{1}{N-1}\sum_{i=1}^{N}(\Gamma_{i,j} - \Psi_j)^2$$

$$Z_j = (\Gamma_{i,j} - \Psi_j)/\sigma_j$$

// Keep iterating until two factors or more found to be inter-correlated

while $j \in Z', j < 2$ **do**

  // Compute covariance matrix C from Z

  $$C = \frac{1}{M}Z'Z$$

  // Calculate eigen values and eigen vectors

  $$Cv = ev$$

  // Rank eigen values in descending order and their eigen vectors

  $$\mathrm{PC}_j = \mathtt{SORT}(e, v, '\mathit{descending}')$$

  // Examine correlation between factors and PCs

  for $j = 1 : M$ **do**

    for $k = 1 : M$ **do**

      $r_{j,k} = \mathtt{CORR}(PC_j, Z_k)$

    end

  end

  // find Z factors that are correlated PCs with a correlation greater than 0.5

  $$Z'_j = \mathtt{FIND}(Z_j, |r > 0.5|)$$

end

2.3.2. Hierarchical Cluster Analysis

An unsupervised ML approach, termed the hierarchical cluster analysis (HCA), connects "factors" and based on their distance, groups are formed. Among biosignals, HCA can visualize and quantify dissimilarities. To provide a hierarchical cluster, the Euclidean distance $d = ||a - b||_2$ was implemented, also referred to as the dendrogram. "Average" is utilized here as the linkage criterion to determine the distance between all factors as a function of the pairwise distances between observations,which is defined as $d(W, v) = \sum \frac{d(w[i],v[j])}{|w|\times|v|}$, for all data points i and j, where $|w|$ and $|v|$ are the cardinalities of clusters $|w|$ and $|v|$, respectively.

2.3.3. Centroid-Based Clustering

A well-known and relatively simple centroid-based algorithm for clustering, known as the K-means clustering (centroid-based clustering, CBC) is used here. The number of factors F is divided into K disjoint clusters. The statistical means of a group of factors form clusters. In other words, the factors with minimum distance between them and their statistical mean formulate an independent cluster. To find the minimum distance between a group of factors and their corresponding statistical mean, the within-cluster sum-of-squares, also known as inertia, is commonly used. Inertia is defined as: $||f_i - \mu_j||^2$, where f_i is

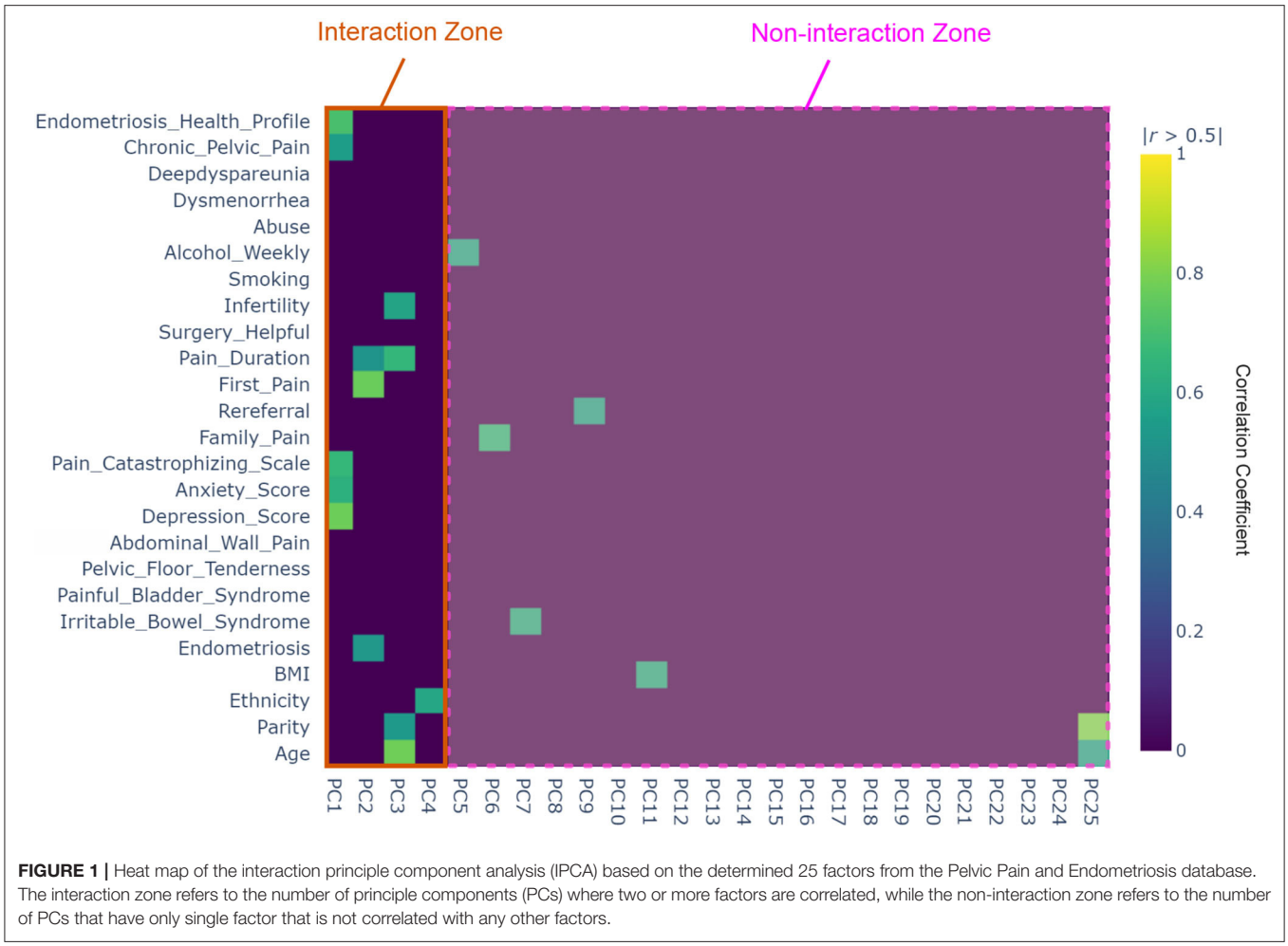

FIGURE 1 | Heat map of the interaction principle component analysis (IPCA) based on the determined 25 factors from the Pelvic Pain and Endometriosis database. The interaction zone refers to the number of principle components (PCs) where two or more factors are correlated, while the non-interaction zone refers to the number of PCs that have only single factor that is not correlated with any other factors.

all values in factor i, and μ_j refers to the mean of all factors in cluster j. It is highly recommended to apply PCA before CBC clustering to reduce dimensionality and visualize the results in two dimensions.

2.3.4. Ensemble Method

We used the "majority voting" rule to combine conceptually different interaction recommendations by different methods. In other words, in majority voting, the consistent interactions suggested by different clustering methods are the ideal and more meaningful interactions. For example, if the recommendations are

1. IPCA $\longmapsto$ C1 (f_1, f_3), C2 (f_5, f_7), C3 (f_8, f_9, f_{10})
2. HCA $\longmapsto$ C1 (f_1, f_3, f_5), C2 (f_8, f_9)
3. CBC $\longmapsto$ C1 (f_1, f_3, f_6), C2 (f_6, f_7)

then the ensemble decision is that f_1 and f_3 interact strongly and more strongly correlated among all other factors.

2.4. Software

We used Python 3.6.5 software and Matlab 2018b software to analyze the data.

3. RESULTS

The significance of the 25 principal components extracted from the database are shown in **Figure 1**. Most of the variance is explained by PC1 (40%), which reflects the relevance and importance of factors correlated with PC1; 23% of the variance is explained by PC2, and lastly ∼17% of variance is explained by PC3. It can be seen that PC1 is the most important, followed by PC2. PC25 shows to be the least important.

A correlation matrix heat map, shown in **Figure 1**, demonstrates the interaction between factors and all PCs. Diagonal entries are equal to one. There are four 25 × 25 blocks. The correlation matrix for PCs in the top right block contains zeros, confirming that the principle components (PCs) are mutually orthogonal, and hence are not correlated. Correlation between all factors is shown in the bottom-left block, and thus, there no correlation was reported.

As seen in **Figure 1**, the 25 × 25 heatmap contains interesting results about the factors interaction. IPCA involves two steps: First, it identifies the most strongly interacting factors, following which IPCA is run again on theses selected factors. In the first

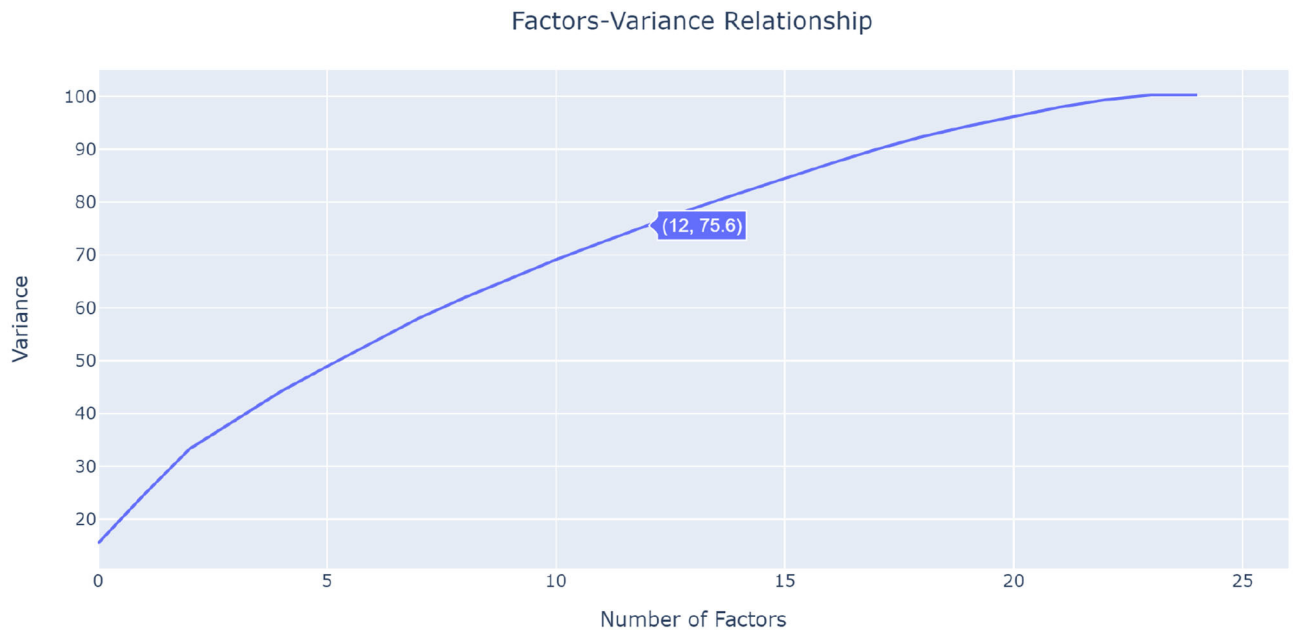

FIGURE 2 | Number of dimensions needed to account for 75% of the total variance.

step, IPCA identified 12 factors that are interacting with each other: age, parity, endometriosis, depression score, anxiety score, Pain Catastrophizing Scale, first pain, pain duration, infertility, CPP, and Endometriosis Health Profile. These 12 factors are located within the first four PCs (PC1–4), which are located in the interaction zone in **Figure 1**. The IPCA algorithm found that there is no interaction between PCs and factors after PC4; therefore, it used only the factors associated with the first four PCs. This was confirmed by running a cumulative sum for all PCs. **Figure 2** visualizes the cumulative sum of PCs and shows that 12 dimensions (i.e., PCs) are needed to account for 75% of the total variance, which is above the 70% cut-off point (14) for determining the optimal number of PCs. Note that the non-interaction zone shown in **Figure 1** contains only individual factors that are not interacting with other factors.

The last step of IPCA shows the interactions between the previously determined 12 factors. As shown in **Figure 3**, the first column shows significant correlation-based interactions between PC1 and CPP, Endometriosis Health Profile, Pain Catastrophizing Scale, anxiety score, and depression score. Chronic pelvic pain, Endometriosis Health Profile, Pain Catastrophizing Scale, anxiety score, and depression score factors move in the same direction. Clinically, this suggests that higher CPP severity, worse quality-of-life, and more anxiety, depression, and pain catastrophizing, all correlate with each other.

In the second column age, and age at first pain, and parity are moving together, which is another kind of AI produced by IPCA, and indicates that younger age, earlier age at first pain, and nulliparity are all correlated. In contrast, these variables are strongly inversely correlated with infertility, which suggests that having never tried for pregnancy is associated with younger age, earlier age at first pain, and nulliparity. The third column shows that endometriosis and pain duration move in the same direction, and both are inversely correlated with age at first pain. This is another level of AI produced by IPCA, and demonstrates that those with younger age at first pain have a longer pain duration and also are more likely have a confirmed diagnosis of endometriosis. PC4–12 show no interactions between factors.

The interaction based on hierarchical clustering of factors is shown in the dendrogram. A hierarchy is built that progressively merges the independent factors to generate clusters. The 25 factors were used, and the process works based on determining how close each set of two factors are. The factors were clustered according to their similarity, as shown in **Figure 4**. By visually inspecting **Figure 4** shows that the changes in the CPP and Endometriosis Health Profile are similar and are both clustered with anxiety score, depression score, and Pain Catastrophizing Scale. This is a kind of AI recommendation produced by HCA, and it suggests that the anxiety score, depression score, and Pain Catastrophizing Scale are good correlates for CPP and Endometriosis Health Profile. This finding is in agreement with the IPCA finding, as shown in **Figure 3**.

As can be seen in **Figure 4**, the most interesting recommendation produced using HCA is clustering alcohol, ethnicity and infertility as one group, such that infertility was correlated with non-Caucasian ethnicity and less alcohol use. Note that HCA was able to detect a non-linear correlation compared to the traditional linear bivariate correlation that was

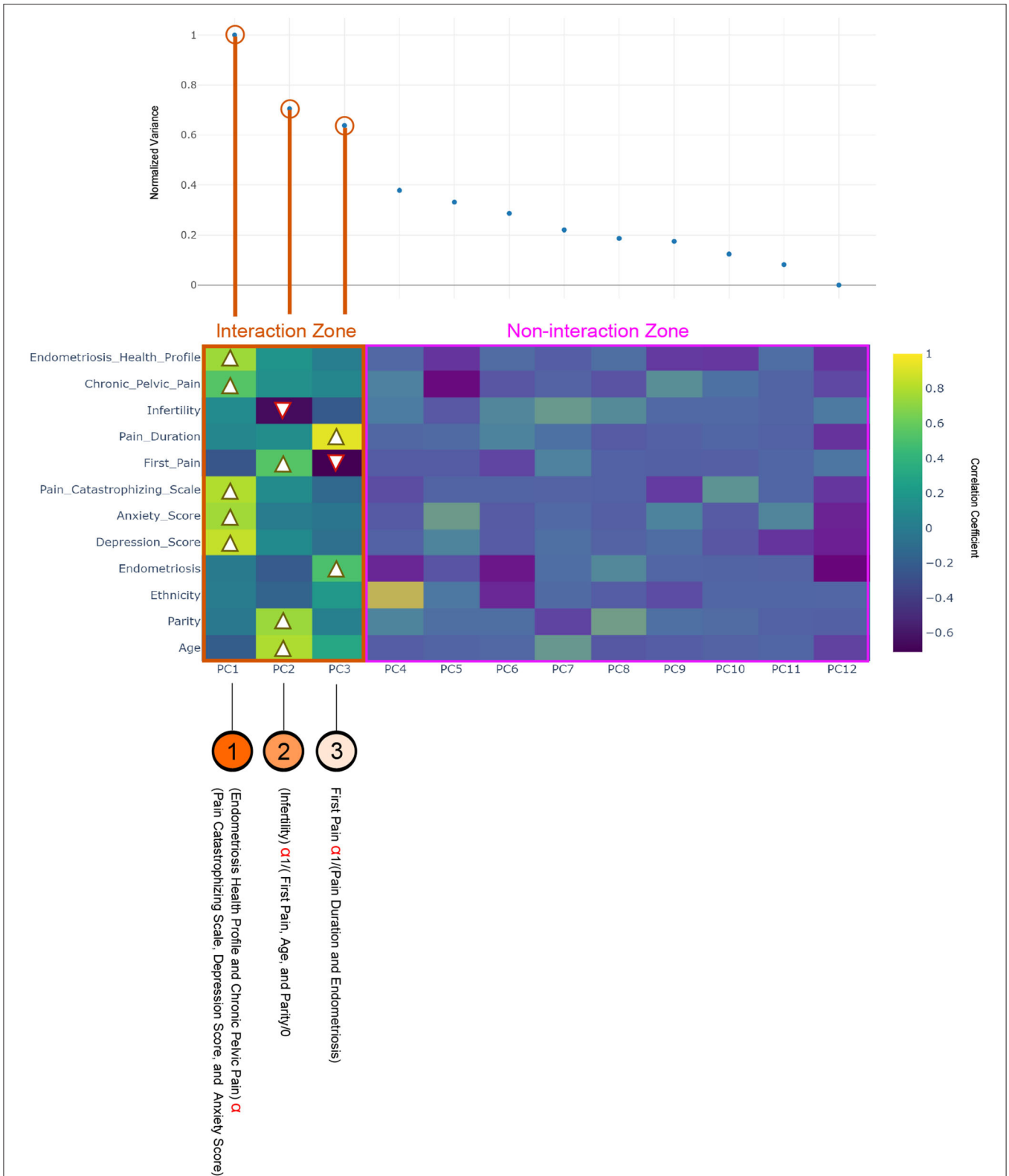

FIGURE 3 | Heat map of the interaction principle component analysis (IPCA) between 12 factors using the Pelvic Pain and Endometriosis database, and the black arrow indicates the direction of the correlation. The interaction zone refers to the number of principle components (PCs) where two or more factors are correlated, while the non-interaction zone refers to the number of PCs that have only single factor that is not correlated with any other factors.

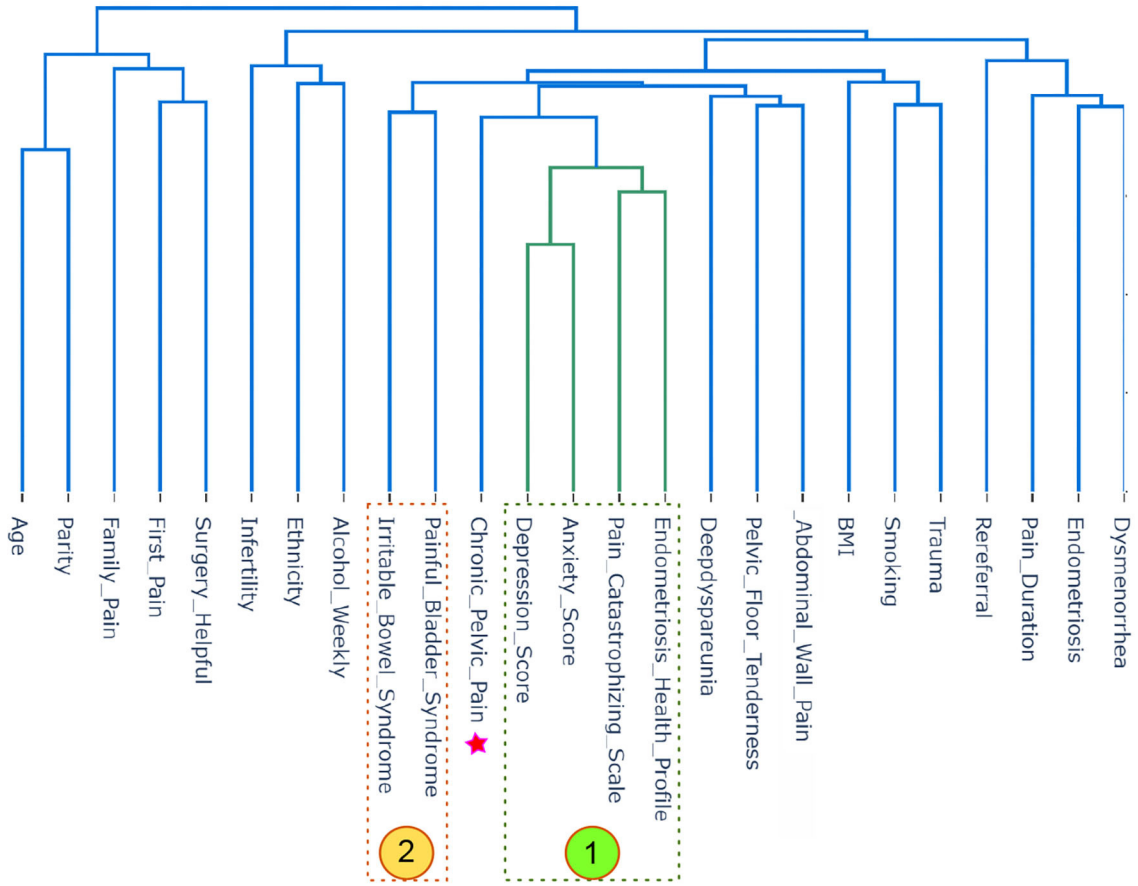

FIGURE 4 | Hierarchical cluster analysis (HCA) analysis of all 25 factors in the Pelvic Pain and Endometriosis database. The star represents the factor of interest, which is chronic pelvic pain (CPP). Four factors in zone #1 are directly associated with CPP, while two factors in zone #2 are indirectly associated with CPP.

not able to detect a correlation between ethnicity and alcohol (i.e., $r = 0.158$), ethnicity and infertility (i.e., $r = 0.068$), and alcohol and infertility (i.e., $r = 0.065$).

Interestingly, as shown in **Figure 4**, HCA grouped BMI with trauma and smoking, suggesting a correlation between the three (patients who are smoking and have been traumatized have higher BMI in this database). In fact, HCA showed that trauma and smoking are together directly associated with BMI. Note that HCA was able to detect a non-linear correlation compared to the traditional linear bivariate correlation that was not able to detect a correlation between trauma and BMI (i.e., $r = 0.0176$), trauma and smoking (i.e., $r = 0.217$), and BMI and smoking (i.e., $r = 0.108$).

The third geometrical interaction for our factors is CBC, which represents an alternative clustering method. Initially, CBC requires the desired number of clusters to process the data. We tested the inertia and found that the ideal number of clusters that reduces the distance between factors and their centroids is five. Then CBC was set up with five clusters, with respect to CPP, CBC clustered anxiety score, Pain Catastrophizing Scale, depression score, and Endometriosis Health Profile as one cluster, the first cluster on the left side of **Figure 5**. This finding is in agreement with the IPCA and HCA findings, as shown in **Figures 3**, **4**, respectively.

Also, CBC clustered four factors (age, family history of chronic pain, parity, and alcohol use per week) as one cluster with respect to CPP. Note that IPCA and HCA confirmed the correlation between age and parity (older age and higher parity/more deliveries), and CBC is in agreement with this finding. In addition, CBC clustered nine factors (BMI, infertility, trauma, deep dyspareunia, smoking, irritable bowel syndrome, painful bladder syndrome, pelvic floor tenderness, and abdominal wall pain) as one cluster with respect CPP. Note that CBC clustered BMI with trauma and smoking, confirming the effects of BMI on smoking and trauma, which is in agreement with HCA finding. Moreover, CBC clustered age when first experienced pain with surgery being helpful (younger age when first experienced pain associated with surgery having been helpful).

It is worth mentioning that IPCA found CPP to be interacting with Pain Catastrophizing Scale, anxiety score, depression score, and Endometriosis Health Profile. HCA showed an indirect (placing both in the same group, but not close to each other) association between irritable bowel syndrome, painful bladder syndrome, and CPP. Both irritable bowel syndrome and painful

FIGURE 5 | Centroid-based clustering (CBC) analysis of all 25 factors in the Pelvic Pain and Endometriosis database using five clusters. Four factors (Pain Catastrophizing Scale, anxiety score, depression score, and Endometriosis Health Profile) are formulated in one cluster with respect to chronic pelvic pain (CPP). Also, four factors (age, family history of chronic pain, parity, and alcohol use per week) are formulated in one cluster with respect to CPP. Nine factors [body mass index (BMI), infertility, trauma, deep dyspareunia, smoking, irritable bowel syndrome, painful bladder syndrome, pelvic floor tenderness, and abdominal wall pain] are formulated in one cluster with respect to CPP. Interestingly, age at first pain and history of surgery are correlated with each other with respect to CPP.

bladder syndrome were placed on the left side of CPP in a different group, as shown in **Figure 4**. Interestingly, HCA showed an association between abdominal wall pain, pelvic floor tenderness, and deep dyspareunia. This points toward the importance of musculoskeletal contributors (abdominal wall trigger points and myofascial pelvic pain syndrome of the pelvic floor) to sexual pain.

4. DISCUSSION

In this study, we utilized ML approaches to characterize factors that are correlated with CPP. To achieve this, we compared our results with those of a previously published study on the same dataset. The previous study (3) from our group on independent associations with factors suggested that seven factors were correlated with CPP: BMI, abdominal wall pain, pelvic floor tenderness, Pain Catastrophizing Scale, painful bladder syndrome, smoking, and history of adult trauma. However, these results showed the independent importance of each factor for chronic pelvic pain assessment. Our simultaneous analysis using INTENSE found that CPP and Endometriosis Health Profile are correlated with depression score, anxiety score, and Pain Catastrophizing Scale.

It was notable that endometriosis was not associated with chronic pelvic pain, as reported in our previous study (3) using regression analyses. However, in this current study, IPCA found an interesting collective relationship, where PC3 shows that those with younger age at first pain are more likely to have had surgery for endometriosis, which was reported as helpful. This was an interesting kind of AI observation, produced using IPCA. Clinically this makes sense, as patients with earlier onset pain and longer pain duration are more likely to undergo surgery to confirm the diagnosis and treat the endometriosis.

Our ML approach was also able to identify other unique relationships that were not apparent with routine regression analyses on the same dataset (3). For example, higher BMI was associated with a history of life trauma (15) and smoking (**Figure 3**). While the factors underlying this relationship are complex, one hypothesis is that life trauma could predispose to smoking as well as lifestyle habits that give risk to obesity. This hypothesis warrants further study.

Another interesting finding was that HCA clustered abdominal wall pain and myofascial pelvic pain of the pelvic floor musculature with deep dyspareunia (sexual pain) (**Figure 4**). This points to the importance of musculoskeletal factors in the etiology of sexual pain specifically, among women with pelvic

pain. The same relationship with musculoskeletal factors was not seen for dysmenorrhea (menstrual cramps), indicating that dysmenorrhea may have a different pathophysiology compared to sexual pain. These unique aspects of the etiology of different types of pelvic pain, discovered using ML, also warrant future study.

A limitation of the study is the inherent heterogeneity of chronic pelvic pain, where multiple underlying diagnoses can be present. While the sample size (> 500) helps to capture this heterogeneity in part, additional multi-center research is needed with even larger sample sizes given the complex multifactorial nature of chronic pelvic pain.

5. CONCLUSION

In this study, we have described our evaluation of the impact of chronic pelvic pain on various factors using machine learning approaches. INTENSE can to detect complex relationships between different factors for chronic pelvic pain, without the need for any previous training or knowledge, and is a completely unsupervised interaction method. The results of the ML methods showed agreement on the significant correlation between chronic pelvic pain and Endometriosis Health Profile-30, depression score, anxiety score, and Pain Catastrophizing Scale. Other unique relationships were also identified with ML, which provide data to drive future research.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: available upon request. Requests to access these datasets should be directed to PY, paul.yong@vch.ca.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of British Columbia. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ME, CA, CW, MB, and PY conceived the study and drafted the manuscript. ME developed the INTENSE algorithm. All authors approved the final manuscript.

FUNDING

This work was supported by the Canadian Institutes of Health Research, grant numbers: CIHR MOP-142273 and PJT-156084. PY was supported by a Health Professional Investigator Award from the Michael Smith Foundation for Health Research.

REFERENCES

1. Jarrell JF, Vilos GA, Allaire C, Burgess S, Fortin C, Gerwin R, et al. Consensus guidelines for the management of chronic pelvic pain. *J Obstet Gynaecol Can*. (2005) 27:869–910. doi: 10.1016/s1701-2163(16)30993-8
2. Mathias SD, Kuppermann M, Liberman RF, Lipschutz RC, Steege JF. Chronic pelvic pain: prevalence, health-related quality of life, and economic correlates. *Obstet Gynecol*. (1996) 87:321–7. doi: 10.1016/0029-7844(95)00458-0
3. Yosef A, Allaire C, Williams C, Ahmed AG, Al-Hussaini T, Abdellah MS, et al. Multifactorial contributors to the severity of chronic pelvic pain in women. *Am J Obstet Gynecol*. (2016) 215:760.e1–14. doi: 10.1016/j.ajog.2016.07.023
4. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42:377–81. doi: 10.1016/j.jbi.2008.08.010
5. Yong PJ, Williams C, Houlihan E, Yager H, Britnell S, Lau B, et al. Development of a centre for interdisciplinary care of patients with pelvic pain and endometriosis. *BC Med J*. (2013) 55:244–7. https://bcmj.org/articles/development-centreinterdisciplinary-care-patients-pelvic-pain-andendometriosis
6. Yong P, Sutton C, Suen M, Williams C. Endovaginal ultrasound-assisted pain mapping in endometriosis and chronic pelvic pain. *J Obstet Gynaecol*. (2013) 33:715–9. doi: 10.3109/01443615.2013.821971
7. Drossman DA. The functional gastrointestinal disorders and the Rome III process. *Gastroenterology*. (2006) 130:1377–90. doi: 10.1053/j.gastro.2006.03.008
8. Meijlink JM. Interstitial cystitis and the painful bladder: a brief history of nomenclature, definitions and criteria. *Int J Urol*. (2014) 21:4–12. doi: 10.1111/iju.12307
9. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x
10. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archiv Intern Med*. (2006) 166:1092–7. doi: 10.1001/archinte.166.10.1092
11. Osman A, Barrios FX, Kopper BA, Hauptmann W, Jones J, O'Neill E. Factor structure, reliability, and validity of the pain catastrophizing scale. *J Behav Med*. (1997) 20:589–605. doi: 10.1023/A:1025570508954
12. Jones G, Kennedy S, Barnard A, Wong J, Jenkinson C. Development of an endometriosis quality-of-life instrument: the Endometriosis Health Profile-30. *Obstet Gynecol*. (2001) 98:258–64. doi: 10.1016/S0029-7844(01)01433-8
13. Elgendi M, Menon C. Machine learning ranks ECG as an optimal wearable biosignal for assessing driving stress. *IEEE Access*. (2020) 8:34362–74. doi: 10.1109/ACCESS.2020.2974933
14. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans R Soc A Math Phys Eng Sci*. (2016) 374:20150202. doi: 10.1098/rsta.2015.0202
15. Latthe P, Mignini L, Gray R, Hills R, Khan K. Factors predisposing women to chronic pelvic pain: systematic review. *BMJ*. (2006) 332:749–55. doi: 10.1136/bmj.38748.697465.55

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 Digital Health

ORIGINAL RESEARCH
published: 01 February 2021
doi: 10.3389/fdgth.2020.613725

Estimating a Sleep Apnea Hypopnea Index Based on the ERB Correlation Dimension of Snore Sounds

Limin Hou[1], Qiang Pan[1], Hongliang Yi[2]*, Dan Shi[1], Xiaoyu Shi[1] and Shankai Yin[2]*

[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China, [2] Department of Otolaryngology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

This paper proposes a new perspective of analyzing non-linear acoustic characteristics of the snore sounds. According to the ERB (Equivalent Rectangular Bandwidth) scale used in psychoacoustics, the ERB correlation dimension (ECD) of the snore sound was computed to feature different severity levels of sleep apnea hypopnea syndrome (SAHS). For the training group of 93 subjects, snore episodes were manually segmented and the ECD parameters of the snores were extracted, which established the gaussian mixture models (GMM). The nocturnal snore sound of the testing group of another 120 subjects was tested to detect SAHS snores, thus estimating the apnea hypopnea index (AHI), which is called AHI_{ECD}. Compared to the AHI_{PSG} value of the gold standard polysomnography (PSG) diagnosis, the estimated AHI_{ECD} achieved an accuracy of 87.5% in diagnosis the SAHS severity levels. The results suggest that the ECD vectors can be effective parameters for screening SAHS.

Keywords: apnea hypopnea index, correlation dimension, non-linear acoustic characteristics, snore sound, sleep apnea hypopnea syndrome

OPEN ACCESS

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Qinglin Meng,
South China University of Technology, China
Jian Guo,
RIKEN Center for Computational Science, Japan
Xiaodong Huang,
South China University of Technology, China

***Correspondence:**
Qiang Pan
qpan@laas.fr
Hongliang Yi
yihongl@126.com

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

Received: *03 October 2020*
Accepted: *18 December 2020*
Published: *01 February 2021*

Citation:
Hou L, Pan Q, Yi H, Shi D, Shi X and Yin S (2021) Estimating a Sleep Apnea Hypopnea Index Based on the ERB Correlation Dimension of Snore Sounds. Front. Digit. Health 2:613725. doi: 10.3389/fdgth.2020.613725

INTRODUCTION

Snoring is one of the most important symptoms of Sleep Apnea Hypopnea Syndrome (SAHS) and carries much information for diagnosing the upper airway disorder (1). Snoring sounds can be recorded by a non-contact microphone using acoustical property analysis for the screening of SAHS (2, 3). The pitch and spectral characteristics of snoring have been widely applied (4, 5). The total airway response for a snore was extracted to examine SAHS by a higher-order statistics algorithm (6). Multiclass classification of snoring was acquired on the acoustic analysis of snore sounds (7). A genetic algorithm was applied to select the better features that can be extracted from the time and spectral domains of full-night recordings to determine the Apnea Hypopnea Index (AHI) value (8). The rhythmic variations in the snores were described to assess the AHI (9). Hidden Markov models with Mel frequency cepstral coefficients (MFCC) were used to classify subjects into different ranges of the AHI (10). Our previous work used snore spectral information to estimate the AHI (11). Traditional time and frequency analysis and the classic method for snore sounds were adopted in the studies mentioned above.

However, the irregular and turbulent airflow that is produced within the upper airway tissue vibrations that cause the snore, such as the intensity of respiratory airflow, vibration on the soft palate, thick tongue root, and epiglottic hypertrophy, etc. could be non-linear (12).

It was suggested that linear analysis methods were limited and that more useful information could be obtained using chaos theory to analysis the snore (13, 14). The largest Lyapunov exponent and entropy were calculated to classify snore-related sounds with multiclass system (15).

In this paper, a new correlation dimension was proposed to analyze the non-linear properties of snoring sounds for automatic AHI prediction. In contrast to the conventional correlation dimension, the all frequency region was divided into multi-sub-bands on an equivalent rectangular bandwidth (ERB) scale, and the correlation dimensions were calculated in each sub-band. Therefore, ERB correlation dimension (ECD) vectors were extracted rather than a single correlation dimension. The gaussian mixture model (GMM) was applied to build the ECD vector models. Whole-night snore sounds of patients were detected in our experiments, and then, the AHI_{ECD} values were estimated. The early experimental studies have been published in Chinese journals, when the experiment is the number of 60 snorers in reference (16). This research continues to now increased to 120 snorers. In other words, the experimental results of adding 60 people dropped slightly from the original 90 to 87%. It illustrates the robustness of new features. This study further adds a comparison with the classic feature MFCC. Compared to the polysomnography (PSG) diagnosis, our non-linear features achieved higher accuracy than the MFCC based snore spectrum information in the severity levels of the SAHS. The ECD vectors were found to characterize various severity levels of snores.

ERB CORRELATION DIMENSION

The phase space reconstruction technique has been widely used in the field of chaos and fractal theory (17), and it has been used in some applications in medical and speech signals (18–20). It could be more comprehensive disclosure of snore implied information by transforming them to high-dimensional space. The general representation of a snore is a time series. Let a one-dimensional discrete series $s(n)$ be denoted by the snore signal, that is get by sampling rate Fs.

Based on the Takens embedding theorem (21), the phase space reconstruction could transform a one-dimensional time series into a high-dimensional phase space vector $Y \epsilon R^m$ as in Equation (1).

$$Y = [Y_1 \quad Y_2 \quad \cdots \quad Y_i \quad \cdots \quad Y_I]$$

Where

$$Y_i = [s(n_i) \quad s(n_i + \tau) \quad \cdots \quad s\left(n_i + (m-1)\tau\right)]^T \qquad (1)$$

Here, τ is the time delay, and m is the embedding dimension. The reconstruction vector Y is an m-dimensional vector with I phase points. The appropriate time delay was selected according to the autocorrelative function (AR function) (22). The time delay τ is an integer multiple of the sampling interval: $\tau = n/Fs$.

The purpose of choosing the embedding dimension is to make the original chaos attractor and the reconstructed attractor topology equivalent. We used the false nearest neighbor (FNN) method to determine the embedding dimension m (23). As the embedding dimension m increases, the orbit of the chaotic motion will gradually open, and the false nearest neighbors will be gradually eliminated, until the trajectory tends to be stable and the proper m is obtained (24). When the frame length was > 150 ms, the slope of the correlation integral curve increased very slowly. Finally in our snore work, the time delay of 0.75 ms and the embedding dimension of 15 were confirmed by the above method with a frame length of 150 ms.

The traditional correlation dimension has only a single parameter, it is difficult to make a more comprehensive analysis of complicated signals. Based on the ERB scale related to auditory perception (25, 26), several sub-bands were divided from the whole frequency band of the snore signal. Phase space reconstruction was performed in each of the sub-band signals of the snore. Then, the correlation dimension on these sub-bands were calculated, which obtained auditory sub-band ERB correlation dimension (ECD) vectors. The flow chart of extracting the ECD is shown in **Figure 1**.

Equation (2) is the correlation integral $C_q(I, r)$ of the qth subband, which calculates the probability that the distance of paired (Y_{iq}, Y_{jq}) is smaller than r.

$$C_q(I, r) = \frac{1}{I(I-1)} \sum_{i,j=1}^{I} \theta(r - |Y_{i,q} - Y_{j,q}|) \qquad (2)$$

Where $\theta(\cdot)$ is the Heaviside function, and if $x < 0$, $\theta(x) = 0$; if $x > 0$, then $\theta(x) = 1$. The correlation dimension D is estimated based on the ratio of the logarithm of the correlation integral and the logarithm of the distance r, as in Equation (3).

$$D_q = \frac{\ln \boldsymbol{C_q}(\boldsymbol{I}, \boldsymbol{r} + \Delta \boldsymbol{r}) - \ln \boldsymbol{C_q}(\boldsymbol{I}, \boldsymbol{r})}{\ln(\boldsymbol{r} + \Delta \boldsymbol{r}) - \ln \boldsymbol{r}} \qquad (3)$$

Therefore, the correlation dimension of the qth subband was calculated by Equations (2, 3) based on the Grassberger-Procaccia (GP) algorithm (24). Finally, we get the ECD vector by arranging and integrating ERB subband' correlation dimension, as in Equation (4).

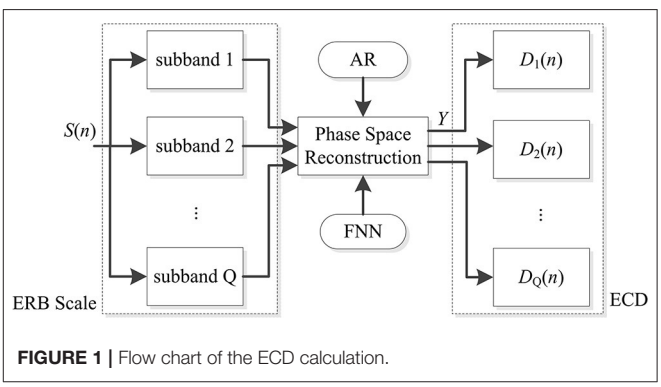

FIGURE 1 | Flow chart of the ECD calculation.

$$\mathbf{ECD}\,(n) = [D_1\,(n) \quad D_2\,(n) \quad \cdots \quad D_q\,(n) \quad \cdots \quad D_Q(n)]\,(4)$$

In this study, the gap of the adjacent sub-band was one bandwidth of the ERB scale, the frequency range of 60 Hz to 4 kHz was divided into 24 sub-bands, that is $Q = 24$, and the 24-dimensional vector of the auditory sub-band ECDs were extracted.

Moore and Glasberg proposed the relationship between frequency and ERB scale (25, 26), as in Equation (5).

$$ERB\,(f) = 6.23f^2 + 93.39f + 28.52$$

$$ERBS\,(f) = 11.17628 * \ln(1 + \frac{46.06538f}{f + 14678.49}) \qquad (5)$$

where f is physical frequency in kHz. ERB(f) is the calculated rectangular bandwidth of the equivalent filter in Hz. ERBS(f) is the ERB scale in physical frequency f in Hz.

SNORE ECD FEATURES

Snore Data

Snore sounds were recorded in the sleep monitoring laboratory in the Department of Otolaryngology of Shanghai Jiao Tong University Affiliated Sixth People's Hospital by a non-contact ambient microphone, and simultaneously, polysomnography (PSG) diagnosis was performed. The recording uses a non-contact microphone Sony EM-C10, which is hung on the head of the bed, about 30 cm away from the patient's nose and mouth. The recording sound card is Creative Audigy 4 Value, the desktop computer is Dell Inspiration 570, the recording software Adobe Audition 3.0, the sampling frequency is 8 kHz sampling, 16 bit quantization, and saved as WAV audio files. The recording duration is 7 h from 10:30 p.m. to 5:30 a.m. the next morning. In this test experiment, the half hour before the beginning and the end are removed, and 6 h of recording are used. The details of recording for snore sounds were the same as literature (11).

The AHI_{PSG} was the apnea hypopnea index as diagnosed by the gold standard PSG. The severity levels of SAHS were determined using the AHI value. The subjects with AHI $>$ 30, 15 $<$ AHI $\leq$ 30, 5 $\leq$ AHI $\leq$ 15, and AHI $<$ 5 were classified as severe (S), moderate (M), mild (L) SAHS, and non-SAHS (N), respectively (27).

The 213 subjects were consecutively recruited. In our experiment training phase, the snore episode was cut artificially from the sound of overnight recordings by a non-contact microphone on the bedhead, which included 93 subjects from 213. According to synchronized PSG nocturnal monitoring data, there were two types of snore episodes that were labeled. One was snoring sound labeled snore events by PSG diagnosis, which was only resounding and occurred periodically. The other was a loudly snoring sound appearance behind apnea or hypopnea events labeled by PSG. We called the former a simple snore (SIMP) and the latter a SAHS snore (SAHS). These are shown in **Figure 2**.

Another 120 subjects from 213 were as a test data set by their overnight recording of sounds. We removed the starting 30 min and the ending 30 min of recording. The remaining 6 h audio signal (11) were used for our test experiments as shown in **Table 1**.

The Largest Lyapunov Exponent

The largest Lyapunov exponent (LLE) of snores has been calculated to measure the rate of local divergence of nearby trajectories in the state space from dynamical systems theory (28). The LLE of all type snores are shown in **Figure 3**. The LLE of the four types of simple snore and SAHS snore are all positive. A few of the severe types have the Lyapunov exponent of SAHS snore $<$ 0, accounting for only 2.4% of the severe type of snoring episodes. The LLE of other types of snores did not appear negative, which also shows that the chaos of snoring is universal. This conclusion is consistent with other researches (13, 15, 29).

The mean of the LLE distribution of simple snore is greater than the LLE distribution of SAHS snore in same severity level. This phenomenon is common in the four types of snore signals. In moderate and severe levels, the difference between the two means of SIMP snore and SAHS snore is increasing. The results reveal that the orbital divergence speed of SIMP snore is greater than that of the SAHS snore, and is consistent with the other study (29). The LLE distribution suggests that unconscious airflow from simple snoring may have more freedom to roam,

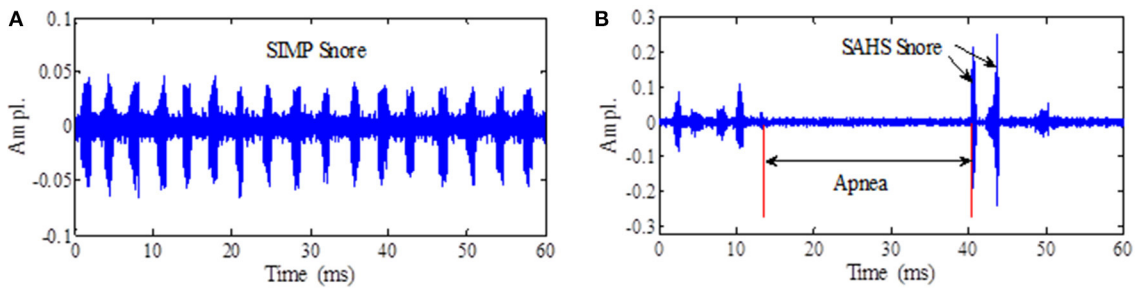

FIGURE 2 | (A) The simple snore wave. **(B)** The SAHS snore wave.

TABLE 1 | Snore data for training and test.

	N	L	M	S
Training Data				
Subjects (number)	10	23	24	36
Gender (M/F)	9/1	21/2	23/1	36/0
Age (years)	42.1 ± 8.5	46.2 ± 12.4	40.4 ± 13.2	45.6 ± 12.5
AHI_{PSG} (events/h)	2.4 ± 1.4	10.8 ± 3.5	24.5 ± 3.8	57.0 ± 16.8
SIMP Episodes (number)	339	919	480	430
SAHS Episodes (number)	55	376	480	916
Test Data				
Subjects (number)	30	30	30	30
Gender (M/F)	19/11	26/4	25/5	27/3
Age (years)	29.9 ± 8.6	40.7 ± 12.4	43.0 ± 13.8	38.9 ± 11.7
AHI_{PSG} (events/h)	1.9 ± 1.6	9.3 ± 3.0	22.2 ± 4.3	62.5 ± 17.8
Recording length (minutes)	360 × 30	360 × 30	360 × 30	360 × 30

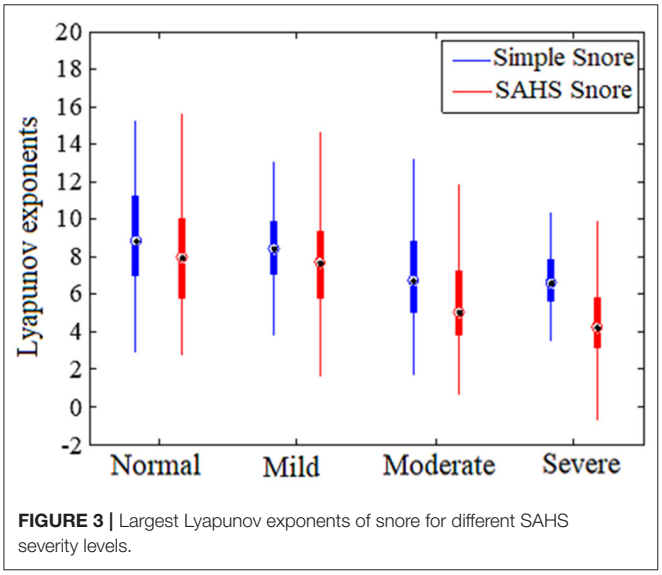

FIGURE 3 | Largest Lyapunov exponents of snore for different SAHS severity levels.

while SAHS snoring may form a certain trend of airflow after being squeezed in the narrow upper airway.

ECD Calculation

According to the illustration in **Figure 1**, the ECD of the snore from the training data in **Table 1** were calculated. The distribution of the ECD vectors in each sub-band of the SIMP snore and SAHS snore of the N, L, M, and S levels, respectively, are shown in **Figures 4A–D**. The ECD vectors distinctively increased with the aggravation of the SAHS severity level in the middle and high-frequency sub-bands. Moreover, the distributions of the ECD vectors were not exactly the same for the SIMP snores and the SAHS snores at the same severity level, and the ECDs of the SAHS snores were always higher than those of the SIMP snores.

In our study, the SIMP and SAHS snores of four levels (N, L, M, and S) were modeled using the Gaussian Mixture Model (GMM), which formed eight types, including N-simp, N-sahs, L-simp, L-sahs, M-simp, M-sahs, S-simp, and N-sahs. The ECDs of the training data in **Table 1** were extracted to model eight GMMs for the training phase (30), and are showed in **Figure 5**.

RESULTS AND DISCUSSION

Results

Mixture Number of GMM were assigned 2, 12, 12, and 8 for both SIMP and SAHS snore of N, L, M, and S level, respectively. GMM was solved by expectation-maximization (EM) algorithm. Two-fold cross-validation method was employed to evaluate the performance of clustering and classification of GMM regarding training data in **Table 1**. For each type of snore, the rate of being classified as different types is shown in **Figure 6.**

According to the PSG clinical diagnosis definition, AHI is the number of respiration events per hour of sleep. The ECD-calculated AHI is AHI_{ECD} as in Equation (6).

$$AHI_{ECD} = \frac{Number\ of\ sleep\ respiratory\ events}{Duration\ of\ nocturnal\ sleep}\ \textbf{events/h} \quad (6)$$

Figure 5 shows the testing phase, there were another 120 participants for the testing data, which consisted of 30 subjects for each severity level among N, L, M, and S in **Table 1**. Firstly, automatic endpoint detection was performed for snore signals of whole-night recordings to detect the snore sounds (16, 31). Thus, we obtained candidate respiratory events based on the unique rhythm of snores (16, 31). Then, the ECD vectors of these candidate respiratory events were extracted, and we calculated the probabilities of matching with eight GMMs. On the basis of the Bayesian maximum posterior probability rule, the maximum posterior probability winner among the eight GMMs was the snore type. When some snore episodes in a candidate respiratory event were classified as any SAHS snores among N, L, M, and S levels by the GMM, that candidate respiratory event was a true sleep respiratory event. Finally, the AHI_{ECD} score was estimated by the number of sleep respiratory events and the nocturnal sleep duration, as in Equation (6).

In the same way, we extracted another feature set that is MFCC, and estimated the AHI_{MFCC} score also. The MFCC is a classical feature and widely used automatic speech recognition. All experiment results in precision and recall were listed in **Table 2**.

Comparisons of AHI_{ECD}, AHI_{MFCC}, and AHI_{PSG} values of each subject, consistency with the severity of the gold standard PSG diagnosis was correctly screened. There are 120 testers in **Table 2**, including 30 people of four different severity levels. As a result of the MFCC classic feature test, 30 people who are non-ASHS can correctly estimate. Twelve of the 30 mild patients were incorrectly classified as non-SAHS or moderate SAHS types. Eleven of 30 moderate patients were wrongly assorted as mild

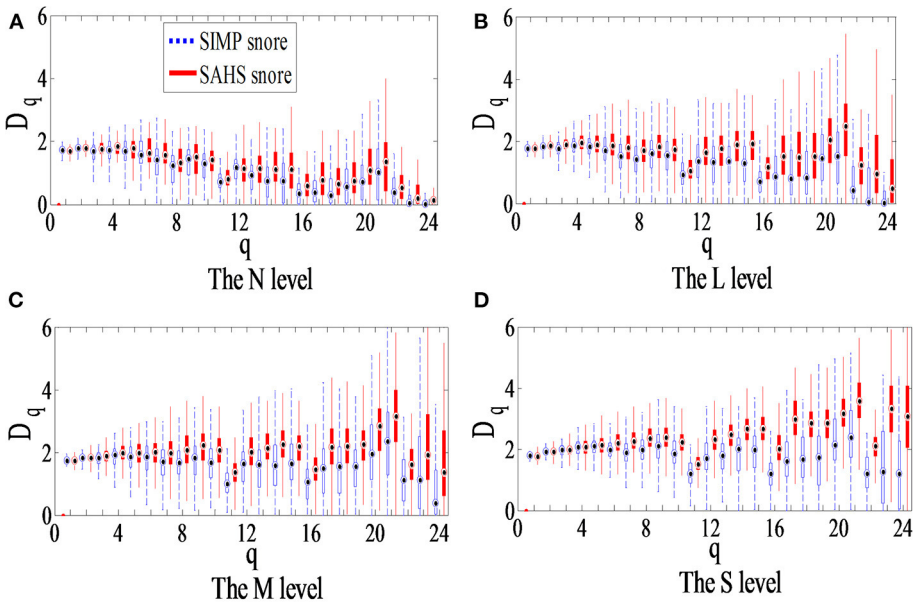

FIGURE 4 | Box plot of ECD vector for different SAHS severity levels. **(A)** The N level, **(B)** The L level, **(C)** The M level, and **(D)** The S level.

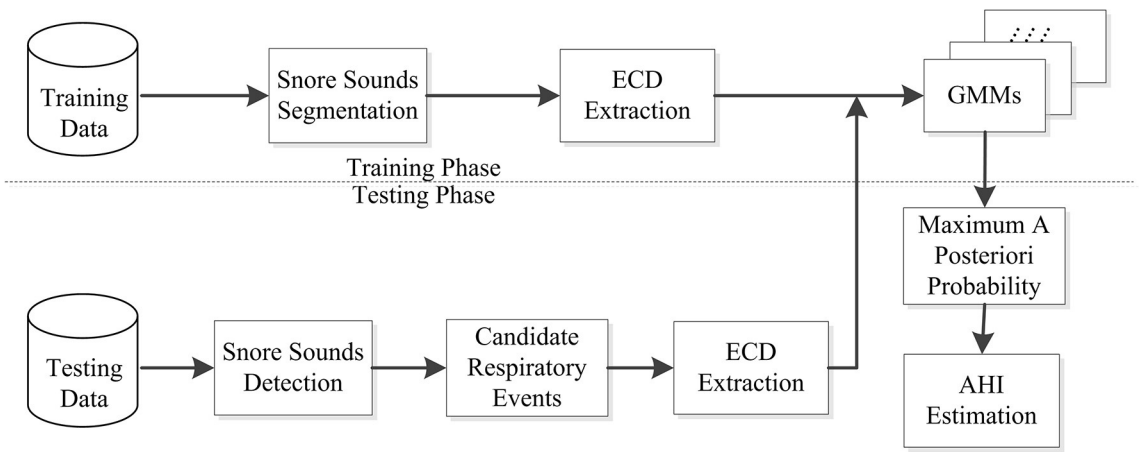

FIGURE 5 | Flow chart of the snore training and testing system.

or severe SAHS types. Thirty patients with severe patients, one of whom were mistakenly estimated as moderate SAHS patients. Compared to AHI_{PSG} in the diagnosis of the SAHS severity level, and the AHI_{MFCC} estimation achieved the mean precision and recall of 79.25 and 80.00%, respectively, as shown in **Table 2**.

As a result of the ECD feature test, two out of 30 non-SAHS people were mistakenly estimated to be mild patients. Thirty patients with mild patients, eight of whom were incorrectly estimated to be non-SAHS or moderate patients. Of the 30 patients with moderate disease, five of them were incorrectly estimated to be mild or severe SAHS patients. Thirty people with severe patients were correctly estimated. The AHI_{ECD} estimation using our proposed method achieved, respectively, the mean precision and recall of 87.74 and 87.50% compared to AHI_{PSG} in the diagnosis of the SAHS severity level as shown in **Table 2**.

The precision and recall of AHI_{ECD} are higher than AHI_{MFCC} in mild and moderate levels especially.

Comparisons results of AHI_{ECD} and AHI_{MFCC}, both features are good at both ends (i.e., non-SAHS and severe patients). However, for patients with mild SAHS and moderate SAHS, the number of errors by using MFCC is higher than ECD feature. The precision and recall of AHI_{ECD} are higher than AHI_{MFCC} in mild

and moderate level especially. New fractal features achieve better results than classical spectral features. Relative literature (16), this work increased the number of patients in test experiments from 60 to 120 and adds to compares them with the classic feature MFCC. Therefore, the experimental results of this paper are almost the same as the initial experiments, once again confirming the advantages of the new features.

Discussion

Most of the previous studies of detecting snore used the acoustic characteristics of the speech signal of the active pronunciation (4–11), that is limit. Because the snoring contained more breathing sounds than speech. This airflow has more randomness and is generated by passive vocalization. Thereby, we proposed a new feature from the chaos and fractal theory to characterize the irregular extent of snore.

The AHI_{ECD} by new features is closer to clinical diagnosis results than AHI_{MFCC} by conventional parameters. The distribution scatter plot of AHI_{PSG} vs. AHI_{MFCC}, AHI_{ECD} is shown in **Figure 7**. The black asterisk represents the result of PSG diagnosis AHI_{PSG}, the green pentagram represents AHI_{MFCC}, and the purple circle represents AHI_{ECD}, and the red dotted line represents the boundary of different severity. The cohen's kappa coefficient of AHI_{ECD} and AHI_{PSG} consistency is 0.833, and AHI_{MFCC} and AHI_{PSG} consistency is 0.733.

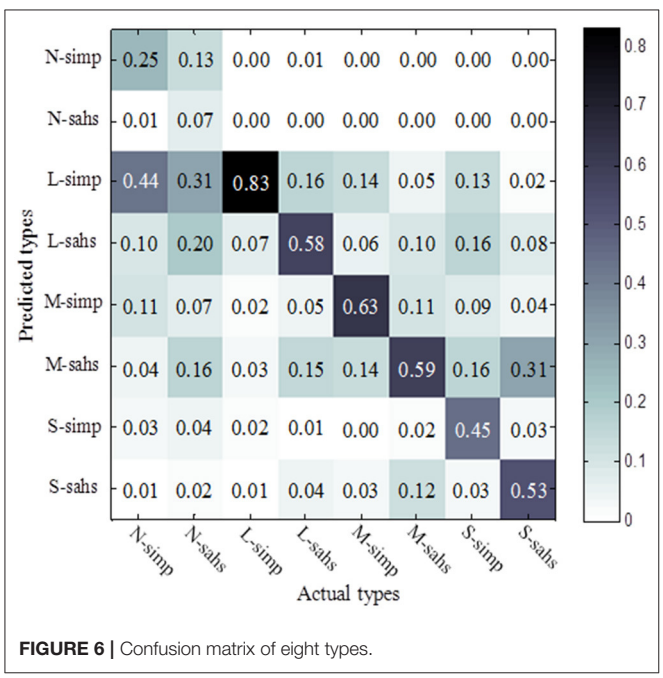

FIGURE 6 | Confusion matrix of eight types.

The Bland-Altman-plot is depicted in **Figure 8**. The ordinate represents the difference of AHI_{PSG} and AHI_{ECD} with pinkish, the difference of AHI_{PSG} and AHI_{MFCC} with green. The mean and variance of difference of AHI_{ECD} and AHI_{PSG} were smaller than AHI_{MFCC}. Compared with PSG, 92.50% (111/120) of AHI_{ECD} falls within the consistency limit of 1.96 times variance, higher than 88.33% (106/120) of AHI_{MFCC}. This further suggests AHI_{ECD} estimated by ERB correlation dimensions is more accordance with AHI_{PSG} than AHI_{MFCC}.

In terms of the severity of SAHS, especially the N-type and the severe type, their frequency spectrum has obvious differences. Therefore, the MFCC parameters maintain a good performance for the judgment of these two types. The ECD feature is the same. However, for the intermediate types of mild and moderate, the accuracy of MFCC's outcome drops sharply. The ECD we proposed is much better than MFCC in these two types.

This paper presents a method to measure the degree of disorder of the snoring signal like noise, which were new features called the ECD vectors. The correlation dimensions of the high frequency sub-bands were larger than those of the low-frequency sub-bands in ECD vector. The maximal correlation dimension appeared in the 21st ERB sub-band as shown in **Figure 2**. This finding suggested that the SAHS snores contain much more irregular and fast-changing components in the high frequency range. The ECD vectors could reveal information that is consistent with the characteristic of the time-domain waveform of the snore. When the upper airway is blocked, the airway becomes shorter. When the upper airway rushes open, the airflow in the narrow area is squeezed, and the turbulent airflow is released. However, the snore spectrum is attenuated to a smaller magnitude in the high frequency range, and thus, it is difficult to give an appropriate description for the mild and moderate level of a snore, which could be to too small to distinguish different level. These non-linear methods are expected to provide useful information for better understanding of irregular snoring sounds (13, 14). MFCC includes only magnitude of snore spectral, but our ECD feature completes information in snore sound. When the upper airway is obstructed, the shortening of the airway leads to an increase in the medium and high frequency components, the airflow in the narrow area is squeezed, and some the rapid

TABLE 2 | Precision and recall of AHI_{MFCC} and AHI_{ECD} compared AHI_{PSG}.

Levels	N	L	M	S	Total Correct	Mean
Subjects(number)	30	30	30	30	120	
Precision of AHI_{MFCC}	85.71%	64.28%	70.37%	96.66%	96	79.25%
Recall of AHI_{MFCC}	100%	60%	63.33%	96.66%	96	80.00%
Precision of AHI_{ECD}	100%	81.48%	75.75%	93.75%	105	87.74%
Recall of AHI_{ECD}	93.33%	73.33%	83.33%	100%	105	87.50%

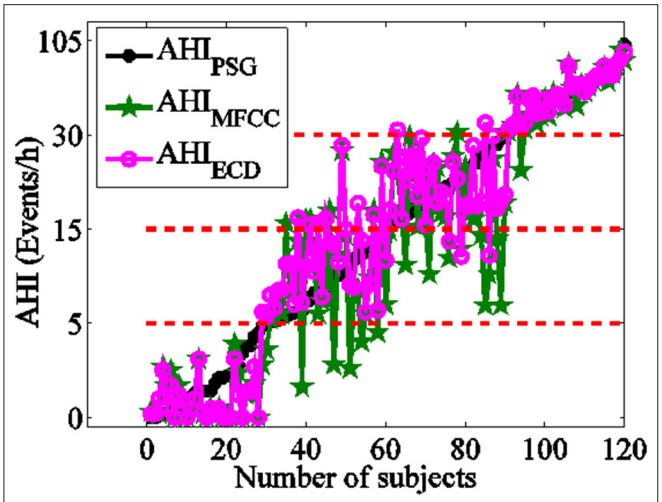

FIGURE 7 | The compared between AHI_{ECD}, AHI_{MFCC}, and AHI_{PSG}.

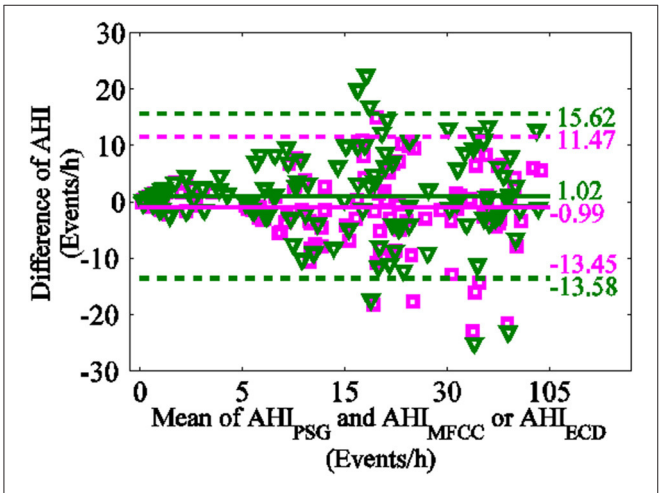

FIGURE 8 | Bland-Altman-plot (the difference of AHI_{PSG} and AHI_{ECD} with pinkish, the difference of AHI_{PSG} and AHI_{MFCC} with green).

change component increases. The Fourier transform shows a characteristic of global decline and local prominence. Compared with MFCC, the ERB enlarged partially and highlighted the anomaly of the mid and high frequency components.

Inspired by the non-linear frequency scale and MFCC characteristics of the Mel spectrum, we use ERB to set the sub-band frequency interval to 8, 4, 2, and 1 ERB bandwidth, so that 3, 6, 12, and 24 subbands are obtained severally in formula (5). The obvious differentiation the snoring sounds of different severity appears when dividing three sub-bands but the details are not enough to distinguish well. As the number of subbands increases, more and more details provide a richer diversity of different severity level of SAHS. According to the distribution of the auditory filter, as it is divided into about 20 subbands in 4 kHz, a set of features is more effective. We adopted one ERB bandwidth and 24 subbands are obtained in formula (5).

No matter how many take the ERB scale, ECD features exhibits SAHS severity is directly proportional to the relationship, that is, the more severe the SAHS, the faster the ECD rises in the middle and high frequency regions, shown as in **Figure 4**.

However, the calculation of the correlation dimension was time-consuming. This limitation requires us to optimize the algorithm for the correlation dimension. The nature of the correlation dimension on the number of more subbands may need further study.

CONCLUSIONS

Based on the previous experiment, we prove the chaotic nature of snoring sound by the LLE and perfect a new method for estimating the AHI value of SAHS using the correlation dimension vector for snore sounds, which was superior to the conventional spectrum analysis. The ECD vectors might be closely related to the SAHS severity level and reveal the effect of different SAHS severities on the upper airway. The correlation dimension of the sub-bands reveals the inherent information of the mid and high frequencies, while the Fourier transform has its limitations. Chaos provides many quantitative parameters for exploring the nature of this internal information. It could be a study about correlation between fractal dimension and internal physical properties of sleep respiratory sound. There is a positive effect on the development of a medical supplementary diagnosis and in-home healthcare in the internet era.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Department of Otolaryngology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All the co-authors contributed to this work. LH was responsible for all scheme design and about correlation dimension algorithm work of this paper. QP organized the writing work of this paper and MFCC programming work. HY and SY participated on the clinical diagnosis by PSG and actively discussed in this work. DS programmed the EM algorithm work of this paper. XS programmed the result analysis work of this paper.

FUNDING

This study was funded by the Science and Technology Commission of Shanghai Municipality (No. 13441901600) and National Natural Science Foundation of China (Nos. 61525203 and 61572308).

ACKNOWLEDGMENTS

The authors thank Wang Yong, Wang Chen, Tong Chao, Zhang Weitao, Liu Huancheng, graduate students in Shanghai University for data collection and segmentation. We appreciate the contribution of all subjects who participated in this study.

REFERENCES

1. Pevernagie D, Aarts RM, De Meyer M. The acoustics of snoring. *Sleep Med Rev.* (2010) 2:131–44. doi: 10.1016/j.smrv.2009.06.002
2. Dafna E, Tarasiuk A, Zigel Y. Automatic detection of whole night snoring events using non-contact microphone. *PLoS ONE.* (2013) 12:e84139. doi: 10.1371/journal.pone.0084139
3. Camacho M, Robertson M, Abdullatif J, Certal V, Kram YA, Ruoff CM, et al. Smartphone apps for snoring. *J Laryngol Otol.* (2015) 10:974–9. doi: 10.1017/S0022215115001978
4. Yang Y, Qin Y, Haung W, Peng H, Xu H. Acoustic characteristics of snoring sound in patients with obstructive sleep apnea hypopnea syndrome. *J Clin Otorhinolaryngol Head Neck Surg.* (2012) 8:360–3.
5. Ng AK, Koh TS, Baey E, Lee TH, Abeyratne UR, Puvanendran K. Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea? *Sleep Med.* (2008) 8:894–8. doi: 10.1016/j.sleep.2007.07.010
6. Karunajeewa AS, Abeyratne UR, Hukins C. Multi-feature snore sound analysis in obstructive sleep apnea–hypopnea syndrome. *Physiol Meas.* (2011) 1:83–97. doi: 10.1088/0967-3334/32/1/006
7. Solà-Soler J, Fiz JA, Morera J, Jané R. Multiclass classification of subjects with sleep apnoea–hypopnoea syndrome through snoring analysis. *Med Eng Phys.* (2012) 9:1213–20. doi: 10.1016/j.medengphy.2011.12.008
8. Dafna E, Tarasiuk A, Zigel Y. OSA severity assessment based on sleep breathing analysis using ambient microphone. In: *35th Annual International Conference of EMBC.* Osaka:IEEE (2013). p. 2044–7. doi: 10.1109/EMBC.2013.6609933
9. Ben-Israel N, Tarasiuk A, Zigel Y. Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults. *Sleep.* (2012) 9:1299–305. doi: 10.5665/sleep.2092
10. Herath DL, Abeyratne UR, Hukins C. Hidden Markov modelling of intra-snore episode behavior of acoustic characteristics of obstructive sleep apnea patients. *Physiol Meas.* (2015) 12:2379–404. doi: 10.1088/0967-3334/36/12/2379
11. Xu H, Wei S, Yi H, Hou L, Zhang C, Chen B, et al. Nocturnal snoring sound analysis in the diagnosis of obstructive sleep apnea in the Chinese Han population. *Sleep Breath.* (2015) 2:599–605. doi: 10.1007/s11325-014-1055-0
12. Azarbarzin A, Moussavi Z. Nonlinear properties of snoring sounds. *Int Conf Acoust.* (2011) 4316–9. doi: 10.1109/ICASSP.2011.5947308
13. Sakakura A. Acoustic analysis of snoring sounds with chaos theory. *Int Congr Ser.* (2003) 1257:227–30. doi: 10.1016/S0531-5131(03)01170-1
14. Mikami T. Detecting nonlinear properties of snoring sounds for sleep apnea diagnosis. In: *2nd International Conference on Bioinformatics and Biomedical Engineering.* Shanghai:IEEE (2008). p. 173–6. doi: 10.1109/ICBBE.2008.621
15. Ankişhan H, Yilmaz D. Comparison of SVM and ANFIS for snore related sounds classification by using the largest Lyapunov exponent and entropy. *Comput Math Methods Med.* (2013) 2013:238937. doi: 10.1155/2013/238937
16. Hou LM, Shi D, Liu HC, Zhang WT. Screening of SAHS snore based on ERB correlation dimension. *J Appl Sci.* (2017) 2:181–92. doi: 10.3969/j.issn.0255-8297.2017.02.005
17. Grassberger P, Procaccia I. Dimension and entropy of strange attractors from a fluctuating dynamic approach. *Phys D.* (1984) 1–2:34–54. doi: 10.1016/0167-2789(84)90269-0
18. Hekmatmanesh A, Mikaeili M, Sadeghniiat-Haghighi K, Wu HP, Nazeran H. Sleep spindle detection and.prediction using a mixture of time series and chaotic features. *Adv Electr Electron Eng.* (2017) 3:435–47. doi: 10.15598/aeee.v15i3.2174
19. Subramaniam K, Clark AR, Hoffman EA, Tawhai MH. Metrics of lung tissue heterogeneity depend on BMI but not age. *J Appl Physiol.* (2018) 2:328–39. doi: 10.1152/japplphysiol.00510.2016
20. Hou LM, Deng DC, Wang SZ. Improvement of speaker identification performance using nonlinear feature. *Patt Recog Artif Intell.* (2006) 6:776–81. doi: 10.1360/crad20061223
21. Takens F. Detecting strange attractors in turbulence in dynamical systems and turbulence, Warwick (1980). *Lect Notes Math.* (1981) 898:366–81. doi: 10.1007/BFb0091924
22. Rosenstein MT, Collins JJ, De Luca CJ. Reconstruction expansion as a geometry-based framework for choosing proper delay times. *Phys D.* (1994) 1–2:82–98. doi: 10.1016/0167-2789(94)90226-7
23. Cao L. Determining minimum embedding dimension from scalar time series. *Springer US.* (2002) 2:43–60. doi: 10.1007/978-1-4615-0931-8_3
24. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. *Phys D.* (1983) 1–2:189–208. doi: 10.1016/0167-2789(83)90298-1
25. Moore BC, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J Acoust Soc Am.* (1983) 3:750–3. doi: 10.1121/1.389861
26. Glasberg BR, Moore BCJ. Derivation of auditory filter shapes from notched-noise data. *Hear Res.* (1990) 1–2:103–38. doi: 10.1016/0378-5955(90)90170-T
27. Editorial Board of Chinese Journal of Ot, Subspecialty Group of Pharyngolaryngolog. Guideline for the diagnosis and surgical treatment of obstructive sleep apea hypopnea syndrome. *Chin J Otorhinolaryngol Head Neck Surg.* (2009) 2:95–6. doi: 10.3760/cma.j.issn.1673-0806.2009.02.003
28. Rosenstein MT, Collins JJ, De Luca CJ. A practical method for calculating largest Lya punov exponents form small data sets. *Phys D.* (1993) 65:117–34. doi: 10.1016/0167-2789(93)90009-P
29. Kizilkaya M, Ari F, Demirgunes DD. Analysis of snore sounds by using the largest Lyapunov exponent. *J Concr Appl Math.* (2011) 9:146–53.
30. Nakamura K, Toda T, Saruwatari H, Shikano K. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Commun.* (2012) 1:134–46. doi: 10.1016/j.specom.2011.07.007
31. Hou LM, Zhang CH, Yin SK, Yi HL, Meng LL. *A Device of Screening OSAHS Based on Recording Snore Sound.* Chinese Patent CN103735267A. Amsterdam: Elsevier Science (2014).

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 Digital Health

ORIGINAL RESEARCH
published: 17 February 2021
doi: 10.3389/fdgth.2020.609349

A Dataset of Pulmonary Lesions With Multiple-Level Attributes and Fine Contours

Ping Li[1†], Xiangwen Kong[2†], Johann Li[2†], Guangming Zhu[2], Xiaoyuan Lu[1], Peiyi Shen[1], Syed Afaq Ali Shah[3], Mohammed Bennamoun[4] and Tao Hua[5]*

[1] Shanghai BNC, Shanghai, China, [2] Embedded Technology & Vision Processing Research Center, Xidian University, Xi'an, China, [3] College of Science, Health, Engineering and Education, Murdoch University, Perth, WA, Australia, [4] School of Computer Science and Software Engineering, The University of Western Australia, Perth, WA, Australia, [5] Pet Center, Huashan Hospital, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Kezhi Li,
University College London,
United Kingdom

Reviewed by:
Tao Chen,
Virginia Tech, United States
Jian Guo,
RIKEN Center for Computational Science, Japan
Zhimin Liu,
Janssen Pharmaceuticals, Inc., United States

***Correspondence:**
Guangming Zhu
gmzhu@xidian.edu.cn

[†] These authors have contributed equally to this work and share first authorship

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

Received: 23 September 2020
Accepted: 09 December 2020
Published: 17 February 2021

Citation:
Li P, Kong X, Li J, Zhu G, Lu X, Shen P, Shah SAA, Bennamoun M and Hua T (2021) A Dataset of Pulmonary Lesions With Multiple-Level Attributes and Fine Contours. Front. Digit. Health 2:609349. doi: 10.3389/fdgth.2020.609349

Lung cancer is a life-threatening disease and its diagnosis is of great significance. Data scarcity and unavailability of datasets is a major bottleneck in lung cancer research. In this paper, we introduce a dataset of pulmonary lesions for designing the computer-aided diagnosis (CAD) systems. The dataset has fine contour annotations and nine attribute annotations. We define the structure of the dataset in detail, and then discuss the relationship of the attributes and pathology, and the correlation between the nine attributes with the chi-square test. To demonstrate the contribution of our dataset to computer-aided system design, we define four tasks that can be developed using our dataset. Then, we use our dataset to model multi-attribute classification tasks. We discuss the performance in 2D, 2.5D, and 3D input modes of the classification model. To improve performance, we introduce two attention mechanisms and verify the principles of the attention mechanisms through visualization. Experimental results show the relationship between different models and different levels of attributes.

Keywords: deep learning, radiology, pulmonary dataset, classification, attention

1. INTRODUCTION

Lung cancer is caused by tumors which leads to the fastest increase in morbidity and mortality. It has a significant negative impact on the health of subjects. Therefore, the early diagnosis of lung lesions is of great significance for the treatment of lung cancer.

The early form of lung cancer is categorized as pulmonary nodules, which are clinically examined using computed tomography (CT). The characteristics of pulmonary nodules in CT images are diverse, which results in a large workload for radiologists to diagnosis the disease and leads to the subjective assessment of features. Therefore, accurate and quantitative analysis of the appearance characteristics of lung nodules is very essential for doctors to determine whether the nodules will grow into malignant tumors.

In recent years, with the development of deep learning technology (1), lung nodule diagnosis has made unprecedented progress in detection (2–7), segmentation (8–11), classification (2, 6, 12–15), and registration (16, 17) tasks. In order to improve the performance of the model, there is a great need of large datasets and accurate annotation of pulmonary lesions.

There are many publicly available datasets of pulmonary nodules. However, there are some shortcomings in the existing datasets, and the diversity of lesions cannot be balanced in these

datasets. For example, LIDC/IDRI (18) has rich attributes, however, it only marks nodules, and the prediction of other pulmonary diseases cannot be performed.

In this paper, we propose a dataset of lung lesions that could help the development of a pulmonary computer-aided diagnosis system. Our dataset is multi-centered, data-diversified, and informative. The proposed dataset is rich in lesion types and covers most of the signs of lung lesions. The lesions of the dataset are labeled with contours and attribute annotations by experienced radiologists using a professional tool. The attribute annotations are composed of nine attributes that are most useful for pathological assessment. In order to make the selected attributes hierarchical, we have selected multi-level attributes:

- **Low-level attributes**: Margin, spiculation, etc, which can be judged basically by the local features of the lesion;
- **Middle-level attributes**: Pleural indentation, vessel convergence, etc, which need to be judged by the relationship with the surrounding tissue around the lesion or cavity and calcification, which need to be judged by the relationship between local features and global features of the lesion;
- **High-level attributes**: The type and the location of the lesion, which requires to be judged by the abstract features of the entire lesion.

In order to describe the proposed dataset clearly, we first count the characteristics of our dataset, define the data storage format and data annotation rules for our dataset. We then propose the contours annotation format. We also focus on the correlation between the attributes of the lesions. In order to study the relationship between multiple attributes, we calculated the probability of a total of 27 categories of 9 different attributes using the chi-square test and conditional probability, and infer the correlation with the attributes by probability.

In order to illustrate the practical significance of our dataset, we discuss several applications that could be studied using our dataset, and then select the attribute classification for further study. First, we model the attribute classification and then explored the performance of the 2D, 2.5D, and 3D input modes on the accuracy of the model. Through experiments, we demonstrate that there is implicit competition between multiple attributes, we, therefore, use two attention mechanisms to filter different feature activations for different attributes. Our experiments show that the attention mechanisms have different effects on attribute classification.

2. RELATED WORK

In this section, we briefly discuss the existing datasets of lung nodules and the relevant classification methods.

2.1. Lung Nodule Datasets

2.1.1. LUNA16 Dataset

The LUNA16 (4) dataset was designed for the Open Pulmonary Nod Challenge, which screened 888 CT volumes from a large dataset LIDC/IDRI as challenge data. Their slice thickness is within 2.5 mm and the nodule size is greater than 3 mm, which was annotated by more than 3 experimental doctors using tow-phase annotation. The detection annotations of a nodule in LUNA16 use the center coordinates and diameter of the inscribed circle of the nodule. In contrast, we use the gravity center coordinates as the center coordinates of the nodule and the longer geometric moment as the diameter to generate the world coordinates. For small round nodules, the two datasets are not much different, but the need is to detect large lesions with irregular shapes and our proposed approach achieves better results for large lesion detection.

2.1.2. LIDC/IDRI Dataset

The LIDC/IDRI (18) dataset labels each nodule with a contour and nine attributes. Besides the benign and malignant nodules, the other eight attributes are all the appearance attributes of the nodules. In contrast, in our dataset, two of the attributes are the basic attributes of the lesion, five are appearance attributes, and two have relationships with the tissue surrounding the lesion in context. These attributes are richer and can better represent a lesion.

2.1.3. LISS Database

The LISS (19) database has 271 CT volumes, including 677 abnormal regions. These abnormal regions are divided into nine categories, which are called common CT imaging signs of lung disease (CISLs). In other words, there is only one CISLs label for each abnormal region. Although it can better help medical scholars learn a certain type of disease (12), it is not very good for CAD system development, because it cannot capture the relationship between disease signs.

2.1.4. ILD Database

The ILD (20) database has 108 image series with more than 1946 ROIs. This dataset is a multimedia collection of cases of interstitial lung disease (ILDs). These ROIs are divided into 13 categories, which are lung tissue patterns from histological diagnoses of ILDs. The lesions in the ILD dataset are large, and the annotations are all high-level attributes. The dataset does not focus on a certain nodule, but on the pathology presented by a piece of tissue.

2.2. Lung Nodule Classification

The classification of lung nodules based on deep learning can be divided into two types of methods: one is to judge the benign and malignant lung nodules. Some methods directly predict the benign and malignant nodules by CT images, and other methods use different attributes of the nodules as the auxiliary basis to judge the benign and malignant nodules, such as (21–23). The other type of method has classified the disease, such as DeepLung (2) or LISCs classification (12). Dey et al. (21) have built a network that produces multiple outputs from multi-scale features to judge the benign and malignant nodules. Nibali et al. (22) has made a three-column configuration to fuse the features generated from three axes. Song et al. (14, 23) proposed methods that split the whole image into patches and predict the lesions. In contrast, Gao et al. (13) have used the whole image for classification. With the development of computationally efficient computers, the 3D models such as (24) has achieved an impressive performance in nodule classification. He (12) proposed a method to generate images for data augmentation,

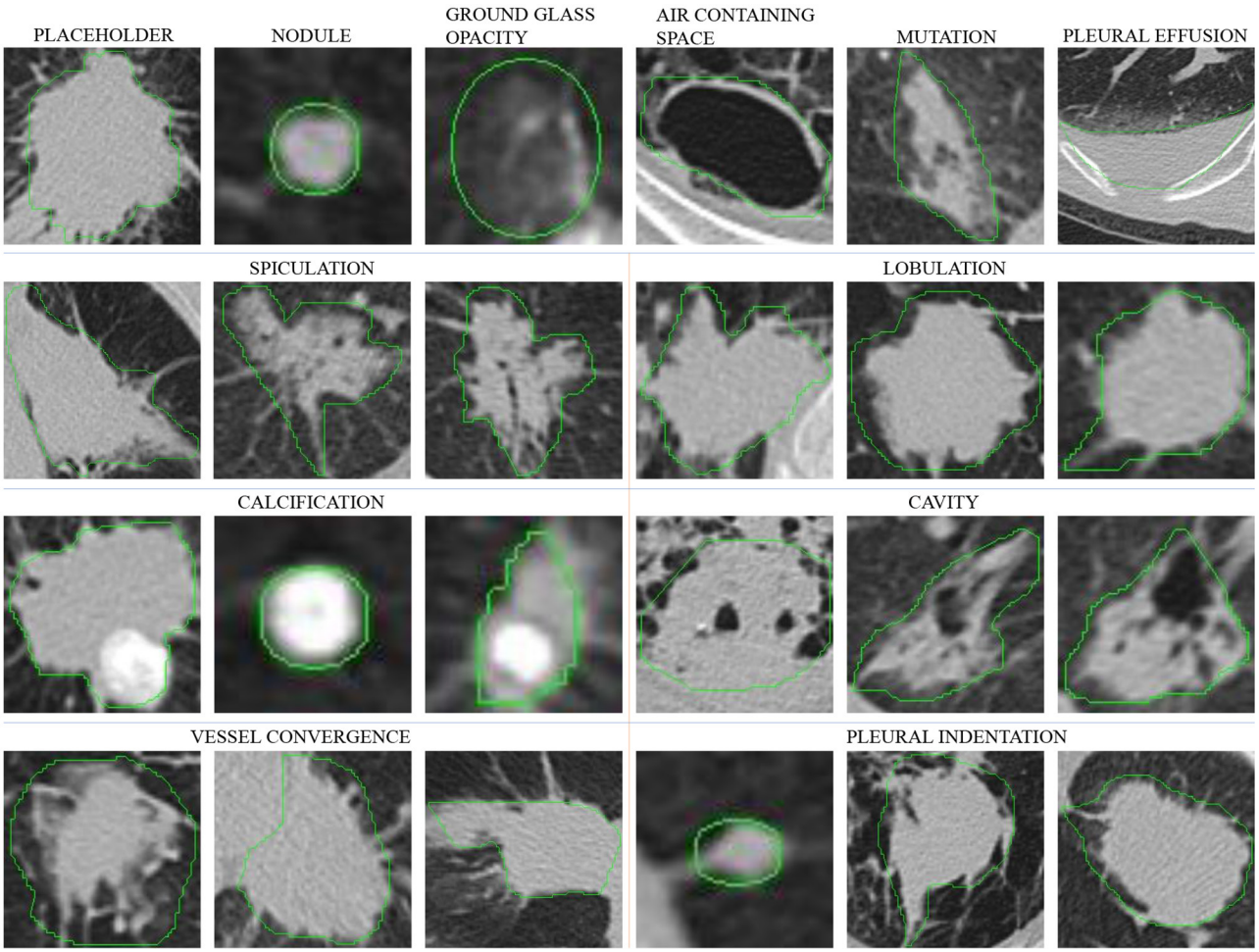

FIGURE 1 | Lesions in our dataset. Except for some small nodules, which are marked with a circle, such as the second image in the first row, other lesions are marked by a very close contour. The six images in the first row are different types of lesions, and in the second to fourth rows, each set of three images are spiculation, lobulation, calcification, cavity, vessel convergence, and pleural indentation.

which achieved a good improvement in performance. Zhu et al. (2) detected the position of the nodules first, then cropped the sent the nodules before feeding it into a classification model to predict one of nine attributes.

Multi-attribute classification is a problem to classify multiple targets using one model. There are currently two approaches to solve this problem. The first is to regard it as a classification task with a fixed number of categories, and solve attribute correlation in one model by using multiple branches to decompose the relationship between multiple targets onto each branch. The second is to treat it as a multi-label classification task, with the positive attribute as the label of the lesion, then each lesion has a floating number of labels, and the labels are decoupled using different methods. In this paper, we use the first method to classify different attributes in a model using a fixed number of branches, and use two attention mechanisms to help decouple the correlation among the attributes.

3. LUNG LESION DATASET

In this section, we provide a description of our dataset. CT data were collected from four hospitals. The body parts examined are mainly the chest and abdomen. Among them, the chest CT was mostly thin (less than 3 mm), and the abdomen CT was mostly thick (greater than or equal to 5 mm). **Figure 1** shows examples of lesions in our dataset. As shown in **Figure 1**, except for some small nodules, which are marked with circles, such as the second image in the first row, other lesions are marked by a very close contour.

Table 1 shows the parameter comparison of our dataset with several other public datasets. Same with LUNA16, our dataset annotates lesion with contour, which is shown in **Figure 1**. Compared with box and polygon, contour annotation has more generalization ability to different tasks, such as location, detection, and segmentation. At the same time, though the

TABLE 1 | The statistical result of comparing our dataset parameters with other datasets.

Dataset	Annotation	Lesion attributes	Multiple categories	Scans	Lesion amount	Lesion size (mm)	Slice thickness (mm)	Pixel spacing (mm)
LUNA16	contour	9	✓	888	1,186	3.25–32.27	0.45–2.50	0.461–0.976
LISS (2D)	Box	9	×	252	511	–	5.0	0.42–1.00
LISS (2D)	Box	9	×	19	166	–	1, 1.25	0.60–0.87
ILD	Polygon	13	×	108	1,946	–	1.00–2.00	0.40–1.00
Ours	Contour	9	✓	694	5,113	0.83–191.32	1.00–2.00	0.176–0.977

number of scans in our dataset is not the largest, the number of lesion annotations and the range of lesion size in our dataset are. These annotations support more robust models. Moreover, the thickness of the slices of our datasets is relatively uniform, especially compared to LUNA 16. It reduces unnecessary processing of the data and makes it easier to use.

3.1. File Storage and Annotation Format

The raw data obtained from the hospital contains some sensitive information of subjects, and the data collected from different hospitals are stored in different ways, making the data difficult to use directly for analysis. Therefore, we first desensitize the data by removing subjects' sensitive information and retain only the necessary information, such as weight. Then, we store the CT volumes and annotation files as described below.

We define the directory structure to store files as follows:

`ct_type/hospital/year/month/day/subject_id/series_id.`

The directory with series_id SE01 stores the CT data with DICOM format, and the directory with series_id SE01_01_0*n* stores the contour annotation file aid_loc.anno, where *n* is the identification number of the doctor who annotated the scans; aid is the number of the annotation in the CT for correspondence with the attribute information; loc is the slice number in the CT volume, and the description in the DICOM file is SliceLocation (0020, 1041). An anno file represents an annotation. Each anno file has a different aid, but two anno files can have the same loc, indicating that the two annotations are in the same slice. It uses a dictionary to store the annotation information we need to use in the CAD tasks. The keywords of the anno format are SeriesID, NoduleSerialNumber, InstanceNumber, Origin, Dimension, Spacing, Coords, XMin, XMax, YMin, YMax. Among them, SeriesID is a unique number of a DICOM volume which described as SeriesInstanceUID (0020, 000E), NoduleSerialNumber and InstanceNumber are aid and loc, respectively as mentioned above, Origin, Dimension, Spacing are the information from DICOM volume, Coords is the contour coordinate of this annotation, and its value is relative to the size of this slice. (XMin, YMin), (XMax, YMax) are the coordinates of the lower left and upper right corners of the bounding box of this annotation.

The CT volumes in our dataset contain lesions, while those without lesions have been removed by manually screening of RIS reports. For repeated subject numbers, such as two volumes of one subject, we map one of them to a new subject number and retain the correspondence to restore the original number.

3.2. Two-Phase Annotation Process

We use a two-phase annotation process to label the lesions. We label the contours of the lesions in the first phase, then label the attributes of the lesions in the second phase.

3.2.1. Contour Annotation Criterion

The contours are marked by experienced radiologists. In order to save the doctor's time and to increase the density of the lesion, we first manually screen the RIS report, retain the CT volume with the lesion in the description, and remove the volume without the lesion from the dataset. In order to standardize the process of marking the lesions, we have prescribed a rule for marking lesions with the doctor as follows:

- Mark all visible lesions;
- If the lesion is too small to draw the contour, circle the lesion with a circle tool;
- If the lesion is larger than one slice, mark the lesion every three consecutive slices;
- Draw a contour as close as possible to the edge of a lesion.

After the marking process, we perform a secondary screening to remove the annotations which are too discontinuous to be processed as contours. Then, we convert the annotations into anno format and mark lesion numbers. In this way, the contour annotations and the attribute annotations correspond with respective file names.

3.2.2. Attribute Annotation Criterion

After discussed with the doctor, we selected nine attributes that are commonly used in clinical diagnosis as attribute annotations for the dataset. A detailed description of these attributes will be provided in section 5.2. Each lesion is independently labeled by a doctor, and we record the doctor's number for each lesion that can be used to identify the doctor if an error is discovered in the annotation.

In order to simplify the labeling of attributes, we implement an attribute labeling tool to collect and manage labels. We associate the slice of the contour with the lesion number so that it is convenient to label the attributes with the corresponding slice. When the attribute information is marked, the corresponding subject number and label number are recorded to correspond to the contour number. It should be noted that the contour annotation and the attribute annotation are not one-to-one

matched. Some problematic contour annotations are filtered out in the previous step, and no attribute annotation is performed. Finally, we only select lesions with both contour and attribute annotations into the dataset. The number of attributes is reported in **Table 1**. As can be noted, the categories of some attributes are very unbalanced. This brings great challenges to the performance of our attribute classification algorithm.

4. ATTRIBUTES AND PATHOLOGY

We initially selected 15 attributes that are commonly used in clinical diagnosis, and then selected 9 attributes for our dataset based on their importance. The number of categories of these attributes is not balanced and the distributions are not independent. Here we briefly describe the importance of these attributes in clinical diagnosis and then discuss the correlation between attributes from the statistics point of view.

4.1. Attributes Description

Among the 9 attributes we selected, besides the basic attribute, lesion type, and lesion location, there are vessel convergence and pleural indentation which represent the relationship between the lesion and the surrounding tissue. On the other hand, margin, calcification, lobulation, spiculation, cavity represent the apparent features of the lesion. The description of the significance of these nine attributes is as follows.

4.1.1. Lesion Type

The first row of **Figure 1** shows six different lesion types. For the lesion type, we choose *placeholder*, *nodule*, *ground glass opacity*, *air containing space*, *mutation*, and *pleural effusion*. The difference between placeholders and nodules is that the lesions with a diameter of less than 30 mm are nodules, and those larger than 30 mm are placeholders. Except for the difference in size, the other attributes of the two lesion types are roughly similar. The air containing space is different from the cavity in pathology. The air containing space (**Figure 1**, the fifth image in the first row) is a pathological enlargement of the physiological cavity in the lung, while the cavities (**Figure 1**, the last three images in the third row) often appears in nodules or placeholders. In the air containing space lesions, the wall of the lesion is thinner and more uniform, mostly occurring in the subpleural area, and the size varies greatly. This means that the location of the air containing space is fixed and there are no apparent attributes such as spiculation and lobulation.

4.1.2. Lesion Location

The location of the nodule is represented by five categories of lobes, including the *right upper* lobe, the *right middle* lobe, the *right lower* lobe, the *left upper* lobe, and the *left lower* lobe. Statistics show that the occurrence of lesions has little relationship with the location. The lesion location is only a basic attribute of the lesion, and it cannot be used as a basis for judging its pathological nature. Some lesions are large and span multiple lung lobes, so we mark them as 0, and do not include it in the five categories above.

4.1.3. Margin

The margin attribute describes whether the outer boundary of a nodule is clear. We defined two main categories for this attribute: *clear* and *unclear* margin. Though the margin of a benign mass is often smooth, while that of a malignant mass is often unclear, inflammation may also cause an unclear margin of placeholder. Therefore, it cannot be used as the sole basis for judging benign and malignant lesion, and needs to be judged in combination with other attributes.

4.1.4. Calcification

The calcification attribute describes lesions whose density is significantly higher than other soft tissues in the mediastinal window, usually with CT values above 100 Hu. The first three images in the third row of **Figure 1** show lesions of calcification. The white region in the images represents calcification. Calcification is a pathologically metamorphic lesion, which is more common in the healing stage of ductal tuberculosis lesions in the lung tissue or lymph nodes; calcification can also occur in tumor tissues or cyst walls. Usually, the greater the proportion of calcification in the lesion, the greater the likelihood of its being benign. Based on this, we classify the calcification attributes into three categories: *no*, *partial*, and *total* calcification.

4.1.5. Lobulation

The lobulation attribute indicates that the nodule or mass grows at different speeds in various directions or is blocked by the surrounding structure. The contours may have a plurality of arcuate protrusions, and the curved phases are concave cuts to form a lobulated shape. The last three images in the second row of **Figure 1** show the lesions of lobulation. We can clearly see the convex part of the masses. We simply define two categories for this attribute: *with* and *without* lobulation.

4.1.6. Spiculation

The spiculation attribute is characterized by a radial, unbranched, straight, and strong thin line shadow extending from the edge of the nodule to the periphery, and the proximal end of the shadow is slightly thicker. The first three images in the second row of **Figure 1** show lesions of spiculation. As shown in **Figure 1**, the burrs of the lesion are often not circled in the scope of annotation. The spiculation is not connected to the pleura, and distinct from the pleural depression. We classify the spiculation attributes into *no*, *short* and *long* spiculation; 5 mm burrs are called short spiculation, and larger than 5 mm burrs are called long spiculation. The pathological basis of the burr is the fiber band in which the tumor cells infiltrate into the adjacent bronchial sheath and local lymphatic vessels, or the tumor promotes connective tissue formation. Benign nodular inflammatory pseudotumor, tuberculoma can also be seen burrs, but longer, softer, more often formed by hyperplastic fibrous connective tissue. The possibility of lung cancer should be considered when there is a burr in solitary lung nodules.

4.1.7. Cavity

The cancerous cavities are mostly located in the anterior segment of the upper lobe and the basal segment of the lower lobe. Most

of the cavities larger than 3 cm in diameter are tumors. Most cancerous cavities present an irregular or lobulated outer edge and irregular inner edge. Those with a wall thinner than 4 mm are mostly benign lesions, and those thicker than 15 mm are mostly malignant lesions. The last three images in the third row of **Figure 1** show the lesions of a cavity. We simply defined two categories for this attribute: *with* and *without* cavity.

4.1.8. Vessel Convergence

The vessel convergence attribute appears on the slices as one or more vessels around the pulmonary nodule that touch with, cut or pass through the placeholder at its edge. The appearance of vessel convergence is related to the size of the placeholder or nodule. The lesions less than 1 cm in diameter have fewer vessel convergence signs. The first three images in the last row of **Figure 1** shows the lesions of vessel convergence. Images of the cavities and vessel convergence are similar, because the blood vessels look like cavities when they are transacted. A multi-vessel-directed lesion presents vessel convergence, which leads to a higher chance of malignancy. In particular, the phenomenon that one blood vessel leads to a nodule or tumor is not only seen in malignant nodules, but also in benign lesions such as tuberculosis, inflammatory pseudotumor, or hamartoma. We simply defined two categories for this attribute: *with* and *without* vessel convergence.

4.1.9. Pleural Indentation

The typical pleural indentation shows a small triangular shadow or a small trumpet shadow on the visceral surface of the visceral pleura. The bottom of the triangle is on the inside of the chest wall, the tip points on the nodule, and the nodule and the triangle shadow can be connected by a linear shadow. The last three images in the last row of **Figure 1** shows the lesions of pleural indentation. Peripheral lesions of the pleural indentation are often accompanied by other imaging signs. The pathological basis and imaging manifestations of pleural indentation in benign and malignant lesions are different. We simply define two categories for this attribute: *with* and *without* pleural indentation.

4.2. Correlation Between Attributes

In order to evaluate the correlation between attributes, we used the chi-square test. We assume that if the two attributes are independent of each other, their data distribution should not affect each other, which means that the proportional relationship between the categories of one attribute is the same under each category of the other attribute. If the chi-square test value calculated by the two attributes is greater than the statistical significance, there is a correlation between the two attributes. The approximate calculation equation for the chi-square test statistic is as follows:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{1}$$

where f_0 is the actual number of observations and f_e is the expected number of times. The larger the value of f_e, the Equation (1) approximately obeys the chi-square distribution. To simplify the calculation of the chi-square test, we used a variant of Equation (1):

$$\chi^2 = \sum \frac{\left(f_{xy} - \frac{f_x f_y}{N}\right)^2}{\frac{f_x f_y}{N}} = N\left(\sum_{x=1}^{R}\sum_{y=1}^{C}\frac{f_{xy}^2}{f_x f_y} - 1\right) \tag{2}$$

where f_x and f_y represent the number of samples of the categories of two different attributes x and y, respectively, R and C are the number of categories of f_x and f_y, and the total number of attributes is N. The degree of freedom df of the independence test is calculated as follows:

$$df = R \times C - R - C - 1 = (R-1)(C-1) \tag{3}$$

We use the data shown in **Table 2** and select a significance level of 0.05 for calculation. **Figure 2A** shows the result of the chi-square test. As the results show, there is a strong correlation between the three attributes of margin, speculation, and lobulation. Meanwhile, there is a strong correlation between vessel convergence and spiculation, margin, lobulation and lesion type, pleural indentation, and margin.

To further explore the specific relationship between the various categories of attributes, we calculated the conditional probability between a total of 27 categories for all attributes. The equation for calculating the conditional probability is as follows:

$$P(X|Y) = \frac{P(XY)}{P(Y)} \tag{4}$$

where $P(X)$ and $P(Y)$ represent the probabilities of two categories X and Y, $P(X|Y)$ represents the probability of X to occur when Y is present, and $P(XY)$ represents the probability of co-occurrence for X and Y. The value of $P(X|X)$ is 1, which is represented by white color in **Figure 2B**. We calculated the conditional probability between each of the two categories. As shown in **Figure 2B**, the white color represents a probability of 1 and the black color represents a probability of 0, while the lighter gray color represents higher conditional probability values.

According to the statistical results, there is a strong correlation between different lesion types and other attributes. For the placeholder, their margins are almost unclear, the degree of lobulation is more obvious, the degree of spiculation and the degree of pleural indentation are the highest among other lesion types. The nodules, ground glass, and mutation categories have a small number of spiculation and lobulation, and more features of vessel convergence and pleural indentation. For cavity and pleural effusion, they almost have no other attributes and their margins are all clear.

The margin attribute is highly correlated with lobulation, vessel convergence, and pleural indentation. When vessel convergence and pleural indentation are present, they are often accompanied by lobulation, and the margin is not very clear. The calcification attribute is concentrated in the nodules, and the cavity is also related to the margin and lobulation.

TABLE 2 | The distribution of each attribute category used for experiments.

Attribute	Categories	Lesions	Attribute	Categories	Lesions	Attribute	Categories	Lesions
Lesion type	Placeholder	675	Lesion location	Right upper	496	Calcification	None	1,902
	Nodule	728		Right middle	151		Partial	62
	Ground glass opacity	220		Right lower	286		Total	50
	Air containing space	153		Left upper	374	Cavity	Without	1,924
	Mutation	208		Left lower	271		With	90
	Pleural effusion	30	Margin	Clear	887	Vessel Convergence	Without	1,461
Spiculation	None	1,198		Unclear	1,127		With	553
	Long	307	Lobulation	Without	1,015	Pleural Indentation	Without	1,222
	Short	509		With	999		With	792

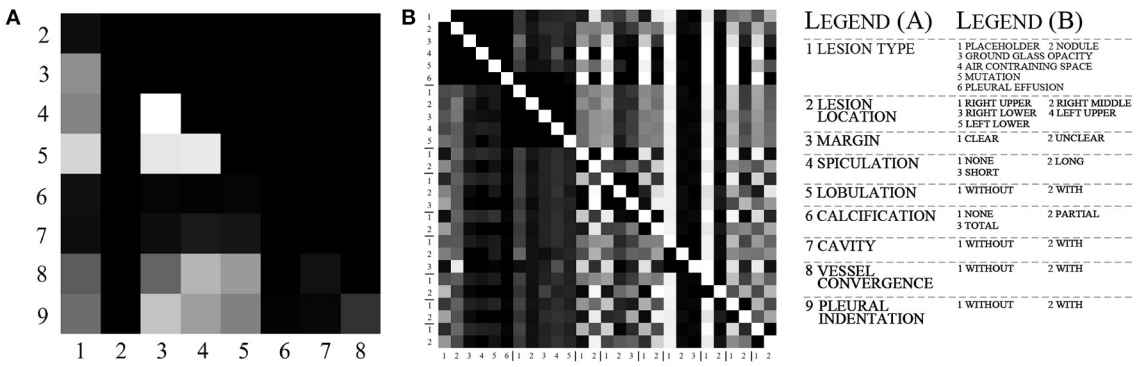

FIGURE 2 | The visualization of chi-test and conditional probabilities. **(A)** is the visualization of the chi-test result. **(B)** is the visualization of conditional probabilities. The brighter grid means that the attributes indicated by its row and column numbers are more relevant. For **(A)**, the meaning of labels 1–9 is listed in legend, and for **(B)**, the legend lists the meaning of label in each group.

5. TASKS OF DATASET

Our dataset is rich in data and diverse in annotations, which means that our dataset can be used for several tasks and aid in the development of CAD systems. We recommend using our dataset for the following tasks:

(1) **Detection**: Some of the lesions in our dataset are smaller than 30 mm, which are nearly circular and suitable for lung nodules detection. This can be helpful for the initial diagnosis of lung cancer.
(2) **Segmentation**: The lesions larger than 30 mm are all marked with precise contours. These lesions are more complex in shape and are suitable for the lung lesion segmentation task. This can be helpful for volume measurement and further treatment.
(3) **Classification**: Multiple attributes of the lesion are suitable for multi-task lung disease prediction. This can be helpful to judge benign and malignant tumors.
(4) **Reconstruction**: At present, medical datasets are small, and their size is not enough for deep learning. Our dataset has various types of data, and we can use real data to train generative adversarial networks to generate synthetic data.

In this paper, we focus on exploring the correlation between attributes. We, therefore, perform multi-attribute classification and report our experimental results in section 6.

5.1. 2D, 2.5D, 3D Modes for Classification

In order to study the importance of the input mode for the model, we use different data dimensions for the same data and the model for classification experiments.

We use three input modes including 2D, 2.5D, and 3D. Assuming that the size of a CT volume is $H \times W \times C$, which corresponds to the three axes of X-Y-Z, the diameter of a lesion is d, the three input modes are expressed as follows:

5.1.1. 2D Mode

The lesion is cut out from the grayscale slice in which it is located with a length d of side, and fed to a 2D network for prediction. The input size is $d \times d \times 1$. The 2D input mode can retain the lesion at the spatial structure in the X-Y direction, but the context information in the Z direction cannot be captured.

5.1.2. 2.5D Mode

The grayscale image of the lesion and the five images above and below are cut out by the bounding box, and fed to the 2D network for prediction. The number of input channels is 5, and the input size is $d \times d \times 5$. Compared to the 2D input mode, the 2.5D

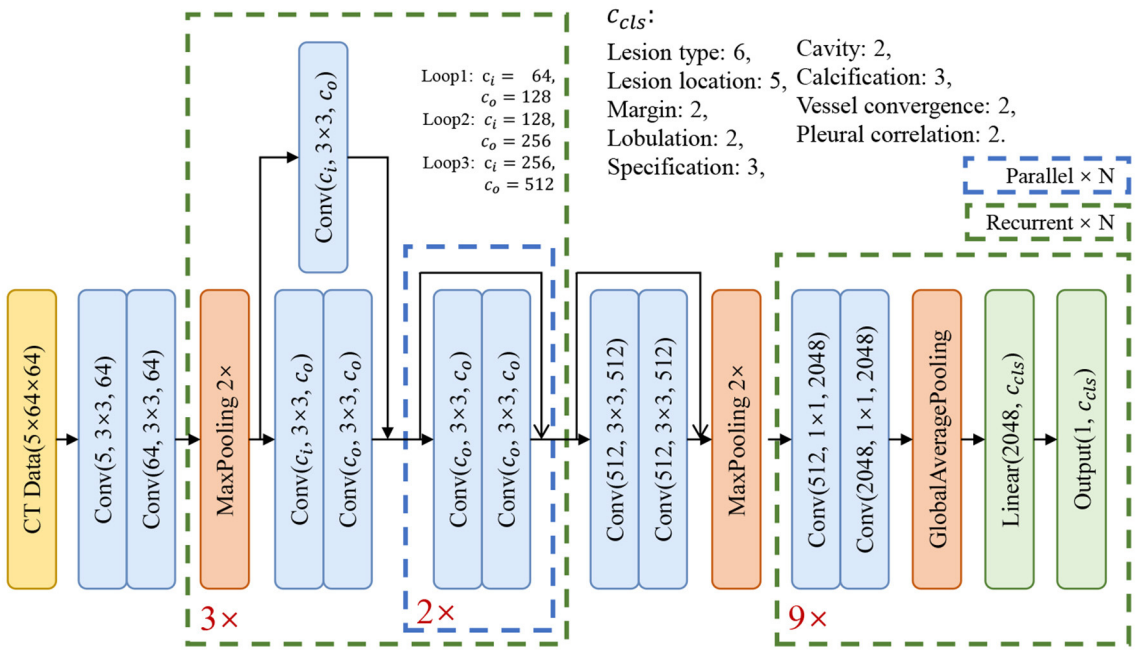

FIGURE 3 | The structure of our basic classification model.

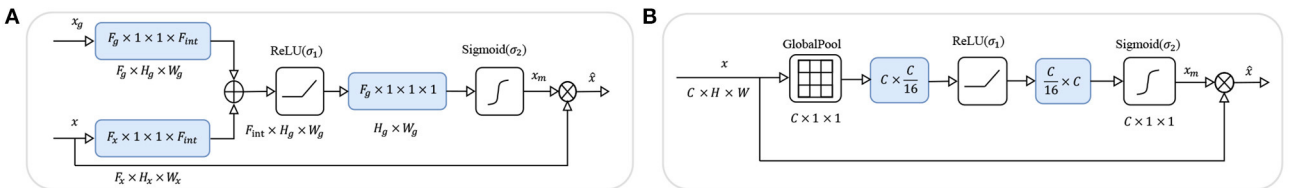

FIGURE 4 | The structures of the two attention modules. In the figures above, blue boxes represent the convolutional layers, ⊕ represents the element-wise sum, and ⊗ represents the spatial-wise reweight in **(A)** and channel-wise reweight in **(B)**.

input mode is supplemented by a fixed number of slices in the Z direction.

5.1.3. 3D Mode

In the X-Y-Z direction where the lesion is located, the bounding box ($d \times d \times d$) is cropped and fed to the 3D network for prediction. 3D network can capture the correlation on the Z-axis of the whole lesion by convolution. Compared with 2D, the information of 2.5D is more detailed, but the amount of 3D network parameters is more than that of 2D network, which can cause the deep learning model to overfit as the size of training data is small.

The architecture of our basic model is shown in **Figure 3**. In order to extract the relationship of nine attributes, we use a ResNet-based network (25) to extract the characteristics of the nodule and then use nine classification branches to predict nine attributes independently. We will explain the details and the results in section 6.1.

5.2. Two Attention Mechanisms

Through the experiments, we found that there is an implicit competition between multiple attributes during training. In the training phase, when the loss value is stable, the accuracy of some attributes increases while the accuracy of other attributes decreases. To solve this problem, we add an attention module in front of each attribute classifier to focus the activation on the features which are useful for classification. In this way, different input features for attributes are extracted, which could mitigate the conflict between attributes. Inspired by (26–28), we employed soft-attention and self-attention, commonly used mechanisms that compute a weight matrix used to filter noise and to focus on important features. These two attention mechanisms are described below in our model, and **Figure 4** shows the structure of the two attention modules.

5.2.1. Soft-Attention Module

As shown in **Figure 4A**, we add a soft-attention module (26) before feeding the features into the classifier to filter out

shallower features with deeper features. While preserving the spatial structure, the attention module extracts a mask from the features to suppress noise which is not related to the attribute to improve accuracy.

Assuming that feature map $x \in \mathbb{R}^{N\times C_x\times H\times W}$ from the basic model is the input feature for the attention model, and feature map $x_g \in \mathbb{R}^{N\times C_g\times H\times W}$ is from a deeper layer as the gate, we firstly use 1×1 convolutional layer to get the same number of channels C_g for both the features, then sum the features x and x_g together and add a non-linear transform ReLU which can be formulated as $\sigma_1(x) = max(0, x)$. So far, the feature x is mixed with richer semantic information x_g, and we use a 1×1 convolutional layer to fuse the channel information and retain the spatial information, and get a mask x_m with a value of $[0, 1]$ through the sigmoid function which can be formulated as $\sigma_2(x) = \left(1 + e^{-x}\right)^{-1}$. Finally, we use the mask x_m to spatial-wise reweight the feature map x and get the output feature $\hat{x}$. After filtering by the soft-attention module, the features $\hat{x}$ are re-weighted by high-dimensional semantic information in the spatial dimension, which is more conducive to multi-attribute classification.

5.2.2. Self-Attention Module

As shown in **Figure 4B**, we add a self-attention module (27, 28) before the features and fed to the classifier to squeeze the spatial structure of a feature map into one vector with spatial information. Then, we gather and filter the information to enhance the activation related to that attribute, and add the information to the original feature map to enhance the feature.

TABLE 3 | Performance of the basic model on the 3D, 2.5D, and 2D modes.

Attributes	Categories	Accuracy			Sensitivity			Specificity		
		3D	2.5D	2D	3D	2.5D	2D	3D	2.5D	2D
Lesion type	Placeholder	0.8636	0.8182	**0.8864**	0.7451	**0.8780**	0.7959	0.9500	0.9344	**0.9561**
	Nodule	0.7460	0.6984	**0.7619**	0.8545	0.8462	**0.9057**	0.8621	0.8288	**0.8636**
	Ground glass opacity	**0.8800**	0.8000	0.8000	**0.8462**	0.8333	0.8000	**0.9793**	0.9640	0.9638
	Air containing space	0.9375	**1.0000**	0.9091	**0.9375**	0.4400	0.7143	0.9935	**1.0000**	0.9933
	Mutation	0.8235	**0.9286**	0.8571	0.8235	**0.8667**	0.8571	0.9805	**0.9932**	0.9866
	Pleural effusion	1.0000	1.0000	1.0000	**1.0000**	**1.0000**	0.7500	1.0000	1.0000	1.0000
Margin	Clear	0.8437	0.8523	**0.8636**	**0.8617**	0.8427	0.8261	0.8052	0.8243	**0.8310**
	Unclear	**0.8267**	0.8133	0.7867	0.8052	0.8243	**0.8310**	**0.8617**	0.8427	0.8261
Spiculation	None	**0.8672**	0.8083	0.8083	0.9407	**0.9604**	**0.9604**	**0.6792**	0.6290	0.6290
	Long	0.3333	0.6000	**0.7333**	**0.4167**	0.2571	0.2558	0.9371	0.9531	**0.9667**
	Short	**0.7143**	0.2857	0.4286	0.4878	0.2963	**0.6316**	**0.9385**	0.8529	0.8889
Lobulation	Without	0.8972	0.8990	**0.9091**	0.9143	0.9271	**0.9375**	0.8333	0.8507	**0.8657**
	With	0.8594	0.8906	**0.9062**	0.8333	0.8507	**0.8657**	0.9143	0.9271	**0.9375**
Calcification	None	**0.9937**	0.8684	0.7763	0.9464	**0.9565**	0.9516	**0.6667**	0.2000	0.1282
	Partial	0.0000	**0.2500**	**0.2500**	0.0000	**0.1176**	0.0741	0.9529	**0.9589**	0.9559
	Total	0.6667	**1.0000**	**1.0000**	**1.0000**	0.3750	0.2500	0.9941	**1.0000**	**1.0000**
Cavity	Without	**0.9819**	0.9557	0.9367	0.9702	0.9742	**0.9801**	0.0000	0.1250	**0.1667**
	With	0.0000	0.2000	**0.4000**	0.0000	0.1250	**0.1667**	0.9702	0.9742	**0.9801**
Vessel convergence	Without	**0.9618**	0.8618	0.8780	0.8936	0.8548	**0.9076**	**0.8333**	0.5641	0.6591
	With	0.6250	0.5500	**0.7250**	**0.8333**	0.5641	0.6591	0.8936	0.8548	**0.9076**
Pleural indentation	Without	**0.8500**	0.8125	0.7946	0.8430	0.8922	**0.9082**	0.6400	**0.6557**	0.6462
	With	0.6275	0.7843	**0.8235**	0.6400	**0.6557**	0.6462	0.8430	0.8922	**0.9082**
Lesion location	Right upper	**1.0000**	0.7083	0.7083	1.0000	1.0000	1.0000	**1.0000**	0.8793	0.8793
	Right middle	**0.8571**	0.7500	0.7500	**1.0000**	0.3000	0.3000	0.9934	**0.9929**	**0.9929**
	Right lower	**1.0000**	0.8750	0.8333	**0.9630**	0.6562	0.7143	**1.0000**	0.9746	0.9672
	Left upper	**1.0000**	0.7143	0.8571	0.9118	**0.9524**	0.9231	**1.0000**	0.9380	0.9677
	Left lower	0.9348	**0.9783**	0.9565	**1.0000**	0.8491	0.8462	0.9739	**0.9897**	0.9796
	Average	0.7513	0.7511	**0.7816**	**0.7671**	0.7006	0.7184	**0.8305**	0.7995	0.8116

Bold value means better performance, compared between 2D, 2.5D, 3D.

Assuming that feature map $x \in \mathbb{R}^{N\times C_x\times H\times W}$ is generated from the basic model as the input feature of the attention model, we use a channel squeeze and spatial excitation branch to transform x to extract the spatial information and reweight the origin x with the transform of itself. We use a global pool which can squeeze x to a vector $z \in \mathbb{R}^{N\times C_x\times 1\times 1}$. Then use two fully connected layers to transform the vector z to $\hat{z} = W_1\left(\sigma_1(W_2\cdot z)\right)$ with $W_1 \in \mathbb{R}^{C\times C/16}$ and $W_2 \in \mathbb{R}^{C/16\times C}$ and the activation σ_1. We also use the non-linear function σ_2 to transform the values to $[0, 1]$ to get the channel mask x_m. Finally, we use the x_m to channel-wise reweight the feature map x and get the output feature $\hat{x}$. After filtering by the self-attention module, the features $\hat{x}$ are re-weighted by the information after squeeze and excitation in the spatial dimension, which is more conducive to multi-attribute classification.

6. EXPERIMENTAL RESULTS

In this section, we first verify that the proposed model can learn the correlation between attributes, and then empirically select the best input mode, and verify the attention mechanism on this input mode.

We used part of the data with a thickness of 1.0–2.0 mm in our experiments, which has 355 CT volumes and 2014 lesions labeled with 9 attributes in our dataset. The dataset has been split into 8:2 as the training set and validation set, with 1,847 lesions in the training set and 163 lesions in the validation set. During training, we randomly select 30% of the data for data augmentation i.e., random flip and rotation. As **Table 2** shows, the number of categories in the dataset is unbalanced, which could affect the convergence of the model. We use weighted cross

TABLE 4 | Performance of the basic, soft-attention, and self-attention models on the 2D mode.

Attributes	Categories	Accuracy			Sensitivity			Specificity		
		Basic model	Soft-att	Self-att	Basic model	Soft-att	Self-att	Basic model	Soft-att	Self-att
attention Lesion type	Placeholder	**0.8864**	0.8182	0.8636	**0.7959**	0.7347	0.7755	**0.9561**	0.9298	0.9474
	Nodule	**0.7619**	0.6825	0.7302	0.9057	0.8776	**0.9787**	**0.8636**	0.8246	0.8534
	Ground glass opacity	0.8000	**0.9200**	0.8800	**0.8000**	0.7419	0.7857	0.9638	**0.9848**	0.9778
	Air containing space	0.9091	**1.0000**	0.9091	0.7143	**0.7857**	0.7143	0.9933	**1.0000**	0.9933
	Mutation	0.8571	0.9286	**1.0000**	0.8571	**1.0000**	0.7368	0.9866	0.9933	**1.0000**
	Pleural effusion	**1.0000**	**1.0000**	0.8333	0.7500	**0.8571**	0.8333	**1.0000**	**1.0000**	0.9936
Margin	Clear	**0.8636**	0.8523	**0.8636**	0.8261	**0.8621**	0.8352	0.8310	0.8289	**0.8333**
	Unclear	0.7867	**0.8400**	0.8000	0.8310	0.8289	**0.8333**	0.8261	**0.8621**	0.8352
Spiculation	None	**0.8083**	0.7750	0.7917	0.9604	**0.9894**	0.9500	**0.6290**	0.6087	0.6032
	Long	**0.7333**	0.5333	0.4667	0.2558	**0.3478**	0.2500	**0.9667**	0.9500	0.9407
	Short	0.4286	**0.7500**	0.5357	**0.6316**	0.4565	0.4286	0.8889	**0.9402**	0.8984
Lobulation	Without	0.9091	0.9091	**0.9192**	0.9375	0.9474	**0.9479**	0.8657	0.8676	**0.8806**
	With	0.9062	**0.9219**	**0.9219**	0.8657	0.8676	**0.8806**	0.9375	0.9474	**0.9479**
Calcification	None	0.7763	0.8158	**0.8224**	0.9516	**0.9538**	0.9398	0.1282	**0.1515**	0.1000
	Partial	**0.2500**	**0.2500**	0.0000	0.0741	**0.1111**	0.0000	0.9559	**0.9586**	0.9437
	Total	1.0000	1.0000	1.0000	0.2500	0.2000	**0.3333**	1.0000	1.0000	1.0000
Cavity	Without	0.9367	0.8734	**0.9557**	0.9801	**0.9857**	0.9805	0.1667	0.1304	**0.2222**
	With	0.4000	**0.6000**	0.4000	0.1667	0.1304	**0.2222**	0.9801	**0.9857**	0.9805
Vessel convergence	Without	0.8780	0.7967	**0.8211**	0.9076	**0.9515**	0.9182	**0.6591**	0.5833	0.5849
	With	0.7250	**0.8750**	0.7750	**0.6591**	0.5833	0.5849	0.9076	**0.9515**	0.9182
Pleural indentation	Without	**0.7946**	0.7857	0.7589	0.9082	0.9167	**0.9551**	**0.6462**	0.6418	0.6351
	With	0.8235	0.8431	**0.9216**	**0.6462**	0.6418	0.6351	0.9082	0.9167	**0.9551**
lesion Location	Right upper	0.7083	0.7083	0.7083	**1.0000**	**1.0000**	0.9714	**0.8793**	**0.8793**	0.8783
	Right middle	0.7500	0.7500	0.7500	0.3000	**0.3333**	0.3000	0.9929	0.9929	0.9929
	Right lower	0.8333	0.8333	**0.8750**	0.7143	0.6897	**0.7241**	0.9672	0.9669	**0.9752**
	Left upper	0.8571	0.8214	**0.9286**	0.9231	0.9200	**0.9630**	0.9677	0.9600	**0.9837**
	Left lower	0.9565	0.9565	0.9565	0.8462	0.8302	**0.8980**	0.9796	0.9794	**0.9802**
	Average	0.7816	**0.8032**	0.7763	**0.7184**	0.7183	0.7155	0.8116	0.8117	**0.8128**

Bold value means better performance, compared between different models.

entropy loss to reduce the impact of data imbalance during the training phase.

In the experiments, each model has four blocks. The first one is a convolutional block and the other three are residual blocks. At the end of the model, there are nine classifier blocks for the classification of nine attributes, respectively. We use the reweighted logistic loss to balance the numbers of categories. During the training phase, we set the learning rate to 0.01 with warm restart (29) and use SGD to optimize the model. The momentum was set to 0.09, the weight decay was set to 10^{-4} and the batch size was set to 64. Since the model converges quickly, we have trained 200 epochs for each model and choose the model with the smallest validation loss as the best model.

The imbalanced data causes that no valid features can be learned, and results in low sensitivity of the model to this attribute. As shown in **Tables 3**, **4**, categories with too few samples, such as *partial calcification* and *with cavity*, were not recognized. A given category prediction may have the following four cases: TP, True Positive; FP, False Positive; TN, True Negative; FN, False Negative.

To evaluate the imbalanced categories of each attribute, we use three metrics to score the results. **Accuracy** (ACC) is the basic metric to evaluate the result, which can be calculated as:

$$ACC = \frac{TP + TN}{P + N} \tag{5}$$

Sensitivity (SE), also called the true positive rate, means the probability that a sick person is diagnosed as positive, which can be calculated as:

$$SE = \frac{TP}{TP + FN} \tag{6}$$

The larger the SE value, the more sensitive our model is in diagnosing this category.

Specificity (SP), also called the true negative rate, means the probability that a person who is actually not sick is diagnosed as negative, which can be calculated as:

$$SP = \frac{TN}{FP + TN} \tag{7}$$

The larger the value of SP, the more accurate our model is for the diagnosis of this category.

We average out accuracies of all categories for each attribute, and average the scores of all attributes as the final score to represent the performance of the model.

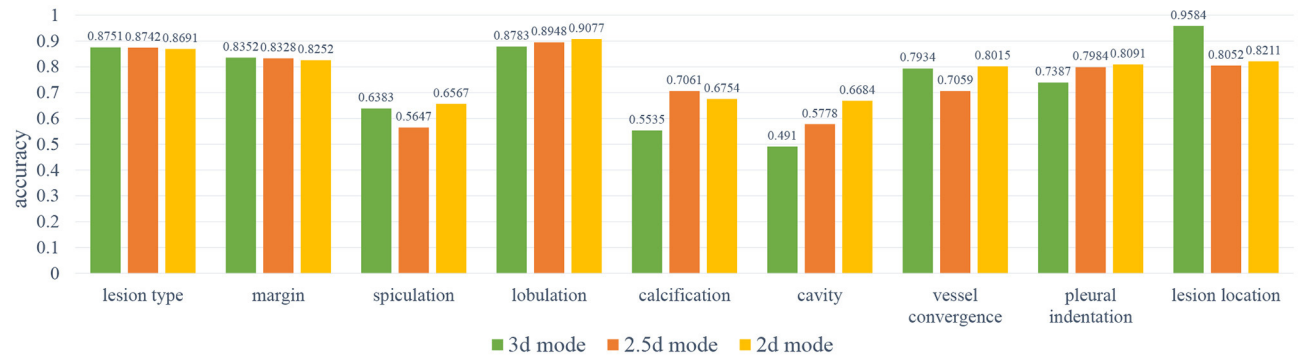

FIGURE 5 | The results of the 2D, 2.5D, 3D input modes. As can be noted, the 3D mode has better results on spiculation, lobulation, cavity, vessel convergence, and pleural indentation.

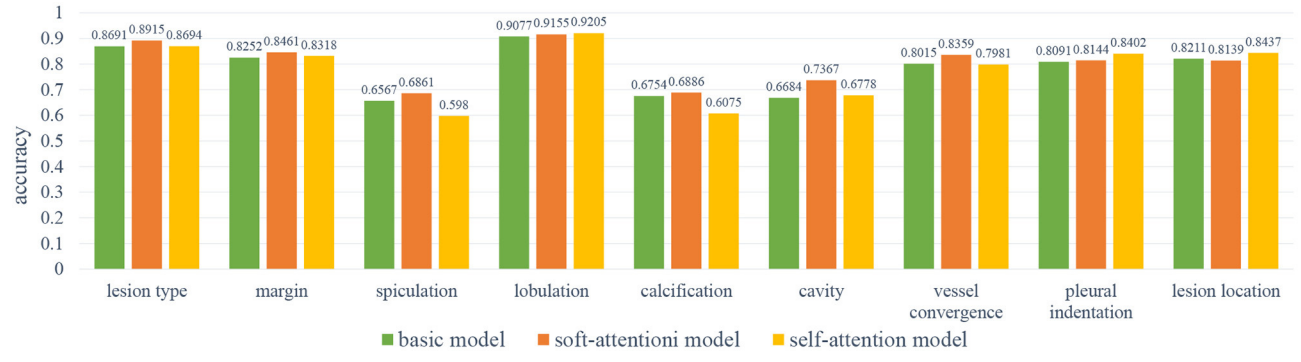

FIGURE 6 | The results of the base model and two attention models. As can be noted, the self-attention module has better results on lobulation, pleural indentation, and lesion location attributes; the soft-attention module has better results on lesion type, margin, spiculation, calcification, cavity, and vessel convergence attributes.

6.1. Results for Input Modes

In order to select the most suitable input mode for the attribute classification of lung lesions, we train the 2D, 2.5D, and 3D model with the same structure described in **Figure 3**. To ensure the fairness of the three models, we do not adjust the hyper-parameters for different models. Each model was trained with 200 epochs and a batch size of 64. To evaluate the performance of the models, we chose the average accuracy of the model with the lowest validation loss as the metric. The average accuracy scores of the 3D, 2.5D, 2D model are 0.7513, 0.7511, and 0.7816; the average sensitivity are 0.7671, 0.7006, and 0.7184; and the average specificity are 0.8305, 0.7995, and 0.8116, respectively. As **Figure 5** shows, the three models have almost the same scores in lesion type and margin, and the model with 2D mode has better scores in spiculation, lobulation, vessel convergence, and pleural indentation. **Table 3** shows the accuracy, sensitivity, and specificity of each category for each attribute. From the experimental results, we note that the higher-level attributes, such as lesion type and lesion location, are more sensitive to the 3D mode and the lower-level attributes, such as spiculation and lobulation, are more sensitive to the 2D mode.

During training, we noticed that the 3D model has more parameters than the 2D models, which led to longer training time and slower convergence. Meanwhile, the 2D model has better average accuracy than the 3D model. So, we chose the 2D mode as the basic model for the following experiments.

6.2. Results for Attention Mechanisms

In order to improve the performance of the basic model, we have used two attention mechanisms to enhance the feature before feeding it to the classifiers. We called the model with the soft-attention module *Soft-Att*, and the model with the self-attention module *Self-Att*. Since the number of parameters of the two attention modules is not large, we use the same hyper-parameters as the basic model to train the two models. Similar to the previous section, we used a batch size of 64 and 200 epochs for training and taking the accuracy of the model with the lowest validation loss as the metric. The average accuracies scores of the basic model, Soft-Att and Self-Att are 0.7816, 0.8032, and 0.7763; the average sensitivities are 0.7184, 0.7183, and 0.7155; and the average specificities are 0.8116, 0.8117, and 0.8128, respectively.

As **Figure 6** shows, the soft-attention module has better results on margin, vessel convergence, lesion type, and spiculation

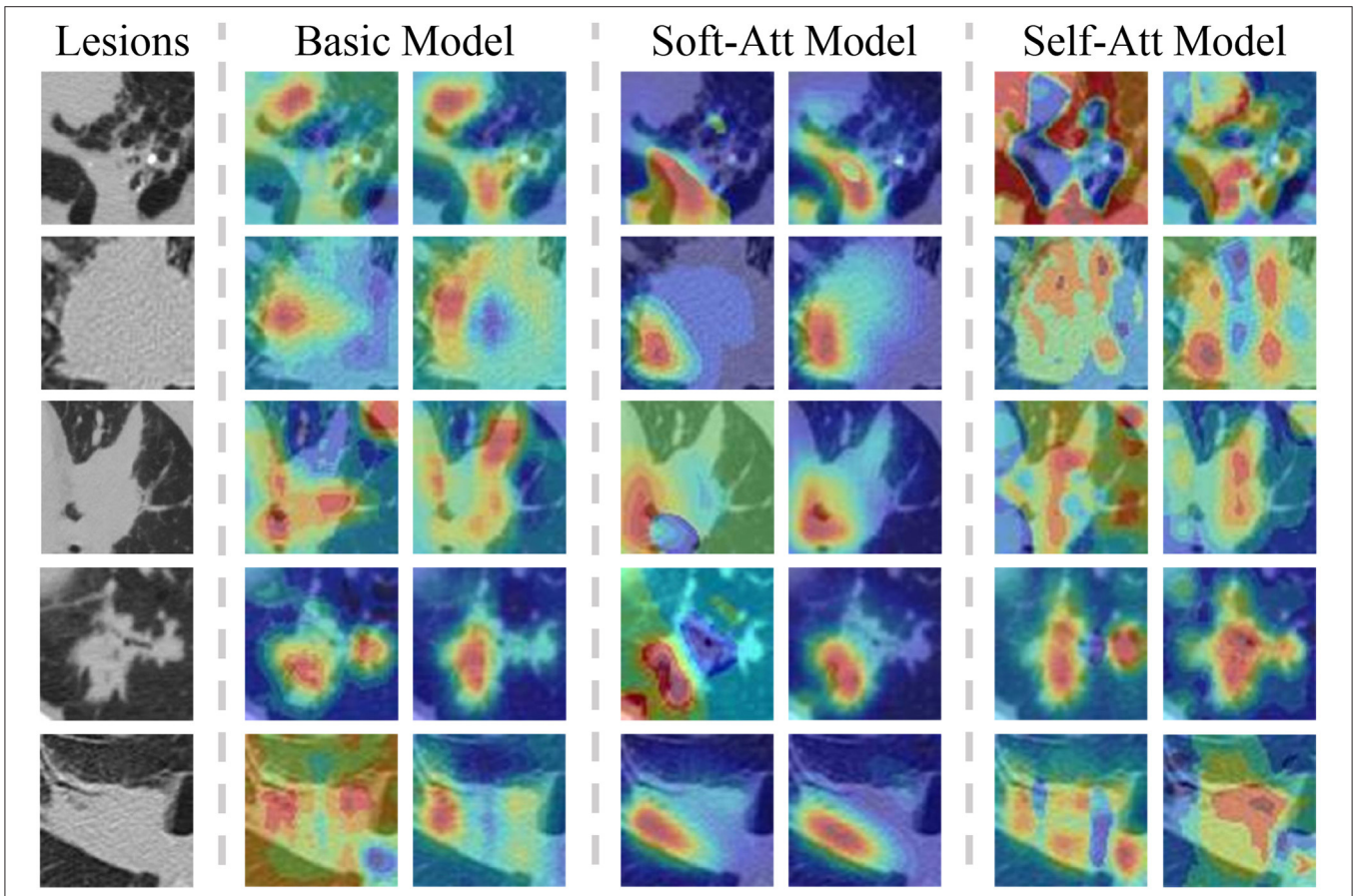

FIGURE 7 | The lesions with two heatmaps from the basic model, soft-Att, self-Att, respectively, in the lesion type and lobulation attributes. The color means the importance of the feature in that position. The red color indicates an important feature. As the figures show, the attention model focuses the features on one point and self-attention spreads the features in the spatial dimension.

attributes, and the self-attention module has better results on lobulation, pleural indentation, and lesion location attributes. Due to the near-zero sensitivity of calcification and cavity attributes, we do not take their accuracy into comparison. As reported in **Table 4**, the two models with attention modules have better performance than the basic model.

The heatmaps in **Figure 7** visualize the attention mechanisms. Compared with the basic model, the red value of soft-attention is concentrated at one point. This is because soft-attention uses higher-layer semantic information to filter the low-layer features, which makes the features spatially smoother and more focused. This is a good feature for high-level attributes because it is concentrated at the point that best reflects the attribute, but it does not fully reflect the local information relationship. Compared with the basic model, the red value of self-attention is more scattered in the spatial dimension. This is because self-attention extracts channel information by compressing spatial information using its own features, and it is more comprehensive in spatial information due to multi-channel fusion. This is a good feature for low-level attributes because its local information relationships are more spatially refined, but because of the noise in the spatial dimension, it may not be appropriate for high-level attributes.

7. CONCLUSION

This paper presents a dataset of lung lesions with fine contour annotation and attribute and explores the correlation between the attributes of the dataset. To demonstrate the contribution of this dataset to the development of CAD systems, we explore two issues of medical data modeling using attribute classification tasks.

One of the issues is the effect of the 2D, 2.5D, 3D input mode on the classification model. The 2D mode works well for low-level attributes that do not require local information relationships between lesions and surrounding tissues, while the 3D mode works better for high-level attributes that require higher contextual relationships. The 2.5D mode is a trade-off between the lightweight of the 2D model and the context information of the 3D model.

The second is the impact of the two attention mechanisms on the model. Soft-attention can better handle the noise in the spatial dimension and concentrate on the features at one point, which is beneficial for the classification of high-level attributes. Self-attention can better integrate the spatial information in the channel dimension, and complement the local relationship in the spatial dimension, which is beneficial for the classification of low-level attributes.

In the future, we mainly want to explore and address the following three issues:

1. For the three categories of cavity, partial calcification, and long spiculation, the sensitivity is almost zero due to the high degree of the category imbalance. We will explore novel methods to improve the accuracy of these three categories.
2. We will use the correlation between attributes to establish a loss function suitable for multi-attribute classification from the statistical learning strategy.
3. There is not a single metric that can well measure the performance of a multi-attribute model. We will build evaluation metrics for multi-task modeling.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the Shanghai Science and Technology Committee (No. 18411952100).

ACKNOWLEDGMENTS

The authors would like to acknowledge all of the contributors to A Dataset of Pulmonary Lesions with Multiple-Level Attributes and Fine Contours.

REFERENCES

1. Altaf F, Islam SMS, Akhtar N, Janjua NK. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*. (2019) 7:99540–72. doi: 10.1109/ACCESS.2019.2929365
2. Zhu W, Liu C, Fan W, Xie X. DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. In: *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*. Lake Tahoe (2018). p. 673–81. doi: 10.1109/WACV.2018.00079
3. Zhu W, Vang YS, Huang Y, Xie X. DeepEM: deep 3D ConvNets with EM for weakly supervised pulmonary nodule detection. In: *Lecture Notes in Computer Science*. Granada (2018). p. 812–20. doi: 10.1007/978-3-030-00934-2_90
4. Setio AAA, Traverso A, de Bel T, Berens MSN, van den Bogaard C, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med Image Anal*. (2017) 42:1–13. doi: 10.1016/j.media.2017.06.015
5. Li Z, Wang C, Han M, Xue Y, Wei W, Li LJ, et al. Thoracic disease identification and localization with limited supervision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT (2018). p. 8290–9. doi: 10.1109/CVPR.2018.00865
6. Masood A, Sheng B, Li P, Hou X, Wei X, Qin J, et al. Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images. *J Biomed Inform*. (2018) 79:117–28. doi: 10.1016/j.jbi.2018.01.005
7. Gonzalez G, Ash SY, Vegas-Sánchez-Ferrero G, Onieva JO, Rahaghi FN, Ross JC, et al. Disease staging and prognosis in smokers using deep learning in chest

computed tomography. *Am J Respirat Crit Care Med*. (2018) 197:193–203. doi: 10.1164/rccm.201705-0860OC
8. Duan J, Schlemper J, Bai W, Dawes TJW, Bello G, Doumou G, et al. Deep nested level sets: fully automated segmentation of cardiac MR images in patients with pulmonary hypertension. In: *Lecture Notes in Computer Science*. Granada (2018). p. 595–603. doi: 10.1007/978-3-030-00937-3_68
9. LaLonde R, Bagci U. Capsules for object segmentation. *arXiv [Preprint]. arXiv:1804.04241*. (2018).
10. Kim J. Lung nodule segmentation with convolutional neural network trained by simple diameter information. In: *Medical Imaging with Deep Learning (MIDL 2018)*. Amsterdam (2018). p. 3–5.
11. Burlutskiy N, Gu F, Wilen LK, Backman M, Micke P. A deep learning framework for automatic diagnosis in lung cancer. *arXiv [Preprint]. arXiv:1807.10466*. (2018).
12. He G. Lung CT Imaging sign classification through deep learning on small data. *arXiv [Preprint]. arXiv:1903.00183*. (2019).
13. Gao M, Bagci U, Lu L, Wu A, Buty M, Shin HC, et al. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Comput Methods Biomech Biomed Eng Imaging Visual*. (2018) 6:1–6. doi: 10.1080/21681163.2015.1124249
14. Song Y, Cai W, Zhou Y, Feng DD. Feature-based image patch approximation for lung tissue classification. *IEEE Trans Med Imaging*. (2013) 32:797–808. doi: 10.1109/TMI.2013.2241448
15. Wang X, Chen H, Gan C, Lin H, Dou Q, Tsougenis E, et al. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Trans Cybern*. (2020). 50:3950–62. doi: 10.1109/TCYB.2019.2935141
16. Lafarge MW, Moeskops P, Veta M, Pluim JPW, Eppenhof KAJ. Deformable image registration using convolutional neural networks. In: Angelini ED, Landman BA, editors. *Medical Imaging 2018: Image Processing. SPIE*. Houston, TX (2018). p. 27. doi: 10.1117/12.2292443
17. Castillo E, Castillo R, Martinez J, Shenoy M, Guerrero T. Four-dimensional deformable image registration using trajectory modeling. *Phys Med Biol*. (2010) 55:305–27. doi: 10.1088/0031-9155/55/1/018
18. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT Scans. *Med Phys*. (2011) 38:915–31.
19. Han G, Liu X, Han F, Santika INT, Zhao Y, Zhao X, et al. The LISS - A public database of common imaging signs of lung diseases for computer-aided detection and diagnosis research and medical education. *IEEE Trans Biomed Eng*. (2015) 62:648–56. doi: 10.1109/TBME.2014.2363131
20. Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti PA, Müller H. Building a reference multimedia database for interstitial lung diseases. *Comput Med Imaging Graph*. (2012) 36:227–38. doi: 10.1016/j.compmedimag.2011.07.003
21. Dey R, Lu Z, Hong Y. Diagnostic classification of lung nodules using 3D neural networks. In: *Proceedings - International Symposium on Biomedical Imaging*. (2018). p. 774–8. doi: 10.1109/ISBI.2018.8363687
22. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. *Int J Comput Assist Radiol Surg*. (2017) 12:1799–808. doi: 10.1007/s11548-017-1605-6
23. Song Y, Cai W, Huang H, Zhou Y, Feng DD, Wang Y, et al. Large margin local estimate with applications to medical image classification. *IEEE Trans Med Imaging*. (2015) 34:1362–77. doi: 10.1109/TMI.2015.2393954
24. Biffi C, Oktay O, Tarroni G, Bai W, De Marvao A, Doumou G, et al. Learning interpretable anatomical features through deep generative models: application to cardiac remodeling. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Lecture Notes in Computer Science*. Granada: Springer (2018). p. 464–71. doi: 10.1007/978-3-030-00934-2_52
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
26. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. *arXiv [Preprint]. arXiv:1804.03999*. (2018).
27. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT (2018). p. 7132–41. doi: 10.1109/CVPR.2018.00745
28. Roy AG, Navab N, Wachinger C. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In: *Lecture Notes in Computer Science*. Granada (2018). p. 421–9. doi: 10.1007/978-3-030-00928-1_48
29. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In: *5th International Conference on Learning Representations, ICLR*. Toulon (2017).

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 Medicine

ORIGINAL RESEARCH
published: 23 February 2021
doi: 10.3389/fmed.2021.637652

Diagnosis of Fibrosis Using Blood Markers and Logistic Regression in Southeast Asian Patients With Non-alcoholic Fatty Liver Disease

Chao Sang[1†], Hongmei Yan[2,3†], Wah Kheong Chan[4], Xiaopeng Zhu[2], Tao Sun[1], Xinxia Chang[2], Mingfeng Xia[2], Xiaoyang Sun[2], Xiqi Hu[5], Xin Gao[2,3], Wei Jia[1,6], Hua Bian[2,3], Tianlu Chen[1]* and Guoxiang Xie[7]**

[1] Shanghai Key Laboratory of Diabetes Mellitus and Center for Translational Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China, [2] Department of Endocrinology and Metabolism, Zhongshan Hospital, Fudan University, Shanghai, China, [3] Fudan Institute for Metabolic Diseases, Fudan University, Shanghai, China, [4] Gastroenterology and Hepatology Unit, Department of Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia, [5] Department of Pathology, Medical College, Fudan University, Shanghai, China, [6] Hong Kong Traditional Chinese Medicine Phenome Research Centre, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China, [7] Human Metabolomics Institute, Inc., Shenzhen, China

Non-alcoholic fatty liver disease (NAFLD) is one of the main causes of fibrosis. Liver biopsy remains the gold standard for the confirmation of fibrosis in NAFLD patients. Effective and non-invasive diagnosis of advanced fibrosis is essential to disease surveillance and treatment decisions. Herein we used routine medical test markers and logistic regression to differentiate early and advanced fibrosis in NAFLD patients from China, Malaysia, and India ($n_1 = 540$, $n_2 = 147$, and $n_3 = 97$) who were confirmed by liver biopsy. Nine parameters, including age, body mass index, fasting blood glucose, presence of diabetes or impaired fasting glycemia, alanine aminotransferase, γ-glutamyl transferase, triglyceride, and aspartate transaminase/platelet count ratio, were selected by stepwise logistic regression, receiver operating characteristic curve (ROC), and hypothesis testing and were used for model construction. The area under the ROC curve (auROC) of the model was 0.82 for differentiating early and advanced fibrosis (sensitivity = 0.69, when specificity = 0.80) in the discovery set. Its diagnostic ability remained good in the two independent validation sets (auROC = 0.89 and 0.71) and was consistently superior to existing panels such as the FIB-4 and NAFLD fibrosis score. A web-based tool, LiveFbr, was developed for fast access to our model. The new model may serve as an attractive tool for fibrosis classification in NAFLD patients.

Keywords: NAFLD, hepatic fibrosis, advanced fibrosis, FIB-4, NFS, logistic regression

OPEN ACCESS

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Weipeng Jiang,
Shenzhen University General Hospital, China
Qiuqiang Kong,
University of Surrey, United Kingdom
Ruolan Huang,
Shenzhen University General Hospital, China

***Correspondence:**
Guoxiang Xie
xieguoxiang@hmibiotech.com
Hua Bian
bianhuaer@126.com
Tianlu Chen
chentianlu@sjtu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *04 December 2020*
Accepted: *22 January 2021*
Published: *23 February 2021*

Citation:
Sang C, Yan H, Chan WK, Zhu X, Sun T, Chang X, Xia M, Sun X, Hu X, Gao X, Jia W, Bian H, Chen T and Xie G (2021) Diagnosis of Fibrosis Using Blood Markers and Logistic Regression in Southeast Asian Patients With Non-alcoholic Fatty Liver Disease. Front. Med. 8:637652. doi: 10.3389/fmed.2021.637652

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD), the manifestation of metabolic syndrome in the liver that is linked to obesity and insulin resistance, is one of the most frequent chronic liver diseases (CLDs) and affects approximately 6–40% of the general population, depending on the population, ethnicity, and diagnostic criteria (1, 2). Most NAFLD patients have simple steatosis without fibrosis. Diverse stages of fibrosis and/or cirrhosis may develop in the context of non-alcoholic

steatohepatitis (NASH). Advanced fibrosis (stage 3–4) is increasingly recognized as the leading cause of hepatocellular carcinoma and liver transplantation (3). Meanwhile, advanced fibrosis is at an increased risk for liver-related and cardiovascular-related mortality (2, 4). As a consequence, patients with NAFLD should be assessed for the extent of fibrosis, especially the presence of advanced fibrosis, because of its prognostic implications.

Liver biopsy is regarded as the gold standard for the diagnosis and monitoring of hepatic fibrosis progression in patients with NAFLD. However, this invasive procedure cannot be performed routinely in a large-scale population due to its inherent shortcomings (5). In the last decade, a number of non-invasive approaches based on blood markers, such as the aspartate transaminase/alanine transaminase ratio (AST/ALT ratio) (6), AST to platelet ratio index (APRI) (7), FIB-4 (based on age, AST, ALT, and platelet (PLT)] (8), NAFLD fibrosis score [NFS; based on age, body mass index (BMI), impaired fasting glycemia or diabetes (DM/IFG), AST/ALT, PLT, and albumin (ALB)] (9), FibroMeter (10), and others (11), have been applied to predict and distinguish the progression of hepatic fibrosis in CLD patients due to their simple operation, few complications, and widespread application (12). Some of them (or their combinations) have been recommended as an auxiliary method for liver fibrosis and cirrhosis diagnosis and monitoring, treatment selection, and risk stratification in some countries and regions (13), although their universality and performances are still waiting for further assessment in larger and special populations (14–16).

Along with the increasing amounts of biomedical data and the popularity of artificial intelligence, machine learning methods have been actively used to develop various tools for disease state assessment (17–19). For example, our group constructed a gradient boosting (GB) machine learning model to stage liver fibrosis and cirrhosis in patients with hepatitis B virus (n = 576) and hepatitis C virus (n = 484) infection (20). Using the same four parameters of the famous scoring system FIB-4, our method showed steady and significant improvements in comparison with FIB-4. In addition, we quantitatively profiled 98 serum metabolites in 1,006 participants (including 504 CLD patients and 502 normal controls) and identified four serum metabolite markers, taurocholate, tyrosine, valine, and linoelaidic acid, which can reliably evaluate the stage of fibrosis by jointly using two machine learning methods, least absolute shrinkage and selection operator and random forest (RF) (21). The prediction models were steadily superior to existing scoring systems, including the APRI, FIB-4, and AST/ALT ratio, with greater sensitivity, specificity, area under the receiver operating characteristic curve (auROC) and area under the precision–recall curve (auPR). However, in further studies and clinical applications, increasing attention has been given to the limitations of machine learning models. First, the computational process of a model is a "black box" to users, and no formula can be given. This ambiguity has impeded its popularity in clinical practice. Second, the overfitting problem is increasingly recognized in patients with diverse backgrounds. Machine learning models usually require a much higher number of training samples and more independent validation sets (to avoid overfitting) than conventional methods due to their complicated structure and a large number of parameters. As large-scale (e.g., over 2,000) samples of liver biopsy-confirmed NAFLD patients are not easy to obtain, complex machine learning methods are considered to be an over-examination for NAFLD patients. Thus, the contradiction between the sample size demand and the poor compliance of patients could not be solved in the short term.

Logistic regression (LR), a simple and classical method, has been used in thousands of studies for disease status assessment. Considering the limitations of machine learning methods and the practical value of LR, in this report, we constructed an LR model for the differentiation between early and advanced fibrosis in NAFLD patients. Our strengths include the following: (1) Three independent cohorts with sample sizes of 540, 147, and 97 were used for model construction and validation; (2) All the patients were evaluated by liver biopsy; (3) Our model used routine medical test markers that can be obtained during routine medical examinations regardless of the medical condition; (4) Diagnostic performances were examined and compared comprehensively with FIB-4 and NFS; and (5) An integrated web tool, LiveFbr, was developed for biological research and clinical application. This paper is organized as follows: Section Materials and Methods introduces the cohorts, data sets, and methodology for model construction and validation. Section Results introduces the basic characteristics of the cohorts, the process of parameter selection and model construction, and the results of model evaluation. Section Discussion summarizes the work and highlights its strengths and limitations.

MATERIALS AND METHODS

Cohorts and Ethics

A total of 784 patients with hepatic fibrosis from three independent cohorts were enrolled in this study. Except for cohort 1, the other two cohorts were collected prospectively from anonymous data sets of existing studies. The discovery set (cohort 1) comprising 540 participants was recruited by authors from Zhongshan Hospital Affiliated to Fudan University, China. Liver biopsy specimens were acquired from all patients who met the diagnostic criteria for NAFL or NASH and underwent liver biopsy (22). Subjects were excluded from the study if they had any of the following conditions: history of cancer, alcoholic intemperance, or other causes of chronic liver disease. Peripheral venous blood samples were taken after a 12-h fasting period. The samples were provided in a de-identified fashion, and the lab staff who prepared the samples were blinded to the clinical information. This study conformed to the ethical guidelines of the 1975 Declaration of Helsinki, and approval was obtained from the Research Ethics Committee of Zhongshan Hospital Affiliated to Fudan University (no. B2013-132, date: November 2013). Written informed consent was obtained from each participant. Validation set 1 (cohort 2), consisting of 147 patients, and validation set 2 (cohort 3), consisting of 97 patients, were recruited by the author from University of Malaya Medical

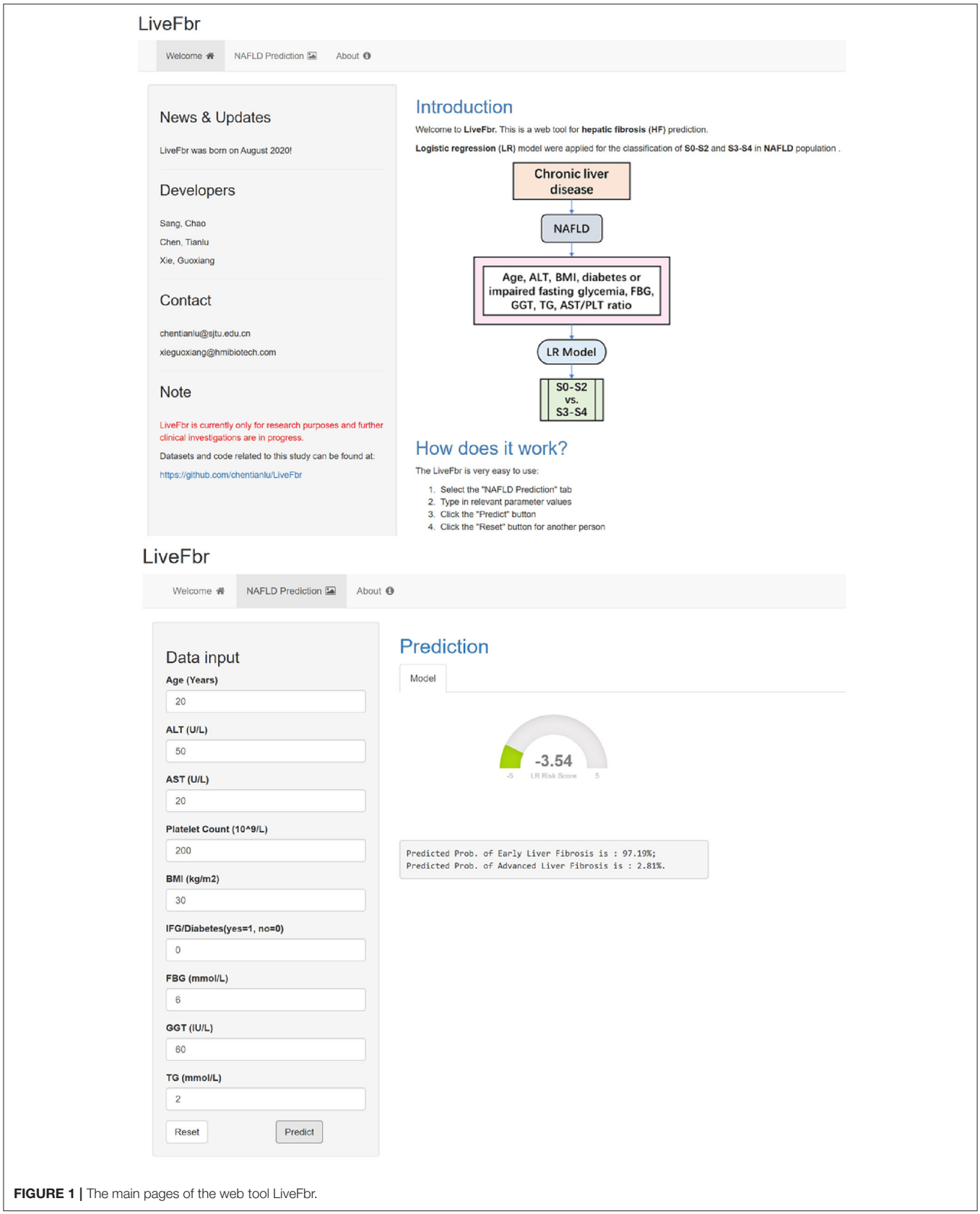

FIGURE 1 | The main pages of the web tool LiveFbr.

Center at different periods (set one was recruited between November 2012 and April 2014, and set 2 began from 2016; for detailed information, please refer to the original publications) (23, 24).

Liver Biopsy

Liver biopsies with ultrasound-guided 1.6-mm-diameter needles were performed by professionally trained operators for patients in the discovery set (cohort 1). For the validation sets (cohorts 2 and 3), percutaneous needle biopsy examinations were performed by one of two experienced operators (WKC and SM) using an 18-G Temno® II semi-automatic biopsy needle (Cardinal Health, Dublin, Ohio, USA) (24). All liver tissue samples of each cohort were examined by an experienced pathologist who was completely blinded to the research design. The non-alcoholic fatty liver disease activity score was used to assess hepatic status based on a standardized histological scoring system (25), namely, included steatosis (0–3), lobular inflammation (0–3), hepatocellular ballooning (0–2), and fibrosis (0–4).

TABLE 1 | Clinical and demographic characteristics of the discovery cohort.

Discovery set	All (n = 540)	Early fibrosis (S0–2) (n = 391)	Advanced fibrosis (S3–4) (n = 149)	p-value
Age (year)	46.76 ± 13.42	44.39 ± 13.44	52.99 ± 11.22	<0.001
ALB (g/L)	4.44 ± 0.41	4.46 ± 0.43	4.37 ± 0.37	0.087
ALT (IU/L)	76.50 ± 49.94	76.25 ± 50.83	77.14 ± 47.69	0.664
AST (IU/L)	47.11 ± 26.40	44.17 ± 25.98	54.81 ± 26.00	<0.001
BMI (kg/m^2)	30.38 ± 5.18	30.23 ± 5.28	30.79 ± 4.87	0.200
FBG (mmol/L)	6.36 ± 2.01	6.10 ± 1.80	7.03 ± 2.38	<0.001
GGT (IU/L)	67.77 ± 60.97	64.98 ± 63.51	75.08 ± 53.25	<0.001
HbA1c (%)	6.61 ± 1.43	6.52 ± 1.44	6.86 ± 1.39	0.001
HDL (mmol/L)	1.11 ± 0.28	1.10 ± 0.26	1.14 ± 0.33	0.115
LDL (mmol/L)	2.95 ± 1.16	3.01 ± 1.20	2.76 ± 1.01	0.134
PLT (10^9/L)	226.70 ± 61.46	235.70 ± 61.20	203.07 ± 55.79	<0.001
TBIL (μmol/L)	12.45 ± 7.08	12.27 ± 7.30	12.93 ± 6.45	0.095
TC (mmol/L)	5.01 ± 1.23	5.06 ± 1.29	4.88 ± 1.06	0.423
TG (mmol/L)	2.02 ± 1.43	2.14 ± 1.58	1.72 ± 0.87	0.001
AST/ALT	0.73 ± 0.38	0.70 ± 0.41	0.80 ± 0.28	<0.001
AST/PLT	0.57 ± 0.39	0.50 ± 0.31	0.75 ± 0.49	<0.001
DM/IFG (no/yes)	233:307	190:201	43:106	<0.001
Sex (M/F)	282:258	221:170	61:88	0.002

Values are expressed as mean ± SD. P-values determined by comparing the characteristics of individuals with early (fibrosis stage 0–2) and advanced fibrosis (fibrosis stage 3–4) were evaluated using an independent-samples t-test or Wilcoxon–Mann–Whitney test. Chi-square test or Fisher's exact test, when appropriate, was used to compare categorical variables.

ALB, albumin; ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; FBG, fasting blood glucose; GGT, gamma-glutamyl transferase; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; PLT, platelet; TBIL, total bilirubin; TC, total cholesterol; TG, triglyceride; DM/IFG, presence of diabetes or impaired fasting glycemia.

Blood Sample Collection and Test

For subjects in the discovery set, routine fasting (12 h) blood samples were collected. Biochemical measurements were performed using standard laboratory procedures. The ALB concentration was examined by the bromocresol green method. Fasting blood glucose (FBG) was assessed by the glucose oxidase method. The level of low-density lipoprotein cholesterol (LDL) was calculated by the Friedewald equation. The concentrations of γ-glutamyltransferase (GGT), high-density lipoprotein cholesterol, total cholesterol, triglyceride (TG), total bilirubin, PLT, ALT, and AST were measured by an automated bioanalyzer (Hitachi 7600, Hitachi, Tokyo, Japan). Glycated

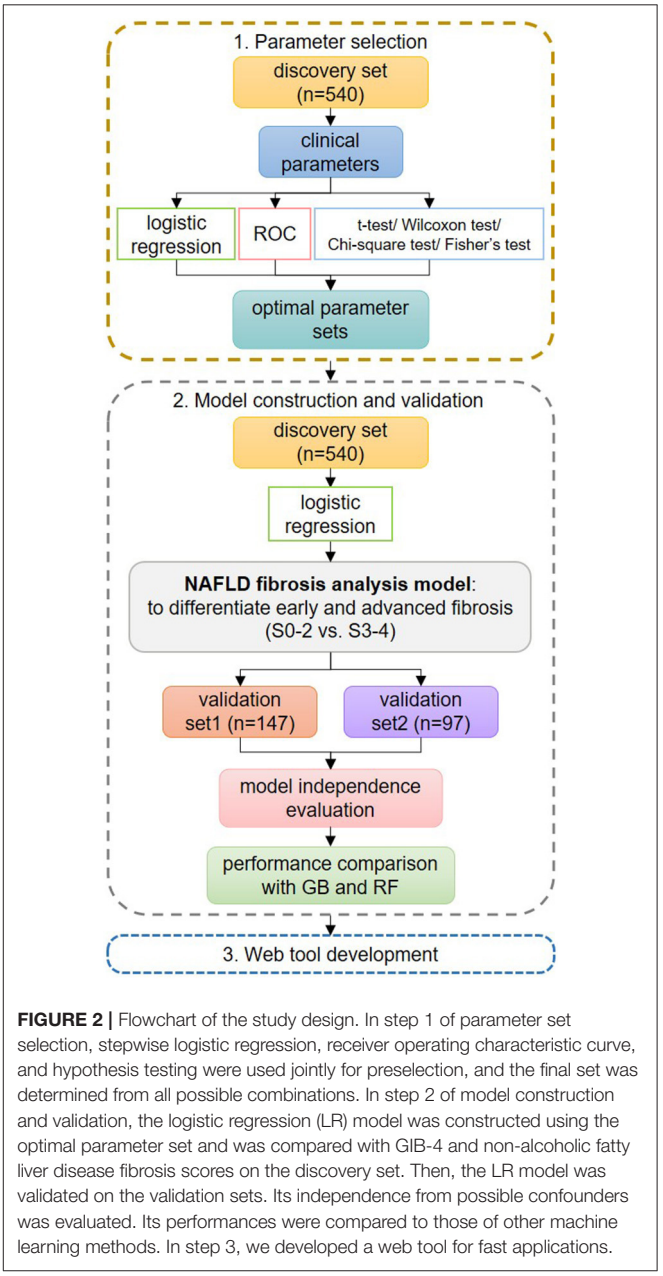

FIGURE 2 | Flowchart of the study design. In step 1 of parameter set selection, stepwise logistic regression, receiver operating characteristic curve, and hypothesis testing were used jointly for preselection, and the final set was determined from all possible combinations. In step 2 of model construction and validation, the logistic regression (LR) model was constructed using the optimal parameter set and was compared with GIB-4 and non-alcoholic fatty liver disease fibrosis scores on the discovery set. Then, the LR model was validated on the validation sets. Its independence from possible confounders was evaluated. Its performances were compared to those of other machine learning methods. In step 3, we developed a web tool for fast applications.

hemoglobin (HbA1c) was estimated by a high-pressure liquid chromatography analyzer (HLC-723 G7, Tosoh Corporation, Japan). Detailed sample collection and test information for the validation sets can be found in the original reports (23, 24).

Model Construction and Validation

Marker Selection

Biological markers are characteristics that are objectively measured and evaluated as indicators of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (26). Marker selection is carried out to eliminate irrelevant or redundant markers (features) and select key features that are truly relevant to the study aim. This step is important to reduce the number of features and to simplify a subsequent model construction. In this study, two steps were taken for marker selection. First, three methods, including stepwise logistic regression, receiver operating characteristic curve analysis, and hypothesis testing [Student's *t*-test for normal parameters, Wilcoxon–Mann–Whitney test for non-normal parameters, and chi-square test or Fisher's exact test (if the expected count is <5 in contingency tables) for categorical parameters] were applied separately for all parameters. The parameters that met two or more conditions (auROC > 0.6, stepwise logistic regression $p < 0.05$, or hypothesis testing $p < 0.05$ between early and advanced fibrosis) were screened out for further selection. Second, all possible combinations among these selected parameters were used to construct numerous LR models. The final optimal parameter set was determined by balancing the number of parameters and the model performances (primarily based on the value of auROC + auPR). The design of our two-step strategy was advanced and effective. The first step reduced the data size and simplified the problem. The second step is time-consuming but necessary, as it is not unusual that a model with fewer parameters performs better than that with more parameters, probably due to the complicated synergistic and competitive relationships among parameters. All these were conducted on the discovery set.

Model Construction and Validation

Based on the optimized parameters, an LR model was established on the full discovery set to differentiate early and advanced fibrosis (S0–2 vs. S3–4). The performances of the LR predictive score were evaluated by ROC and PR curve, auROC, auPR, accuracy, F1 value, and sensitivity (when specificity is 0.8) and were compared with FIB-4 and NFS. The ROC curve is a comprehensive method reflecting sensitivity and specificity. The PR curve is a comprehensive method reflecting recall and precision. auROC and auPR are the area values under these curves. The larger the area is, the better the classification performance. We also employed Wilcoxon tests and box plots to compare FIB-4, NFS, and LR scores in early vs. advanced fibrosis. These results were further validated in two independent validation sets.

To estimate the independence of the LR model on potential confounders, we further applied LR to the predictive score of the model and five parameters that were significantly different between early and advanced fibrosis but were not used in LR model construction.

Considering the good performance of machine learning methods in our previous studies, we constructed an RF and a GB model using the optimal parameter set (with default parameter settings) and compared their performance with that of our LR model.

Code, Data, and Web Tool Availability Statement

R (v 4.0.2) was used for data analysis and figure plotting in this study. The LR, RF, and GB models were built by the stats (v 4.0.2), randomForest (v 4.6–14), and gbm (v 2.1.8) packages, respectively. The data sets and code for result generation

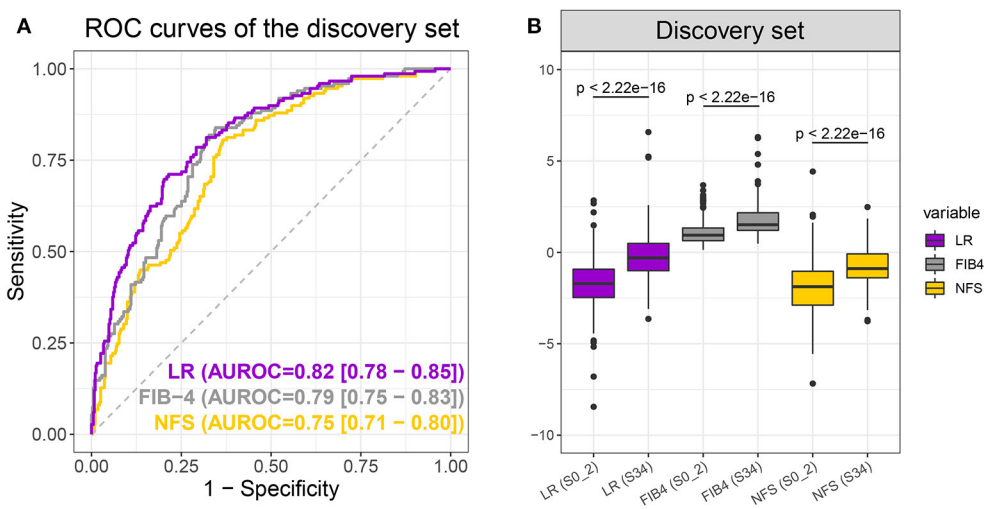

FIGURE 3 | Receiver operating characteristic curves **(A)** of the logistic regression (LR) model (purple), FIB-4 (gray), and non-alcoholic fatty liver disease fibrosis scores (NFS) (yellow) and boxplot **(B)** of LR, FIB-4, and NFS scores when differentiating S0–2 vs. S3–4 on the discovery set. *P*-values were calculated using the Wilcoxon test.

are accessible at https://github.com/chentianlu/LiveFbr. A web-based tool, LiveFbr, has also been developed to provide fast access to our diagnosis system (https://metabolomics.cc.hawaii.edu/software/LiveFbr/, **Figure 1**).

Definitions

The formula of FIB-4 was age × AST (IU/L)/[PLT ($\times 10^9$/L) × $\sqrt{}$ALT (IU/L)] (8). The formula of NFS was $-1.675 + 0.037 \times$ age (years) $+ 0.094 \times$ BMI (kg/m^2) $+ 1.13 \times$ DM/IFG (yes = 1, no = 0) $+ 0.99 \times$ AST/ALT ratio$-0.013 \times$ PLT ($\times 10^9$/L)$-0.66 \times$ ALB (g/dl) (9). The AST/ALT ratio was calculated as AST (IU/L)/ALT (IU/L). The AST/PLT ratio was calculated as AST (IU/L)/PLT ($\times 10^9$/L). The F1 score of a group was calculated as $2PR/(P + R)$, where P and R were the precision and the recall of the group, respectively. The accuracy was calculated as (true positive + true negative)/all samples.

RESULTS

Basic Characteristics of the Discovery Set

A total of 540 biopsy-proven NAFLD patients were involved in model discovery. Two-thirds of the participants, 391 (72.41%), had early fibrosis, and the remaining one-third, 149 (27.59%), were diagnosed with advanced fibrosis. Generally, patients with advanced fibrosis were older, with a higher proportion of females, and had impaired fasting glycemia or the presence of diabetes. In addition, their AST, FBG, GGT, HbA1c, AST/ALT ratio, and AST/PLT ratio levels were higher, and the PLT and TG levels were lower than those of early fibrosis patients (more details are listed in **Table 1**).

Optimal Parameter Set Selection

Two steps were conducted for optimal parameter set selection using all the samples in the discovery set (step 1 in **Figure 2**). After the first step, 14 of the 18 parameters were preselected by logistic regression, ROC, and hypothesis testing: AST, AST/ALT ratio, AST/PLT ratio, DM/IFG, FBG, GGT, PLT, TG, ALT, BMI, LDL, HbA1c, and sex. In the second step, all possible parameter combinations among them were used to construct numerous LR models. Eight parameters were finally selected, balancing the number of parameters used and the values of auPR + auROC, accuracy, and F1 score (**Supplementary Figure 1**). The optimal parameter set consisted of age, ALT, BMI, DM/IFG, FBG, GGT, TG, and AST/PLT ratio.

Model Construction

An LR model was constructed to differentiate early and advanced fibrosis among NAFLD patients using the optimal parameter set on the full discovery set. According to the LR model, the LR score could be obtained as follows: $-5.26952 + 0.041784 \times$ age $-0.01357 \times$ ALT $+ 0.043788 \times$ BMI $+ 0.574987 \times$ DM/IFG $+ 0.089424 \times$ FBG $+ 0.001741 \times$ GGT $-0.490716 \times$ TG $+ 7.738743 \times$ AST/PLT ratio. As **Figure 3A** and **Table 2** show, the auROC and auPR values of our model (0.82 and 0.63, respectively) were higher than those of FIB-4 (0.79 and 0.58) and NFS (0.75 and 0.49), indicating the superiority of the LR model relative to FIB-4 and NFS. We further assessed the group differences in the LR model-generated predictive score and the FIB-4 and NFS scores. All the scores were significantly (Wilcoxon test, $p < 0.05$) different between early and advanced fibrosis (**Figure 3B**). The detailed classification performances of the LR model, FIB-4, and NFS are listed in **Table 2**. As expected, most

TABLE 2 | Performances of the logistic regression (LR) model, FIB-4, and non-alcoholic fatty liver disease fibrosis scores (NFS) in the diagnosis of advanced liver fibrosis.

Method	Accuracy	F1_S0-2	F1_S3-4	auROC	auPR	Specificity	Sensitivity
Discovery set							
LR model	0.78	0.86	0.46	0.82	0.63	0.80	0.69
FIB4_1.45	0.73	0.81	0.52	0.79	0.58	0.80	0.58
FIB4_3.25	0.75	0.85	0.20	0.79	0.58	0.80	0.58
NFS_-1.455	0.68	0.74	0.57	0.75	0.49	0.80	0.47
NFS_0.676	0.74	0.84	0.19	0.75	0.49	0.80	0.47
Validation set 1							
LR model	0.84	0.90	0.60	0.89	0.62	0.80	0.81
FIB4_1.45	0.82	0.88	0.60	0.85	0.60	0.80	0.71
FIB4_3.25	0.79	0.88	0.11	0.85	0.60	0.80	0.71
NFS_-1.455	0.77	0.84	0.59	0.85	0.57	0.80	0.74
NFS_0.676	0.80	0.88	0.17	0.85	0.57	0.80	0.74
Validation set 2							
LR model	0.74	0.82	0.56	0.71	0.61	0.80	0.50
FIB4_1.45	0.65	0.75	0.43	0.63	0.54	0.80	0.38
FIB4_3.25	0.69	0.81	0.12	0.63	0.54	0.80	0.38
NFS_-1.455	0.46	0.45	0.48	0.59	0.39	0.80	0.25
NFS_0.676	0.65	0.78	0.15	0.59	0.39	0.80	0.25

FIB4_1.45 and FIB4_3.25 indicate FIB-4 with different thresholds of 1.45 and 3.25. NFS_-1.455 and NFS_0.676 indicate NFS with different thresholds of −1.455 and 0.676.

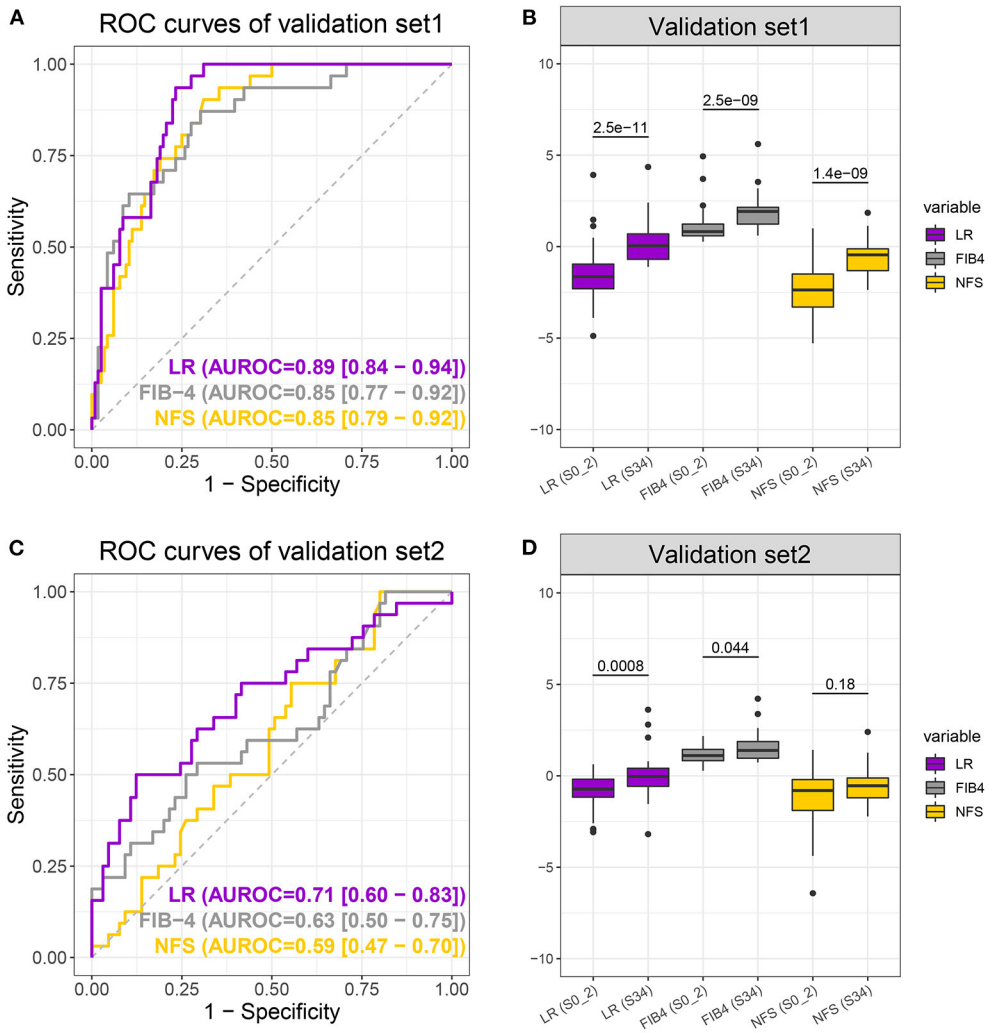

FIGURE 4 | Receiver operating characteristic curves **(A,C)** of the logistic regression (LR) model (purple), FIB-4 (gray), and non-alcoholic fatty liver disease fibrosis scores (NFS) (yellow) and boxplot **(B,D)** of LR, FIB-4, and NFS scores when differentiating S0–2 vs. S3–4 on the validation sets. *P*-values were calculated using the Wilcoxon test.

TABLE 3 | Results of logistic regression (LR) with the LR score only and the LR score + possible confounders.

Dataset	Parameters	*B*	Wald	OR (95% CI)	*P*-value
Discovery set	LR score	1.000	87.602	2.718 (2.225–3.384)	<0.001
Discovery set	LR score + possible confounders	0.981	50.612	2.667 (2.054–3.529)	<0.001
Validation set 1	LR score	1.266	25.085	3.545 (2.258–6.120)	<0.001
Validation set 1	LR score + possible confounders	1.057	7.204	2.879 (1.387–6.554)	0.007
Validation set 2	LR score	0.903	9.159	2.466 (1.461–4.739)	0.002
Validation set 2	LR score + possible confounders	1.139	5.679	3.124 (1.307–8.717)	0.017

Possible confounders were aspartate transaminase (AST), glycated hemoglobin, platelet, AST/alanine transaminase ratio, and sex.

of the criteria of the LR model were the highest compared with those of FIB-4 and NFS.

Model Validation

The LR model obtained by the discovery set was validated in two independent validation sets. Validation set 1 consisted of 147 NAFLD patients, 116 with early fibrosis and 31 with advanced fibrosis, and validation set two consisted of 97 NAFLD patients, 65 with early fibrosis, and 32 with advanced fibrosis. More specific demographic and biological information is available in **Supplementary Table 1**. As expected, the LR model performed best with the highest auROC, auPR, and

sensitivity (when specificity was 0.8) of 0.89, 0.62, and 0.81, respectively, for validation set 1 and 0.71, 0.61, and 0.50, respectively, for validation set 2 (**Figures 4A,C** and **Table 2**). Moreover, the group differences of the LR model were apparently more significant than those of the NFS and FIB-4 in both validation sets (**Figures 4B,D**). In summary, the LR model was consistently superior to FIB-4 and NFS for early and advanced fibrosis classifications.

Model Independence Evaluation

The 14 parameters selected by step 1 were distinctly different between early and advanced fibrosis in the discovery set and were possible confounders for fibrosis staging. Among them, AST, HbA1c, PLT, AST/ALT ratio, and sex were not chosen in our LR model. Hence, logistic regression was applied to the independent assessment of the LR score for these confounders (**Table 3**). The crude OR (95% CI) of the LR score was 2.718 (2.225–3.384) in the discovery set, 3.545 (2.258–6.120) in validation set 1, and 2.466 (1.461–4.739) in validation set 2, with all $p < 0.05$. After adjusting for AST, HbA1c, PLT, AST/ALT ratio, and sex, the LR score was still statistically significant ($p < 0.05$) in the discovery and validation sets, indicating the independence of our model.

Performance Comparison With Other Machine Learning Methods

Two machine learning models, an RF and a GB model, were constructed using the optimal parameter set and the discovery set and then tested by the validation sets. The auROC, auPR, and sensitivity (when specificity was 0.8) of the GB model were 0.83, 0.63, and 0.70, respectively, for the discovery set, 0.83, 0.54, and 0.74, respectively, for validation set 1, and 0.71, 0.60, and 0.47, respectively, for validation set 2. The auROC, auPR, and sensitivity of the RF model were 0.83, 0.76, and 0.68, respectively, for the discovery set, 0.89, 0.59, and 0.81, respectively, for validation set 1, and 0.69, 0.58, and 0.41, respectively, for validation set 2. Comparatively, the LR model had better or comparable auROC, auPR, and sensitivity values than the GB and RF models in the discovery and validation sets.

DISCUSSION

NAFLD has become a significant health problem worldwide; therefore, accurate and reliable assessment of the severity in the NAFLD population is increasingly crucial for treatment decisions and long-term monitoring. A fundamental purpose in the control and management of NAFLD patients is to distinguish those who are more likely to develop significant fibrosis as recently emphasized in the American Association for the Study of Liver Diseases practice guidance, the European Association for the Study of the Liver guidelines, and the Chinese Society of Hepatology guidelines (13, 27, 28). Attempts to establish non-invasive approaches for the stratification of NAFLD patients have yielded various diagnostic panels, indices, and imaging modalities (8, 29, 30) that might be applied in lieu of liver biopsy.

In this study, an LR model was constructed to differentiate early and advanced fibrosis. First, three independent data sets with 784 participants from major ethnic groups in Southeast Asia (Chinese, Malay, and Indian) were used to assess the performance of our model. Our LR model shows admirable diagnostic performance in the discovery and validation sets, although the result in validation set 2 was slightly inferior to that in validation set 1. We carefully compared these data sets and believe that the following differences might lead to different performances: (1) In original studies, validation set 1 was collected for a fibrosis study, and validation set 2 was collected for a steatosis study. The collection criteria for validation set 1 were more similar to those of the discovery set; (2) The patients in validation set 2 were generally older than those in the discovery set and validation set 1; (3) The proportion of patients who had DM or IFG in validation set 2 (no/yes = 10:87) were quite different from that in the discovery set (233:307) and validation set 1 (67:80, **Table 1** and **Supplementary Table 1**). Second, compared with the markers included in FIB-4 and NFS, three additional parameters, FBG, GGT, and TG, were used in our new model. These markers are routine medical test parameters and are also used in other serological diagnostic tools for staging fibrosis or for diagnosing steatosis in patients with NAFLD. Thus, the performance improvement did not come at the cost of the clinical burden. Third, the two-step parameter selection strategy is advanced and practical. In addition to the commonly used difference analysis, all possible combinations of parameters were involved. This is a time-consuming but necessary step to ensure the best solution. Fourth, the performance of our LR model was evaluated comprehensively. Its independence from other parameters was examined. Its diagnostic capability was comparable with some machine learning methods, although LR is sometimes also categorized as a machine learning method.

The limitations of our study include the following: (1) It is well-known that virus infection, NAFLD, heavy drinking, and abnormal immune systems are different etiologies of fibrosis. The patterns of blood parameters and the manner of fibrosis progression in NAFLD patients differ from those in patients with other etiologies. Therefore, our LR model cannot be used directly on other CLD patients. Investigations into different patterns of blood test parameters among CLD patients of various etiologies and the development of general diagnostic tools are ongoing; (2) Longitudinal studies are necessary to further validate the effectiveness and stability of the current findings as well as cross-sectional studies; (3) Our model was validated only by samples from Southeast Asia. Its performances in different data sets were slightly different. Further validation in more and diverse populations is necessary prior to clinical application.

In summary, we constructed a scoring model for the distinction of advanced fibrosis in NAFLD patients. We validated its overall superiority to existing indices and its independence from possible confounders in two independent data sets. The online tool LiveFbr was developed, through which NAFLD patients can obtain auxiliary results of their liver fibrosis severity.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/chentianlu/LiveFbr.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Ethics Committee of Zhongshan Hospital affiliated to Fudan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XG and TC were the principal investigators and designed the study. HB, XG, and WC provided biospecimens and clinical data. TC and CS conducted the data analysis, implemented the methodology, and developed the web tool. HY, XZ, XC, MX, and XS gathered the data and discussed the outcomes. XH was the pathologist. TC, CS, and GX prepared the original draft. WJ, GX, TC, CS, TS, XG, HB, and WC reviewed and edited the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Key Research and Development Program of China (2019YFA0802300 and 2017YFC0906800), the National Natural Science Foundation of China (31972935), the Shenzhen Science, Technology and Innovation Commission [2020(82)], and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01).

ACKNOWLEDGMENTS

We thank the participating hospitals for their assistance in data acquisition and sample collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.637652/full#supplementary-material

REFERENCES

1. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. *J Hepatol.* (2019) 70:151–71. doi: 10.1016/j.jhep.2018.09.014
2. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of non-alcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology.* (2016) 64:73–84. doi: 10.1002/hep.28431
3. Charlton M. Non-alcoholic fatty liver disease: a review of current understanding and future impact. *Clin Gastroenterol Hepatol.* (2004) 2:1048–58. doi: 10.1016/s1542-3565(04)00440-9
4. Sesti G, Sciacqua A, Fiorentino TV, Perticone M, Succurro E, Perticone F. Association between non-invasive fibrosis markers and cardio-vascular organ damage among adults with hepatic steatosis. *PLoS ONE.* (2014) 9:e104941. doi: 10.1371/journal.pone.0104941
5. Regev A, Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, et al. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. *Am J Gastroenterol.* (2002) 97:2614–8. doi: 10.1111/j.1572-0241.2002.06038.x
6. Park SY, Kang KH, Park JH, Lee JH, Cho CM, Tak WY, et al. Clinical efficacy of AST/ALT ratio and platelet counts as predictors of degree of fibrosis in HBV infected patients without clinically evident liver cirrhosis. *Korean J Gastroenterol.* (2004) 43:246–51
7. Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple non-invasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology.* (2003) 38:518–26. doi: 10.1053/jhep.2003.50346
8. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, et al. Development of a simple non-invasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology.* (2006) 43:1317–25. doi: 10.1002/hep.21178
9. Angulo P, Hui JM, Marchesini G, Bugianesi E, George J, Farrell GC, et al. The NAFLD fibrosis score: a non-invasive system that identifies liver fibrosis in patients with NAFLD. *Hepatology.* (2007) 45:846–54. doi: 10.1002/hep.21496
10. Cales P, Boursier J, Ducancelle A, Oberti F, Hubert I, Hunault G, et al. Improved fibrosis staging by elastometry and blood test in chronic hepatitis C. *Liver Int.* (2014) 34:907–17. doi: 10.1111/liv.12327
11. Mansoor S, Collyer E, Alkhouri N. A comprehensive review of non-invasive liver fibrosis tests in pediatric non-alcoholic fatty liver disease. *Curr Gastroenterol Rep.* (2015) 17:23. doi: 10.1007/s11894-015-0447-z
12. Castera L, Friedrich-Rust M, Loomba R. Non-invasive assessment of liver disease in patients with non-alcoholic fatty liver disease. *Gastroenterology.* (2019) 156:1264–81.e4. doi: 10.1053/j.gastro.2018.12.036
13. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. *Hepatology.* (2018) 67:328–57. doi: 10.1002/hep.29367
14. Unalp-Arida A, Ruhl CE. Liver fibrosis scores predict liver disease mortality in the United States population. *Hepatology.* (2017) 66:84–95. doi: 10.1002/hep.29113
15. Le MH, Devaki P, Ha NB, Jun DW, Te HS, Cheung RC, et al. Prevalence of non-alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States. *PLoS ONE.* (2017) 12:e0173499. doi: 10.1371/journal.pone.0173499
16. Yoshihisa A, Sato Y, Yokokawa T, Sato T, Suzuki S, Oikawa M, et al. Liver fibrosis score predicts mortality in heart failure patients with preserved ejection fraction. *ESC Heart Fail.* (2018) 5:262–70. doi: 10.1002/ehf2.12222
17. Berry SE, Valdes AM, Drew DA, Asnicar F, Mazidi M, Wolf J, et al. Human postprandial responses to food and potential for precision nutrition. *Nat Med.* (2020) 26:964–73. doi: 10.1038/s41591-020-0934-0
18. Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat Rev Gastroenterol Hepatol.* (2020) 17:635–48. doi: 10.1038/s41575-020-0327-3
19. Oh TG, Kim SM, Caussy C, Fu T, Guo J, Bassirian S, et al. A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* (2020) 32:878–88.e6. doi: 10.1016/j.cmet.2020.06.005
20. Wei R, Wang J, Wang X, Xie G, Wang Y, Zhang H, et al. Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning. *EBioMedicine.* (2018) 35:124–32. doi: 10.1016/j.ebiom.2018.07.041

21. Xie G, Wang X, Wei R, Wang J, Zhao A, Chen T, et al. Serum metabolite profiles are associated with the presence of advanced liver fibrosis in Chinese patients with chronic hepatitis B viral infection. *BMC Med.* (2020) 18:144. doi: 10.1186/s12916-020-01595-w
22. Bedossa P, Consortium FP. Utility and appropriateness of the fatty liver inhibition of progression (FLIP) algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of non-alcoholic fatty liver disease. *Hepatology.* (2014) 60:565–75. doi: 10.1002/hep.27173
23. Chan WK, Nik Mustapha NR, Mahadeva S. A novel 2-step approach combining the NAFLD fibrosis score and liver stiffness measurement for predicting advanced fibrosis. *Hepatol Int.* (2015) 9:594–602. doi: 10.1007/s12072-014-9596-7
24. Harry S, Lai LL, Nik Mustapha NR, Abdul Aziz YF, Vijayananthan A, Rahmat K, et al. Volumetric liver fat fraction determines grade of steatosis more accurately than controlled attenuation parameter in patients with non-alcoholic fatty liver disease. *Clin Gastroenterol Hepatol.* (2020) 18:945–53.e2. doi: 10.1016/j.cgh.2019.08.023
25. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for non-alcoholic fatty liver disease. *Hepatology.* (2005) 41:1313–21. doi: 10.1002/hep.20701
26. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther.* (2001) 69:89–95. doi: 10.1067/mcp.2001.113989
27. European Association for the Study of the Liver, European Association for the Study of Diabetes, European Association for the Study of Obesity. EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J Hepatol.* (2016) 64:1388–402. doi: 10.1016/j.jhep.2015.11.004
28. National Workshop on Fatty Liver and Alcoholic Liver Disease, Chinese Society of Hepatology, Chinese Medical Association and Fatty Liver Expert Committee, Chinese Medical Doctor Association. Guidelines of prevention and treatment for non-alcoholic fatty liver disease: a 2018 update. *Zhonghua Gan Zang Bing Za Zhi.* (2018) 26:195–203. doi: 10.3760/cma.j.issn.1007-3418.2018.03.008
29. Imbert-Bismut F, Ratziu V, Pieroni L, Charlotte F, Benhamou Y, Poynard T. Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study. *Lancet.* (2001) 357:1069–75. doi: 10.1016/s0140-6736(00)04258-6
30. Brancatelli G, Federle MP, Ambrosini R, Lagalla R, Carriero A, Midiri M, et al. Cirrhosis: CT and MR imaging evaluation. *Eur J Radiol.* (2007) 61:57–69. doi: 10.1016/j.ejrad.2006.11.003

Conflict of Interest: GX was employed by Human Metabolomics Institute 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 Digital Health

REVIEW
published: 25 February 2021
doi: 10.3389/fdgth.2021.639444

Trends in Heart-Rate Variability Signal Analysis

Syem Ishaque, Naimul Khan and Sri Krishnan*

Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada

Heart rate variability (HRV) is the rate of variability between each heartbeat with respect to time. It is used to analyse the Autonomic Nervous System (ANS), a control system used to modulate the body's unconscious action such as cardiac function, respiration, digestion, blood pressure, urination, and dilation/constriction of the pupil. This review article presents a summary and analysis of various research works that analyzed HRV associated with morbidity, pain, drowsiness, stress and exercise through signal processing and machine learning methods. The points of emphasis with regards to HRV research as well as the gaps associated with processes which can be improved to enhance the quality of the research have been discussed meticulously. Restricting the physiological signals to Electrocardiogram (ECG), Electrodermal activity (EDA), photoplethysmography (PPG), and respiration (RESP) analysis resulted in 25 articles which examined the cause and effect of increased/reduced HRV. Reduced HRV was generally associated with increased morbidity and stress. High HRV normally indicated good health, and in some instances, it could signify clinical events of interest such as drowsiness. Effective analysis of HRV during ambulatory and motion situations such as exercise, video gaming, and driving could have a significant impact toward improving social well-being. Detection of HRV in motion is far from perfect, situations involving exercise or driving reported accuracy as high as 85% and as low as 59%. HRV detection in motion can be improved further by harnessing the advancements in machine learning techniques.

Keywords: heart rate variability, wireless sensors, drowsiness, stress, morbidity, exercise, machine learning

OPEN ACCESS

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Tao Chen,
Southeast University, China
Hao Wang,
Shenzhen University General Hospital, China
Shuai Yu,
Fudan University, China

****Correspondence:***
Syem Ishaque
sishaque@ryerson.ca

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

Received: *09 December 2020*
Accepted: *02 February 2021*
Published: *25 February 2021*

Citation:
Ishaque S, Khan N and Krishnan S (2021) Trends in Heart-Rate Variability Signal Analysis. Front. Digit. Health 3:639444. doi: 10.3389/fdgth.2021.639444

1. INTRODUCTION

HRV has been associated with many research studies involving morbidity and mortality, stress, fatigue and athletic performance. HRV is primarily used to assess the function of the autonomic nervous system (ANS), it consists of the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS) which coordinates the activities of the body's unconscious actions as a part of the peripheral nervous system. SNS is known as the fight and flight response, it operates within the middle of the spinal cord and activates in response to stress causing an increase in HR, constriction of blood vessels and an increase in blood pressure in order to maintain homeostasis, a healthy/stable state of the body. PNS is known as the rest and digest mechanism, the activities of the PNS contradicts SNS, it relaxes the heart which slows down the heart rate, lowers stress and decreases blood pressure. SNS and PNS work together to maintain a balance, also known as the sympathovagal balance, allowing humans to be safe and sound or an imbalance would indicate abnormalities associated with the heart (1). Time and frequency domain methods

are two of the most common approaches used to accurately assess the function of the ANS (2). Time domain parameters include features such as: (a) standard deviation of NN (normal R-peaks)- intervals (SDNN), (b) square root of the mean of the sum of the squares of differences between successive NN-intervals (RMSSD) and, (c) proportion of the number of NN-interval difference of successive NN- interval which are greater than 50 ms divided by the total number of NN-interval (PNN50) (3). NN intervals were used instead of RR intervals in order to emphasize the use of normal R-peaks. These methods can efficiently analyze HRV through the analysis of the R-R interval which can indicate changes in the HR due to the activities of the SNS or PNS but it's not a sufficient method to discriminate between the SNS and PNS (3). Frequency domain methods such as LF (0.04–0.15 Hz) and HF (0.15–0.4 Hz), LF/HF ratio are often utilized to differentiate between the activity of the SNS and PNS. LF primarily indicates the activity of the SNS but is also partially associated with the activity of the PNS, while HF indicates the activity of the PNS, and their ratio LF/HF is used to determine the sympathovagal balance (3). These indices have made it possible to detect many abnormalities, diseases and possible indication of mortality due to the distorted activity of the heart and the peripheral nervous system. HRV has been used for various applications in research studies which include: analysis of mental and physical stress, classification of drowsiness and other sleep states, analysis of athletic performance and fatigue, studying the correlation between a sedentary lifestyle and mental/physical well-being and analysis of anxiety and depression and various other morbidities associated with reduced HRV.

Kim et al. (4) presented a review paper to analyze HRV and stress, the study described the physiological function associated with stress, as well as HRV related to specific parts of the brain/heart anatomy responsible for the changes associated with stress. The paper presented information related to the anatomy/physiology behind stress, but neglected trends in wearable devices used for data collection, different types of signal processing algorithms used for HRV feature extraction and analysis, machine learning algorithms used for classification of pathologies, wireless monitoring of HRV to improve the health care system and ultimately patient's health and the various applications associated with HRV research (as shown in **Figure 1**).

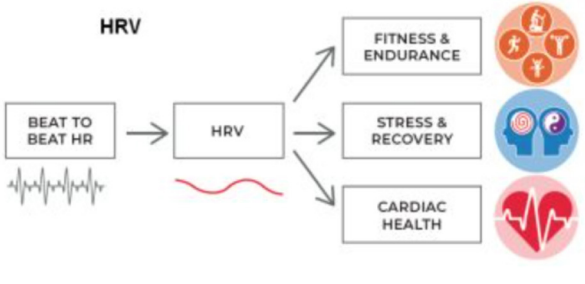

FIGURE 1 | Some important applications of HRV (5).

This article will analyze the various abnormalities associated with HRV, their detection and analysis using an ECG (electrocardiogram), Respiration, GSR and other wearable devices. The impact of pathologies on the human body and mental state as well as the possible gaps that are associated with each research study.

2. METHODS

The literature survey was performed through Ryerson University Library and Archives (RULA) online system. PubMed, IEEE Xplore, Web of Science (WoS), Scopus were the primary search databases directed from RULA. The search was allocated toward HRV studies using ECG, EDA, RESP, PPG signal analysis, few papers involved the analysis of EEG or EOG, but were not considered to present information primarily based on ECG, EDA, RESP, and PPG signal analysis. All the reviewed articles were published after 2010 to present information which is not outdated, except one paper which was used to present the function of time and frequency domain analysis. The relevant papers which were reviewed and summarized described the morbid conditions/situation associated with HRV in depth and in detail, any paper which only briefly discussed HRV were not considered. Papers which primarily focused on factors outside of HRV were also not considered. More than 70 papers were reviewed but most of them were not considered for meta-analysis since they did not provide an in-depth analysis of HRV to examine cardiac pathologies, exercise or drowsiness. Accounting for repetitive topics, 18 major concepts were discussed in depth from 25 articles (as shown in **Figure 2**). The gaps associated with each article were acknowledged and presented.

HRV has a wide range of applications, some of those applications were presented in **Table 1**. The upcoming sections will scrutinize various research experiments which transpired through the analysis of HRV, investigate the changes within a patient's/subject's HRV due to certain activities and morbidities. It will also examine the void and inconsistency of each research study and outline future direction for HRV research, areas which requires more attention in order to become a more efficient procedure which can have a positive impact on people's lives and prevent chaotic outcomes.

3. TRENDS IN HEART RATE VARIABILITY

In this section, we discuss the trends and evolution of HRV from the oldest upto the most recent research conducted. HRV is not a new topic by any means, initial research on this topic was conducted during the early 1940s. Over the years, along with the significance of HRV analysis, feature extraction and modalities used to assess HRV have also evolved.

Features play an important role in discriminating the underlying function associated with any physiological signal. The evolution of features used to analyze HRV is depicted in **Figure 2**. The earliest feature utilized to analyze HRV was HR from time domain. In 1940, Knox studied the variation in HR due to exercise through mean and standard deviation of

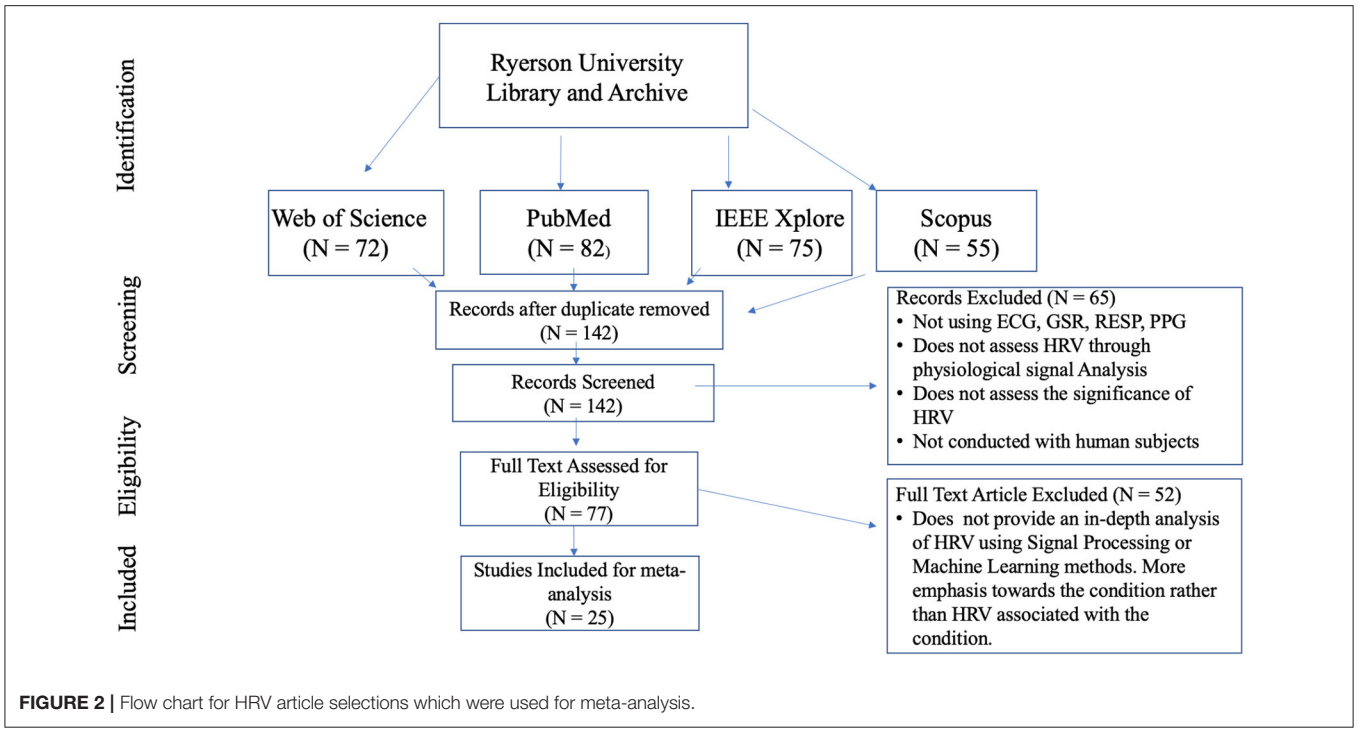

FIGURE 2 | Flow chart for HRV article selections which were used for meta-analysis.

TABLE 1 | Research paper associated with HRV detected using an ECG, type of study, results of HRV, concepts being analyzed.

References	Features	Application	Modality	Notable results	Method of analysis
Rosenberg et al. (6)	LF, HF, LF/HF	1D/2D stress study	ECG	2D accuracy 90%	2D scatter plot.
Blood et al. (7)	LF, HF, LF/HF	Depression	ECG	HRV decreases	Frequency Domain
Molina et al. (8)	RMSSD, LF	Posture	ECG	HRV Reduced	Time Domain
Leti and Bricout (9)	RMSSD, LF	Overtraining	ECG	SNS Dominant	Time, Frequency
Walker et al. (10)	SDNN, HF	Noise	ECG	HRV Reduced	Time, Frequency
Wang et al. (11)	R-R, LF/HF	CHF	ECG	100% acc	SVM, KNN
Huang et al. (12)	LF, HF	Anxiety	ECG	HRV Reduced	LF, HF
Pinheiro et al. (13)	LF, SDNN	MI	ECG	HRV Reduced	Frequency Domain
Toni et al. (14)	LF/HF, LF, HF	CVD	ECG	HRV Reduced	Frequency Domain
Shi et al. (15)	HR, SDNN	Emotion	ECG	LF/HF inc	Time, Frequency
Ponnusamy et al. (16)	RMSSD, HF	Seizure	ECG	HRV Reduced	Time, Frequency
Howells et al. (17)	HF	Bipolar	ECG	HRV Reduced	HF
Rios et al. (18)	R-R, RMSSD	Drowsiness	ECG	HRV Inc	Time Domain
Jung et al. (19)	RMSSD, HF	Fatigue	ECG	HRV Reduced	Time, Frequency
Rahim et al. (20)	LF, HF, LF/HF	Drowsiness	ECG, PPG	HRV Reduced	Frequency Domain
Georgiou et al. (21)	RMSSD, HF	Exercise	ECG, PPG	91–99% acc	Time, Frequency
Gontier (22)	LF, HF, LF/HF	Mind Wander	ECG	LF dec	Time, Frequency
Vicente et al. (23)	LF, HF, LF/HF	Drowsiness	ECG	98% spec	LDA
He et al. (24)	ApEn, LF	Stress	ECG	17.3% err	CNN
Schmidt et al. (25)	LF, HF, ST	Stress	ECG, GSR	80% (3 labels)	Adaboost
Cho et al. (26)	SCL, LF/HF	Stress	GSR, PPG	95% acc	KELM NN

each subject's pulse rate (27). This translated to classification of abnormal variability associated with cardiac pathology. In 1958, Simonson studied the amplitude of the QRS complex (28). He derived the mean and SD associated with normal subjects and differentiated them from patients with cardiac pathology. HRV was more distinguishable using animal studies, due to the level of invasiveness allowed for animals. In 1968, Lynch studied the variation in HRV due to shock applied to dogs (29). The data

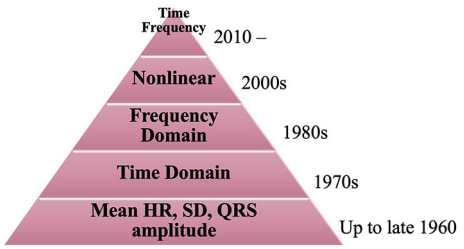

FIGURE 3 | Evolution of feature analysis for HRV from 1940 to 2020.

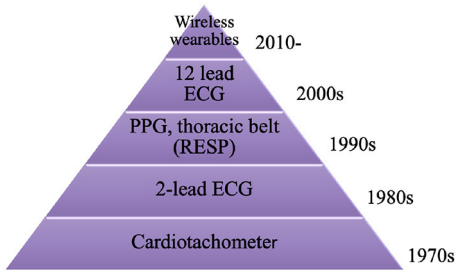

FIGURE 4 | Evolution of medical devices utilized for HRV data acquisition.

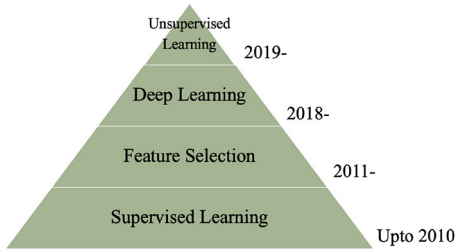

FIGURE 5 | Trends in machine learning for HRV classification.

was analyzed using mean and SD of heart rate. A major change occurred around 1969–1970s, R-R intervals were emphasized for their ability to better analyze HRV from ECG which led to the development of time domain features such as RMSSD, pNN50, and SDNN. In 1977, Rompelman et al. presented a literature which compared the various methods used to analyze HRV and demonstrated that R-R intervals were more accurate for measuring HRV in comparison to HR (30). Researchers didn't just stop there, during the 1990s R-R were deemed less effective in comparison to spectral analysis methods. More studies were conducted, which primarily assessed PSD features such as LF, HF and LF/HF associated with ANS impairment due to cardiac pathologies (30, 31). In 2006, Poincaré plots were introduced to present a visual representation of non-linear scatter plots corresponding to cardiac pathologies and reduced HRV (32). Recently joint time-frequency is a recurring trend which is gaining a lot of attention from researchers (2). It is capable of tracking instant changes in HRV through a shorter period, which can effectively diagnose exercise and cardiovascular diseases. **Figure 3** depicts the evolution of HRV feature analysis from 1940 to 2020.

Figure 4 delineates the evolution of healthcare devices used to detect physiological signals, which can be analyzed to assess HRV. Data collection is the key ingredient which allows researchers to analyze and detect cardiac pathologies associated with an impaired HRV. Upto the 1980s,
cardiotachometer were most commonly used to record a person's electrical signal and record their HR for HRV research (33). Although ECG was developed in 1924, it took about 60 years for them to become affordable for public research. 2 lead ECG's were typically used during the 1980s for HRV research (34). HRV was not just related to heart beat, it also involved blood pressure, mental activity and respiration.

From the 1990s and onwards, HRV research became more diverse. HRV was also analyzed by measuring BP and respiration using PPG and thoracic belt (35). This expanded theories and problems related to impaired HRV, it also added more depth to HRV analysis through information obtained from various physiological signals. Twelve lead ECGs were introduced in 2000, this allowed researchers who were collaborating with clinicians to analyze various cardiac pathologies more effectively (36). The signals obtained were smoother and more efficient in comparison to signals from other ECGs which used fewer electrodes. The current trend involves the use of wearable devices to detect physiological signals, these are much more flexible and portable in comparison to the traditional ECG and PPG devices (25, 37).

Figure 5 describes the common techniques used to classify HRV using machine learning algorithms from 2010 to present. Machine learning has been part of many research studies since the mid 2000's. Although it was initially developed in 1950, supervised methods did not become popular until the 2000s. Literature for machine learning was nothing less than an instant success, within the past decade there have been numerous books, literature, research papers, industrial work and health care innovation based on machine learning. It's hard to pinpoint a specific focus in this domain, so we narrowed the timeline to beyond 2010 and focused on common machine learning topics that were the focus for many research conducted on HRV. Supervised learning has been the most common method to classify various cardiac pathologies and symptoms related to HRV since 2010 (37, 38). Supervised models learn the data and predict labels through learned mapping, which allow models such as DT, LDA, and SVM to predict labels based on corresponding features (39). Many research papers in 2011 revolved around identifying the most important features through feature selection algorithms, in order to obtain better classification accuracy and reduce classification time (39, 40). In addition to automatic diagnosis and classification, researchers have implemented shorter windows to extract features associated with physiological function from real-time (41, 42). Deep learning has been utilized more often for HRV research from 2018 to improve automatic classification through real-time. They are capable of detecting hidden patterns from the input through hidden layers, iteratively

minimizing errors in data prior to classification. This makes the algorithm more efficient for extracting relevant information related to the topic being analyzed, improves classification accuracy and requires less features for real-time classification (24, 43). An emerging trend on the rise from 2019 is the use of unsupervised deep learning to classify mental stress associated with HRV using autoencoder (44). Self organizing map (SOM) is a dimensional reduction method trained through unsupervised learning, which can indicate the most effective features required to classify stress with high accuracy (26).

4. HRV TRENDS FOR DATA COLLECTION

This sections illustrates the various data collection methods used to detect and analyze HRV. **Table 2** reveals the biomedical devices utilized, how they made a significant contribution to the corresponding research and their limitations. Wearable devices are recurrently used in recent HRV research, further indicating the emphasis on remote and wireless monitoring of HRV, in order to make life easier and improve monitoring the health of patients suffering from severe cardiac diseases.

4.1. Smartphones and HRV

Recent smartphones are more than just a device used for communication and listening to music, these devices include embedded sensors, accelerometers, microphones, digital camera, and various apps based on measuring the affective state (neural, emotion, stress) of an individual. These features allowed researchers to conduct valuable experiments which required wireless monitoring of physiological activity, position, speech patterns, facial expression and affective state, in order to analyze stress levels, behavior and emotion at anytime and anywhere, thus promoting better human health and well-being (45, 46).

Prolonged work periods without sufficient rest/recovery periods can reduce happiness and lead to chronic stress due to mental workload (45). Recent development in technology which integrates artificial intelligence/machine learning (AI/ML) provides insight about a persons stress level at work, during social encounters and sleep. Muaremi et al. (45) utilized smartphones to collect audio, communication and physical activity data during work periods and a wearable Wooho chest belt was used to collect HRV data during sleep. They were able to classify stress using HRV features with only 59% accuracy, indicating that although these advancements are quite fascinating and promotes a healthier lifestyle, it wouldn't be considered effective or rational to use such methods to monitor the health of subjects who are suffering from chronic stress or impaired HRV. The most critical aspect of wearable sensors is their inability to produce accurate data. Utilizing such methods would only seem feasible for empirical studies. They are nowhere near the level required to be effective for use by people suffering from stress or impaired HRV. Smartphones are not designed to promote a healthy lifestyle unlike a wearable ECG sensor, using it for the purpose of diagnosing work stress would require further modification of the design, which would make it more adaptable for health care interventions.

4.2. Wearable Devices and HRV

Smartphones and wireless ECG, EEG, and EDA devices would make it possible to detect cardiovascular diseases associated with HRV impairment before it becomes chronic and fatal (46). They make it feasible for health practitioners and people suffering from various cardiovascular diseases (Diebetes, Hypertension) to act proactively and minimize severe outcomes by monitoring their physiological activity throughout the day, including during sleep. Machine learning enable them to predict stress and negative emotions associated with their daily activities, minimizing certain activities may lead to a greater level of productivity and a better sense well-being.

ECG is the most commonly used device with respect to HRV detection (6, 21, 25, 37). Rosenberg et al. (6) utilized a wireless ECG sensor during various situations to measure stress response associated with conference presentations, mental stress test, emergency, and pain. Schmidt et al. (25) utilized Emphatica E4 to measure BVP, EDA, ACC, and TEMP and RespiBAN to detect respiration and ACC (accelerometer). The data collected was used to develop WESAD, a public database which consists of data required to effectively analyze affective states and stress. Cho et al. (26) analyzed HRV, skin conductance (SC)/sweat and skin temperature (SKT) through data collected using a PPG, EDA, and SKT, respectively. They were able to classify stress with high accuracy, using a novel feed forward neural network algorithm and integrated features. Georgiou et al. (21) revealed that wearable devices can detect HRV at rest with 85% accuracy using a PPG and 99% accuracy using an ECG which deteriorates to 85% accuracy during exercise.

Ambulatory detection of HRV is the current resolve for most researchers who hope to make a pragmatic and positive impact on the health and well-being of patients suffering from CVD, hypertension, diabetes, chronic stress and myocardial infarction. Patients suffering from these pathologies need to be monitored throughout the day in order to prevent a serious calamity. Remote monitoring of HRV would undoubtedly benefit senior or chronic patients, who are suffering from cardiovascular diseases but cannot make the effort to visit the hospital all the time, due to the considerable distance and lack of physical ability.

Schmidt et al. (25) were able to classify binary classes of stress by analyzing data collected through wireless sensors with 93.6% accuracy using multinomial logistic regression model. They were able to classify low, mid and high level of stress with 72% accuracy using a random forest algorithm, further demonstrating that chronic stress is hard to predict, although stress can be distinguished from a relaxed state with high efficiency. Cho et al. (26) were able to detect severe stress with wireless PPG, EDA, and SKT sensors from a VR task with 95% accuracy using a kernel based extreme learning machine (K-ELM) algorithm. Although there were numerous studies which classified stress with high accuracies using HRV features, they completely neglected statistical analysis of the data. Machine learning algorithms cannot differentiate between efficient data and errors. They are highly susceptible to biased predictions which arise from biased training datasets, a high classification accuracy can be achieved from erroneous data, if the training data is biased. Physiological signal analysis and statistical analysis

TABLE 2 | Data collection methods, their Pros and Cons.

References	Modality	Pros	Cons
Rosenberg et al. (6)	Wearable ECG	Detect stress with 90% accuracy	Less effective during pain and non-stationary situations
Blood et al. (7)	Holter ECG	Effectively detect depression and HRV	Accuracy of results
Molina et al. (8)	12-lead ECG	Accurate correlation between HRR and HRV	May cause scar
Leti and Bricout (9)	Polar RS 800	Detect fatigue and HRV in motion	Accuracy of Results
Walker et al. (10)	GE Light ECG	Effectively analyze Noise exposure and HRV	Did not detect correlation between noise and BP
Wang et al. (11)	Wearable ECG	Discriminate between CHF and NSR with 91.3% acc	RMSSD is not accurate
Huang et al. (12)	12-lead ECG	Effectively determine HRV due to stroke and hemodialysis	LF/HF ratio is not accurate
Pinheiro et al. (13)	PTB recorder	Determine prognosis of patients following MI	Cannot deduce causality behind results
Toni et al. (14)	Clickholter ECG	Detect HRV in motion due to antidepressants and exercise	LF/HF, RR are not accurate
Shi et al. (15)	RM6240B ECG	Effectively discriminate between HRV of happiness and sadness	RMSSD, pNN50 and SampEn are not accurate
Howells et al. (17)	MP150 Biopac	Accurately analyzed HRV due to meditation and BD wirelessly	Results lacked most ECG measures
Rios et al. (18)	Gear S, PPG	Possibly recognize drowsiness while in motion	No results were obtained
Jung et al. (19)	ECG sensor	Wireless analysis of HRV due to drowsiness and fatigue	Accuracy of results
Georgiou et al. (21)	ECG,PPG	Analyze HRV with 91-99 % accuracy	Accuracy reduces during motion
Gontier (22)	eMotion Faros	Efficiently detect correlation between awareness and HR	Did not find robust correlations
Vicente et al. (23)	eXim Pro	Detect drowsiness while in motion	Detect drowsiness with 62% sensitivity
He et al. (24)	custom ECG	Detect stress using ulta-short epoch	Accuracy of classification was not revealed
Schmidt et al. (25)	RespiBAN Empatica E4	Detect stress with 93% accuracy	May have resulted from overfitting
Cho et al. (26)	Biopac PPG EDA,UIM	Detect stress with 95% accuracy	Not a viable solution in real-life

can provide an effective corroboration that the data utilized were an efficient representation of a subjects physiological function. Venkatesan et al. (47) developed a novel DENLMS adaptive filter for remote health care applications, in order to remove white noise from ECG signals obtained from patients suffering from cardiac arrhythmia. SVM classifier performed better than other ML algorithms and classified normal/abnormal cardiac arrhythmia with 96% accuracy using HRV features extracted from the preprocessed signal through discrete wavelet transform. Although research is seemingly headed toward the right direction, most wearable ECG devices still require much improvement before they can be used to accurately diagnose heart attack or other cardiovascular diseases. Recent smartwatches did not present accurate information about a

subject's HR with respect to their daily life, research studies which used wearable watches to improve weight loss demonstrated that the device produced an ineffective measurement of a person's HR and did not improve weight loss (48). Wearable devices can provide real-time data which can motivate patients to be more careful and promote better self-management in order to prevent chronic outcomes but affordability, adaptability and functionality are still a major concern with wearable devices, especially if they were to be integrated with ML, which poses a major set back and might be the reason that prevents the deployment of such devices. Wearable devices such as a wearable ECG sensor can be utilized to monitor a person's cardiac signal, HR and HRV, which are indicative of chronic outcomes such as myocardial infarction, but they still require further enhancement before they can be considered an effective method for such diagnosis.

4.3. Drowsiness and HRV

Around 10–30% of all road crashes are associated with fatigue and drowsy driving. Recent smart watches and portable ECGs are efficiently being utilized to antedate drowsiness, in order to alert the driver prior to any possible accidents. Accelerometer and gyrometer has been examined to assess the users HRV and physical activity, which allows for the detection of drowsiness/fatigue prior to the transition to stage 1 sleep (drowsiness) (18). There is a high correlation between PPG and ECG in terms of detecting HR, Lee et al. proposed a method to automatically remove noise from PPG using a PPG strap which can be used to accurately detect HR while driving, PSD can be utilized to detect HRV in frequency domain, making it a simple and effective method to detect drowsiness through a persons HR (49). Physiological signals such as an ECG have been described as the most accurate representation of drowsiness in comparison to vehicle based method (lane position of the vehicle) and behavioral method (yawning, eye blinking) (49). Although it has yet to be fully established, wireless ECG sensors might be capable of effectively detecting drowsiness, while the driver is driving. In addition, GSM modules can be utilized to send continuous signals to the control room, DC motor can be used to control the speed of the vehicle upon drowsy detection since the driver's reaction would be distorted, LCD can be used to monitor the driver's condition and LED in the rear side of the vehicle can signal the vehicle behind the drowsy vehicle to slow down (50). Roy and Venkatasubramanian (51) proposed a similar idea which involved using an accelerometer to detect motion, SMS to send an alert message to the control room and microcontroller to process the analog signal prior to its analysis through labVIEW and Matlab. Research based on drowsy driving is still relatively new in comparison to myocardial infarction and hypertension which has been studied for over 30 years, which is one of the biggest reasons for lack of adequate research concerning drowsy driving. A reliable and accurate method to detect drowsiness while a person is driving is still a part of ongoing research, it makes sense in theory but HRV is complex and becomes more intricate to detect in motion such as exercise (only 78.6–85% accuracy in frequency domain) and it is especially worse during drowsy driving (21). Vicente et al. (23) conducted a study which involved truck drivers using a drowsy detection detector as well as a sleep deprivation detector and the accuracy of the results were 0.59 and 0.62 sensitivity, respectively. The results indicate that when a truck is in motion, there are a lot of errors associated with wireless ECG detection, some parts of the signal were blank while in motion. Specificity and predictivity were 0.98 and 0.96 using a drowsiness episodes detector and 0.88 and 0.80 using a sleep deprivation detector, disclosing that detection of the signal was the hardest part during this process, specifying drowsiness/awake state upon detection was very accurate through the data analysis of ECG signals using the linear discriminant analysis (LDA) algorithm. The biggest impediment with regards to drowsy detection is the level of interference associated with electrodes. Electrodes are often attached to a person which can hinder their movement, driving requires constant steering to maneuver the vehicle, which produces error and loss of signal detection. Other methods which involve sensors attached to steering wheels are also hindered by the constant placement of both hands on the steering wheel. Most vehicle based measures are deemed unreliable and inaccurate. Most empirical methods that provide partial results which are somewhat indicative of a person's HRV are often imprecise due the lack of control associated with driving, wireless devices still require sensors to be attached to a person which hinders a person's ability to drive and move freely. Smartwatches which are capable of detecting a person's heart rate would be the least intrusive while driving, but would require extensive modification and testing before it could be considered a valid option to prevent drowsy driving. The cost to develop a smartwatch capable of interpreting a person's HRV and drowsiness would be much greater than the current wireless ECG sensors, making it a less likely solution for drowsiness detection which results in thousands of casualties each year.

4.4. Video Game and HRV

HCI (human to computer interaction) is one of the various methods utilized for stress analysis, cognitive games such as stroop test are often utilized to assess a subjects ability focus while they are subjected to distraction. Fernandes et al. (52) developed a novel method in order to design a video game FlappyHeartPC which used ECG signals as the input, bridging the gap between human physiology and gaming, such interaction might spark more interest within the user for a boring activity (which is relaxing and beneficial for stress reduction health) such as mediation, fishing, or simply analyzing your physiological signal in a lab. The game design includes a tailor belt worn below the chest with electro-textile electrodes was used as the interface between the sensor and the skin, data acquisition required Bitalino (a specialized data acquisition board), python was used to design the signal processing algorithm used to process/filter the input ECG signal, detect QRS complex and calculate HR. Unity 3D was the engine which made the development of the game possible which can utilize HR as the input for certain physiological analysis (52). The video game is a great innovation which can be utilized for science and excitement but it did not have a specific purpose outside of the gaming business. There have been numerous claims by the gaming industry which proclaims that videos can be utilized to stimulate the brain and improve cognitive abilities associated with memory,

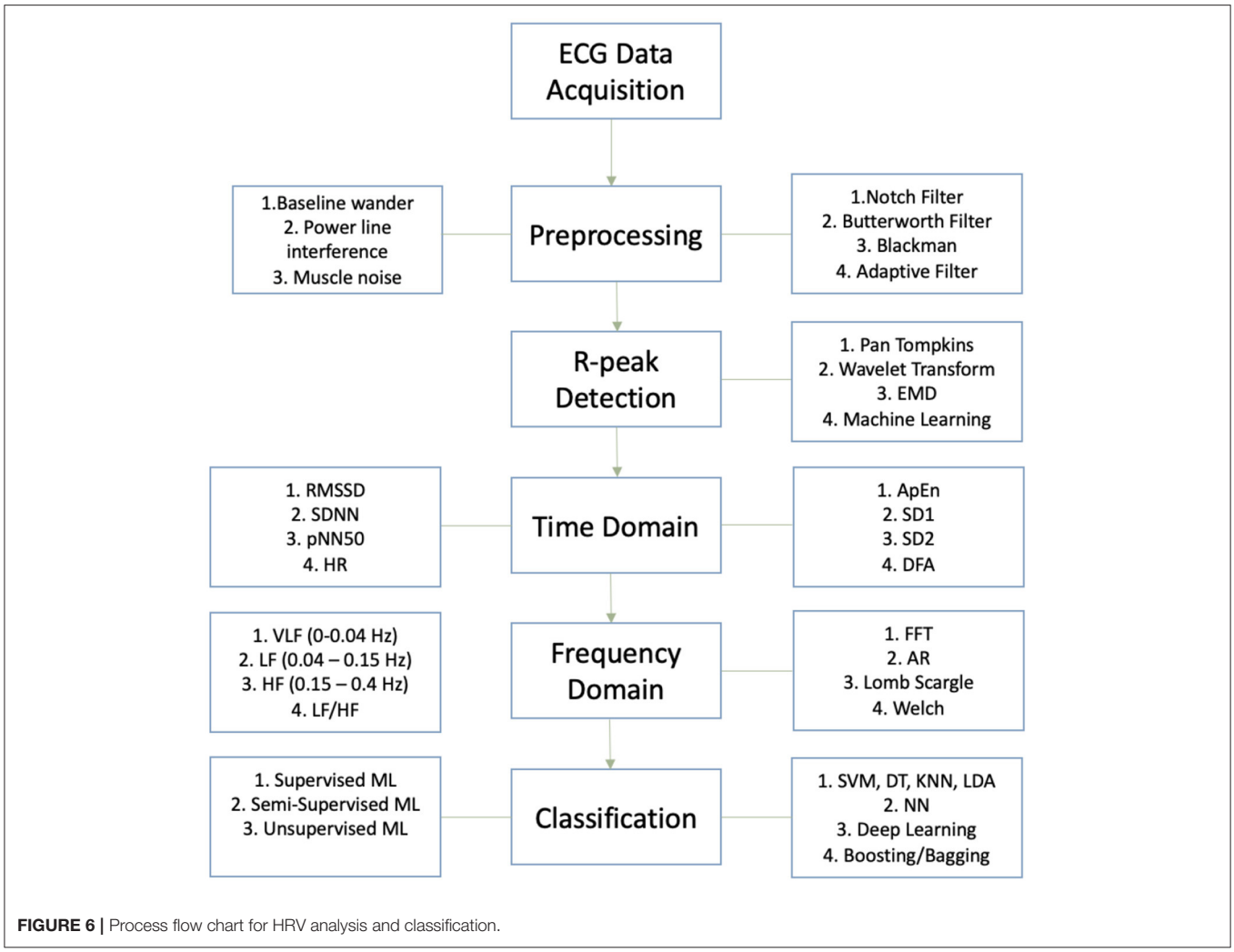

FIGURE 6 | Process flow chart for HRV analysis and classification.

reasoning and processing speed. Unlike 2D video games, 3D video games often allow the user to be notably immersed within the virtual environment and absorb more complex information which stimulates the hippocampus. Analyzing just the heart rate alone would not provide sufficient information to analyze an individual's HRV. Python packages can be used to scrutinize the detected ECG signal through time, frequency and non-linear methods but the extracted data may not be accurate enough to validate the users physiological function. However, it can be utilized to improve human health by implementing a stress detection algorithm into the game. If ML learning can be embedded, there are various possibilities with regards to health care applications such as predicting stress and low HRV, which can also antedate cardiovascular diseases.

5. HRV TRENDS FOR FEATURE ANALYSIS

5.1. HRV and Signal Processing Methods

HRV detection is a complex procedure which requires a series of actions, in order to accurately measure the rate of change associated with the R-R interval obtained from the QRS complex, the raw ECG signal first needs to be filtered, processed and reconstructed. Raw ECG signals need to be filtered in order to remove baseline wander, powerline interference and muscle noise (53, 54). After filtering, the ECG signal is a lot smoother and cleaner, which makes it easier to detect the QRS complex. Researchers have developed and innovated many robust R-peak detection algorithms prior to feature extraction such as: Pan-Tompkins alorithm, wavelet transform algorithm and empirical mode decomposition (EMD) algorithm (55–57). Time domain parameters can be extracted using the R-peaks detected but in order to secure frequency domain parameters, spectral transformation of the QRS complex is required through PSD (power spectral density), which can be obtained through Fast Fourier Transform (represents frequency components), Autoregressive (reduces spectral leakages to improve the resolution of the data), Welch Periodogram and Lomb Scargle Periodogram analysis of the QRS complex. Time domain parameters are statistical evaluations of the ECG signal (presents statistical properties) and frequency domain parameters describe how power (variance) is dispersed as a function of frequency (58). **Figure 6** demonstrates the process required to extract

TABLE 3 | Time and frequency domain features.

Features	Description
HR	The rate of change associated with R-R intervals from HR represents HRV. Increases due to stress
SDNN	The standard deviation of interval between two normal heartbeats (NN). NN measures the total power. Decreases in response to stress. $SDNN = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(RR_j - \overline{RR})^2}$
RMSSD	The root mean square of successive differences between normal heartbeats. Primarily manipulated by PNS activity. $RMSSD = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(RR_{j+1} - \overline{RR_j})^2}$
pNN50	Represents the percentage of the difference associated with NN interval which differ more than 50 ms.It shares a strong correlation with PNS activity, RMSSD, HF
SD1	Non-linear variables derived from the Poincaré plot. Shares a high correlation with HF, RMSSD. Decreases due to stress
SD2	Non-linear variables derived from the Poincaré plot. Shares a high correlation with LF. Increases in response to stress
ApEN	Represents the ratio between SD2 and SD1. Shares a high correlation with LF/HF. Increases due to stress
GSR std	Standard deviation associated with electrodermal activity. Increases during stress
GSR mean	Mean value obtained from measuring the rate of change associated with EDA activity. Increases during stress
Resp Rate	Represents breathing rate, increase in Resp rate leads to increased PNS activity, HF and decreased LF, SNS activity. Increases in response to stress
VLF	Represented within the VLF band (0.0033–0.04 Hz) and it is mediated by SNS activity
LF	Represented through 0.04–0.15 Hz within the PSD, it is mostly used to indicate SNS activity but can specify PNS activity
HF	Represented by the frequency range of 0.15–0.40 Hz and solely indicates PNS activity
LF/HF	Represents ANS activity, increases in response to increased stress and decreased HRV

HRV features from an ECG signal, perform HRV analysis and classify/predict impaired HRV. **Table 3** illustrates the time and frequency domain features used to analyze HRV and their correlation to stress.

5.2. HRV and Stress

As described in **Figure 7**, stress is primarily associated with the activity of the SNS, increased LF (0.04–0.15 Hz) band in frequency domain and reduced HRV. It activates due to perceived danger (such as a deadline, financial worries, exam) and increase in cortisol levels causing the activation of SNS which mobilizes the body's activity under stress in order to react/respond rapidly to any dangerous situations (6). Rosenberg et al. (6) analyzed the levels of stress in response to various situations including public speaking, math, exercise, mediation, pain and cognitive tests. ECG signal obtained through a wireless ECG sensor was processed to measure HR, time and frequency domain features such as SDNN, PNN50, RMSSD, LFn, HFn, LFp, HFp, LFiA, HFiA. LF/HF (normalized, power, instantaneous) were extracted to measure HRV as well as SNS and PNS activity associated with HRV. There are different levels of stress depending on the person's HRV, most often 1D frequency domain methods such as sympathovagal balance (LF/HF ratio) were used since they are more efficient/ simpler than the time domain methods (RMSSD, PNN50, SDNN), which takes longer to assess, although the efficiency of the method can be significantly improved (6, 23). Rosenberg et al. used a 2D scatter plot (LF Vs. HF on a 2D scatter) using multiple variables such as LFn, LFn, LFp, HFp, LFiA, HFiA and their ratio and compared it with the 1D methods (LF/HF ratio or LF, HF computed independently) for different stress tests such as: mental stress, pain, emergency, meditation, and pain. The results concluded that 2D scatter plots were much more efficient than 1D univariate methods, 2D results produced accuracy of 90% or above, whereas 1D methods were around 70%. 1D variables are very linear, unlike stress, they cannot effectively discriminate between 2 tests (such as: Math and exercise) that lead to similar heart rates. However, 2D scatter plots can efficiently differentiate between each ANS activity due to the different activities, resulting in much more efficient results and categorization of ANS activities (SNS and PNS activity) due to different stress states. The accuracy of the experiment is questionable since only 10 participants were used, which is less indicative of the overall population, one individual can have distinct patterns which is not comparable to the rest of the world during exercise or math. Another questionable result would be the result of the HR, exercise should result in a higher HR since the heart starts pumping faster to pump blood to the rest of the body during exercise, in order to match the incremental demand of the exercise, resulting in an increased HR upto 5 min post exercise. LF value during exercise was also rather low, exercise promotes efficient use of one's energy allowing an individual to be more awake/alert throughout the day, which is more associated with the activity within the LF band. A 3D assessment which includes time would probably result in a more comprehensive analysis, effectively specifying the periods associated with increased levels of stress.

5.3. Short-Term Signal Analysis and HRV

Rosenberg et al. (6) have also indicated that time of the epoch used to assess HRV in time domain is very important, 3 min is the minimum epoch that can be used by RMSSD in order to measure fatigue in athletes, but 5 min epoch are optimal for stress analysis, otherwise preprocessing the signal may lead to filtering out valuable information which would result in inadequate, less efficient output and representation of HRV activity associated with stress. Castaldo et al. (41) analyzed HRV using ultra-short

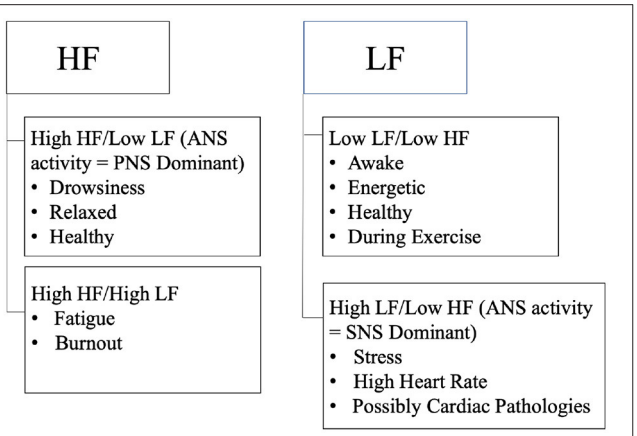

FIGURE 7 | Describes LF and HF associated with stress, drowsiness, awake, and fatigue.

term HRV features in order to assess mental stress in real-time. Theoretically stress is generally associated with perception, it can be due to internal perception such as negative emotions of anger, anxiety, fear, depression and mood swings or it may be induced by external perception of the world around us, such as an upcoming exam, presentation or deadline which causes us to worry, lose sleep and accumulate stress (37). MeanNN, stdHR, HF features resulted in the best accuracy when classified through an automated classifier such as TPOT, which classified HRV using various ML models (SVP, MLP, neighbor search IBK, C4.5, and LDA) and indicated which algorithm was able to classify stress with the highest accuracy (41). Statistical testing is an essential component of every research study, in order to verify, validate and understand the significance of the results obtained. Statistical hypothesis testing legitimizes the efficiency of the results and encourages further expansion of notable methods which can have a significant impact on people suffering from drowsiness, impaired HRV, and cardiovascular diseases (59). Current trends in machine learning hints that there is a bigger initiative for real-time analysis, various algorithms were developed to permit real-time analysis using ultra-short term epochs of 3 min and under (60). In certain cases even 1 min epoch can produce data which can be analyzed to effectively classify HRV using specific features, some features are peripheral, by reducing such features, HRV can be classified in real-time and with higher accuracy (41). Most PSD methods such as FFT, Lomb Scargle periodogram and Autocorrelation are capable of producing useful results which can be used to detect HRV from only 3 min, but it requires the subject to be stationary and stable. Experiments which involve motion (e.g., exercise, driving) produce erroneous results. Most research studies emphasized the use of time domain features to analyse HRV from short-term durations, which is also simpler to extract than frequency domain features. Time domain features are not consistent and often vary, which makes HRV analysis very complicated and flawed. Frequency domain features are more accurate in comparison to time domain methods but do not produce valid data from shorter windows since the rate of change associated with R-R intervals are being compromised as well. Short-term duration does not allow the data to fully grasp the activity of the heart, HRV is derived from the rate of change due to fluctuations in HR, shorter windows produce less data and less accurate results. Short-term duration results in minimizing most of the data which also removes valuable information needed to understand the overall condition of the subject (2, 61). Pre-processing is also limited by short-term durations since most of the data might be filtered out if the data is noisy, which is reasonable from subjects under stress. Short-term data can make a significant contribution to the health of patients suffering from CVD, by allowing them to monitor their heart rate in real-time from a distance using ambulatory ECG sensors but extensive research is needed to find viable solutions which can minimize the motion artifacts and reduce errors.

5.4. Low/Reduced HRV

Lower/reduced HRV transpire as a result of increased SNS activity and reduced PNS. It often infers that higher HR/blood pressure leads to various morbidities and increases the chances of mortality. HRV of patients/subjects suffering from depression is very low, VLF (0.003–0.04 Hz) has been positively associated with depression and it is also one of the strongest indicators of depression (7). Blood et al. were able to make these diagnoses using correlation analysis (scatter plots), which compares the activity of the LF, VLF, and HF due to various symptoms associated with depression. The research study also revealed that low HF (equivalent to low HRV) emanate more anger, sadness, peer problems, and anxiety, while decreased VLF would cause the development of chronic inflammation, and dysregulation of VLF (associated with metabolic process, thermoregulation, renin angiotensin, regulates blood pressure and fluid balance) which would result in more fatigue and depression (7, 27, 62). The research neglected any possible solution to counteract depression, wireless sensors can be incorporated into biofeedback systems in order to monitor a person's HRV and provide feedback to improve their emotional well-being by increasing their HRV. Nexus has developed biofeedback devices which are capable of measuring physiological activity associated with impaired HRV and providing solutions to improve their physiological function. Mendi developed a biofeedback device to strengthen cognitive function associated with low HRV and stress, which can improve depressive symptoms as well. Interaxon also developed a biofeedback device the muse to counteract low HRV and stress through guided meditation. These devices are expensive and would not be considered as a cure for chronic conditions such myocardial infarction but they can improve depressive symptoms which is often associated with prolonged stress and imbalanced physiological parameters associated with impaired ANS activity. Reduced HRV is a risk predictor of heart failure after acute myocardial infarction, a warning sign for diabetic neuropathy, and has been associated with patient suffering from sleep apnea, dilated cardiomyopathy, fetal distress as well as congestive heart failure (11). Decrease in HRV is correlated to reduced SDNN and a shorter R-R interval. Significantly lower LF along with a reduced HRV antedates sudden cardiac death for patients suffering from CHF, due to the impaired activity of

the ANS, which is unable to respond/react accordingly to the treacherous situation. Both time and frequency domain variables (such as SDNN, LFn, HFn, LF/HF) were used as predictor of morbidity/mortality within the study conducted by Wang et al. (11). Moreover lower HRV and vagal tone indicated through low HF, shorter R-R interval and smaller RMSSD values are associated with epileptic seizure (16). A study conducted by Shiro et al. analyzed the correlation between HRV, chronic neck pain and shoulder pain specifically within females (63). Common cause of neck and shoulder pain is repetitive/over work which can cause an increase of intramuscular glutamate and lactate within the traps. Isometric contraction was performed to indicate the effect of muscle load, LF/HF was lower (increased HRV) within the relaxed and pain free subject but it was inactive for the pain group which was attestation of impaired ANS activity (63). Inactivity of LF/HF is not a clear and concise representation of ANS dysfunction, 2D scatter plots may have provided more efficient results. Undetected signals can also produce dormant results, ECG sensors are not as competent when monitoring subjects in motion. Interpolation is capable of estimating rational values which can be used to fill in the missing values. The results would not be perfect but it may produce frequency domain values which can reveal the most likely outcome due to neck and shoulder pain. A research study analyzed HRV due to fatigue, in order to prevent athlete performance burnout and overtraining (9). Competition has been associated with increased LF/HF and SNS dominance, indicating that athlete's may suffer from more fatigue, stress and anxiety during competition (64, 65). Studies revealed that HRV and HF decrease with an increase in age (9). Aerobic training positively impacts HRV and HF, which was indicated through the positive correlation with time domain parameters such as SDNN and RMSSD and HF. Excessive training can cause impairment of the cardiovascular control system, negatively impact a competitors mood/state which has been associated with injury and fatigue, resulting in reduced HRV and HF. Increased SNS activity which is specified through an increase in LF, compensates for reduced cardiac performance and helps recover normal blood flow. High SNS is also associated with fatigue during training which correlates to reduced HRV and HF (64, 65). Two days after the competition, an increased HF suggested a rise in PNS activity and HRV, disseminating that exercise/training improves vagal tone and helps to maintain ANS modulation (64). Unlike Fourier transform which neglects the time-localization information, wavelet transform extract information with respect to time and frequency, which is excellent to detect HRV information which is not stationary. It can detect the instantaneous change associated with HR due to exercise more efficiently than common PSD methods such as fft and AR periodogram which is more effective for frequency domain analysis and stationary processes (2). Missing data and ECG signal recording inactivity is a common problem associated with monitoring HRV in motion and during exercise. Interpolation, reconstruction of large gaps and reconstruction with localized estimation are few methods which can help rectify the data and extract feasible frequency domain features (66). There is a higher probability/occurrence of myocardial infarction associated with older women as a result of lower HRV and ANS dysfunction (13). HRV analysis also revealed that SDNN, RMSSD, triangular index were significantly worse for women than men, additionally reduced HRV is the strongest predictor of myocardial infarction (13, 67). Resting HR is a robust indicator of myocardial infarction and coronary death within women, low HR as well as increased HR associated with depression antedates coronary artery disease. Women and men require different treatments for an accurate prognosis due to sexual dimorphism associated with men and women. Time domain methods are not capable of differentiating between SNS and PNS activity which can make data analysis somewhat biased and based on preconceived assumptions. Statistical t-test or chi squared tests can corroborate the plausibility of the data and help determine whether the results presented are statistically significant (3). Patients suffering from stroke and requiring hemodialysis also indicated a lower HRV, post dialysis presented an increased VLF, LF, TP, and LF/HF ratio (12). VLF is robust in terms of prognosis for CHF. Lower HRV is also associated with adverse cardiac states, increased morbidity and mortality within patients suffering from ESRD (end stage renal disease). Relaxing music such as classical music improved HRV in patients with cardiovascular dysfunction and dementia. Interestingly classical music at high intensity also reduced HRV, although sufficient analysis was not provided. LF was reduced during heavy metal which may indicate that it is harmful and causes increased fatigue. Higher intensity of music increased sympathetic tone on HR, the reaction designate that music is perceived as a threat by the ANS and may induce stress/fatigue (68). The frequency domain data was analyzed via FFT algorithm which is capable of producing miscellaneous results due to its inability to apprehend transient signals through unspecified capture windows. Specific ranges within the capture windows are capable of producing valid results depending on the duration of the transient signal, otherwise it can result in data leakage which distorts the feature values obtained. Bandwidth filtering of the signal was not mentioned, which can lead to aliasing and result in incorrect frequency and amplitude. Do Amaral et al. (68) identified that music can increase or reduce HRV based on the type of music and its impact on HRV. Music therapy involving soothing music improves HR, it has been utilized to improve cardiac function after taking cardiotoxic medication (68, 69). Heavy metal and metal rock reduced HRV and the modulation of the heart indicated through reduced SDNN. Although SDNN is capable of interpreting the overall HRV, it can increase or decrease as a result of decrease in HRV. Its simple to compute but does not provide sufficient information to understand ANS activity associated with reduced HRV (3). Kubios was used to analyze the data, its a software which automatically produces results in time and frequency domain. It uses automatic filters which are likely to produce imprecise results if the signal is very noisy (21).

6. HRV TRENDS USING MACHINE LEARNING

This section discusses the recent studies which classified HRV using machine learning algorithms. **Table 4** demonstrates the

TABLE 4 | Recent publications based on HRV + Machine Learning. The accuracy produced and the theoretical computational cost required by the algorithm.

References	Accuracy (%)	Computational cost	ML algorithm(s)
Castaldo et al. (41)	94,88,94,94	$0(n)$,$0(kd)$,$0(n\log n)$,$0(nd^2)$	MLP, SVM, C4.5,LDA
Cho et al. (70)	90.19	$0(n \cdot k \cdot d)$	CNN
Cho et al. (26)	95	$0(n^4)$	K-ELM
Coutts et al. (71)	83	$0(W)$ $W = 4IH + 4H^2 + 3H + HK$	LTSM
Taye et al. (72)	98.6	$0(W)$ $W = IH + HK$	ANN
Arsalan et al. (73)	92.85	$0(n)$	MLP
Lima et al. (38)	80	$0(n^*\log(n)*d*k)$	Random Forest
Kublanov et al. (74)	91.3,87.8, 87.1,88.2	$0(nd^2)$,$0(kd)$, $0(n^*\log(n)^*d)$, $0(c*d)$	LDA,SVM,DT,NB
Ma et al. (75)	96.58, 98.2	$0(n \cdot k \cdot d)$, $0(n)$	CNN, MLP
Persson et al. (76)	77.5,83.4, 82.4,85.4	$0(nd)$,$0(n^2)$, $0(nt)$,$0(n^*\log(n)^*d^*k)$	KNN, SVM, AdaBoost, RF

accuracy achieved and the computational cost associated each machine learning algorithm. In order to make a significant impact and connect to as many patients as possible, remote monitoring and analysis of HRV needs to improve. Machine learning is revolutionizing society. It is progressing at a very fast rate to make remote monitoring of HRV effective and accessible to everyone. HRV analysis through machine learning is creating a major impact in research and the world at large, making it possible to accurately antedate diseases, lower healthcare cost and help patients make the right decision, with regards to treatments and therapies.

6.1. Stress Classification Through HRV Analysis

Alhitary et al. (37) have indicated that people need a little bit of stress in their life to stay focused, alert and energetic, so that they can solve the problems they face in their daily life. Alhitary et al. (37) also revealed that if people let stress linger around and continue to worry, it can evolve into chronic stress, leading to more anxiety, lack of coordination and reduced level of productivity. If stress is not detected early, it often leads to many heart related diseases such as hypertension and CVD. In addition to increasing the chance of an infection, it is also a major cause of emotional trauma such as depression. Schmidt et al. (25) developed WESAD, a multimodal public dataset using wearble devices, which includes data for stress and affective emotions. They detected the affective states of users through Emphatic machines such as RespiBAN and Empatica E4, which was placed on their chest and wrist, respectively, to assess their neural state (baseline brain activity), stress levels and amusement condition (emotional state, in this scenario humor was induced). Utilizing the machine learning classification algorithm Adaboost, they were able to classify stress/no stress conditions with 93% accuracy using features obtained from physiological signals (e.g., ECG, EDA, Respiration, skin temperature, accelerometer). Adaboost is a boosting classifier which is considered a strong learner, it is made up of cascade of weak learners such as DT. Unlike weak learners, boosting models learn from the training data and iteratively reduce error by adding a weak learner based on the weight associated with the error. It can predict labels with high precision, by adapting to the training data and minimizing errors. It takes longer to train adaboost and it is not effective for learning imbalanced training data (77).

6.2. HRV Analysis Using Random Forest

Lima et al. (38) revealed that research experiments are sometimes unpredictable as LF and LF/HF activity during stress decreased for certain circumstances where stress was detected. Delineating the changes in ANS activity plays a significant role toward preventing CVD and stress. ANS is regulated by the CNS, it comprises multiple neuroanatomical structures. CNS sends a signal to the SA node in order to adjust to physiological arousal, it's also responsible for responding and adapting to environmental changes (38). The structures of the brain influences the activity of the heart. In contrast to the theory that SNS activity increases during stress, LFnu decreased for some subjects during instances of stress. In order to efficiently classify stress and detect the event, they implemented a SVM algorithm which included an optimal hyperplane to separate subjects whose LFnu increased and decreased during stress (38). There was also a contradictory decrease in LF, LF/HF ratio during stress phases. Using time domain HRV features such as: HR, RR-interval and SD1/SD2, they were able to classify stress with 80% accuracy through Random forest (RF) classifier. SCL, SCR and rise time extracted from EDA resulted in 77% accuracy using RF. Stress labels were obtained by comparing the results to a baseline for both experiments (38). These features used to predict stress are not consistent with the theories associated with ANS activity, stress was classified by comparing the results to a baseline signal and HR which always varies was a prominent predictor of stress in this scenario. Classification report which includes TN, TP, FN, FP accuracy behind stress detection would better indicate the reason behind the contradictory results, which varies from standard theories associated with ANS activity (such as: a decrease in contrast to an increase in LF, LF/HF ratio during times of stress). RF is a bagging algorithm which also implements an ensemble of decision trees much like Adaboost. In contrast to most strong learners which are prone to overfitting and memorizing the data, bagging algorithms reduce variance in a data which improves accuracy and reduces overfitting. Most models perform more effectively if features with linear pattern are utilized, RF is a curve based algorithm which can efficiently adapt to non-linear parameters. It also requires a longer training period and a lot of computational power to handle the excessive number of decision trees used (A standard classification process for a RF algorithm is shown in **Figure 8**).

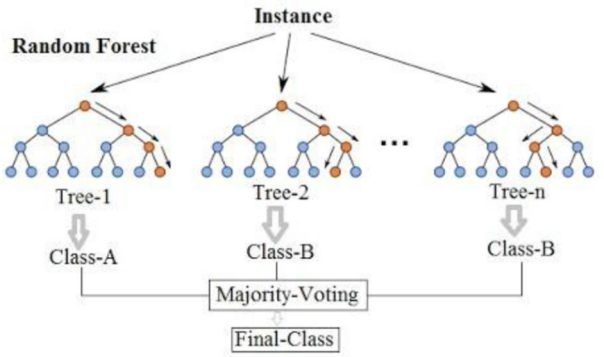

FIGURE 8 | A standard classification process for a RF algorithm. Source: Montantes (78).

6.3. Classifying HRV From ECG and EDA Features Through ML Classification Algorithms

Posada and Bolkhovsky (79) conducted a study to assess Psychomotor vigilance (PVT-measures reaction time), auditory working memory (n-back task), visual search (ship task) through ECG, EDA features, and ML classification algorithms. Lack of sleep due to stress reduces vigilance and the ability of working memory regresses with prolonged lack of sleep. The detection of the activities indicated that PVT, auditory working memory and ship search all had different effects on ANS. SCL, TVsym, and LFn which are SNS biomarkers were the most significant differences associated with EDA and ECG activity during each task. Data was classified using linear kNN, linear SVM, LDA with 66, 66, and 62% classification accuracies, PVT along with ship search was classified with 69% classification accuracy using kNN, while working memory was classified with 69% accuracy using LSVM. The study was conducted upto 24 h, classification after 20 h indicates that ANS activity diminished after 20 h of wakefulness, but surprisingly recovered after 24 h (79). In order to improve the low classification accuracy, feature selection would be an appropriate method to reduce the number of features which are futile. Dimensional reduction methods such as PCA can also be used to classify the data with the most valuable features, which can also reduce model complexity, improve classification accuracy and reduce overfitting (80). Training data is almost of no importance for KNN algorithms, it is an instance based algorithm which cannot derive any discriminative function from the training data, large number of features makes it difficult for the algorithm to derive the distance between each dimension, which also results in a low accuracy. Noisy data-set also hinders performance, outliers and missing data have to be optimized to improve performance. Noisy data also negatively impacts SVM, making feature engineering an essential component to improve performance (81). Noise can produce flawed data which is random and is not normally distributed, if the data set is non-gaussian, it negatively impacts LDA algorithms ability to preserve the complex structure data needed for an efficient classification. Data wrangling is often utilized prior to training/testing a dataset, to minimize outliers, missing data and transform the data-set in order to make it more appropriate, which would make it more efficient and effective for classification using unsupervised models (82). There is a recurring trend between low classification accuracy and irrelevant features, although more data may improve classification accuracy, the appropriate feature selection method is capable of significantly improving the efficiency of the results (83). Ideally more features result in better accuracy, but Taye et al. (72) demonstrated that innovating features based on the specific domain is a much more efficient approach. They were able to reduce 7 dimensions and improve classification accuracy by 26.6% using a novel QRS complex feature engineering method. This is another example of reducing the computational costs while improving the efficiency of the methods. Additional research which combines such methods with wearable devices will allow researchers to dive deeper and further reduce the gap which prevents remote monitoring and diagnosis of HRV from being accessible to everyone in today's healthcare. COVID-19 has really addressed an urgent need for remote health solutions, researchers can revolutionize healthcare by combining ML with HRV in order to reduce stress and cardiac pathologies.

6.4. HRV Associated With Affective Computing, Classified Through NN and SVM

Mobile devices which can monitor health accurately can positively impact a large population of people. This research is targeting more than just CVD and stress, it is expanding to cancer detection, muscle injuries, circadian rhythm and affective emotion (emotion, stress due to age and gender). Rukavina et al. (84) analyzed physiological signals obtained through EMG, EDA, ECG and respiration to distinguish between various affective states based on gender and age. NN and SVM reported the highest classification accuracy using features Mean, Std, fEMG, low valence low arousal (LVLA), low valence high arousal (LVHA), high valence low arousal (HVLA), high valence high arousal (HVHA), and neutral. Mean and std were analyzed to detect skin conductance associated with SNS activity. Valence and arousal state were scrutinized by studying the correlation between neural states and emotions. Performance was evaluated using the leave one out cross validation (LOOCV) method. The classification accuracy was blunted by a small dataset, which can be improved through more trials and additional features (84, 85).

Pathoumvanh et al. (86) revealed that ECG biometrics are different from affective states, they were able to classify HRV conditions with 97% classification accuracy and also achieved 80% robustness study accuracy, using only a single beat ECG feature and LDA algorithm. LDA is a simple model that predicts labels based on the highest probability obtained through Bayes theorem. Fisher's linear discriminant analysis is an extension of LDA which can reduce RMS dimensions and classify data with higher precision. Unlike DT, it's not prone to overfitting (87).

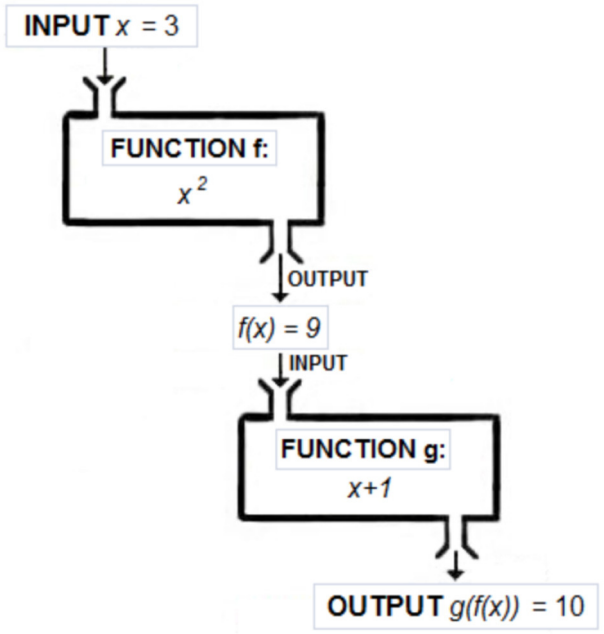

FIGURE 9 | Process of transforming data in a deep learning ELM algorithm. Source: Terry-Jack (88).

6.5. Stress Induced Through VR Environment and Classified Using Extreme Learning Machine (ELM)

Cho et al. (26) were able to classify stress with 95% accuracy using features obtained from three physiological signals (PPG, ECG, EDA) through Kernel based Extreme Learning Machine (K-ELM). K-ELM is based on a single hidden layer feedforward neural network which generates input weights and hidden layer biases, it requires less resources to classify results with high accuracy and leave one out cross validation (LOOCV) was used to evaluate the classifier. KELM is capable of discriminating between classes with high efficiency due to its ability to transform data which is hard to distinguish into linearly separable data while utilizing specific features (as shown in **Figure 9**). However, the features are selected randomly without utilizing an established algorithm like CNN, which makes the results unreliable and random for a specific dataset. The algorithm might not effectively classify other data-set as efficiently. LOOCV takes advantage of one feature to evaluate model performance, it has a high variability despite classifying labels with high accuracy. LOOCV also requires a lot of time to fit and evaluate the data. The experiment unfolds the possibilities which exist for wireless monitoring of stress, accurate results produced from HRV through a wireless device is an indication of phenomenal solution that is yet to be produced in health care due to the lack of efficiency, this is an indication of many possibilities that may arise within the next decade for wireless monitoring of HRV and human health through the use of machine learning and wearable devices.

6.6. Convolutional Neural Network (CNN) Used to Detect Stress Through HRV

Whether it involves stress, CVD or drowsiness detection, one of the limiting factor that exists within most innovations is their inability to perform during real-time applications. He et al. (24) was able to classify cognitive stress using features which were observed in real-time through ultra short 10 s windows. They utilized Lomb scargle periodogram to obtain the PSD from the detected R-peaks. CNN was used to understand the 0.04–20 Hz band from the PSD and extract the relevant features from the input layer. CNN utilizes automatic feature learning for fast and accurate analysis of cognitive stress through HRV features. CNN is similar to other deep learning methods, but it also consists of a convolutional layer in its hidden layer (process flow chart shown in **Figure 10**). It can automatically capture the relevant information from the input unlike other feedforward neural networks, it can reduce the image features to the point where the information becomes very simple to process without losing valuable features required to make an accurate prediction. A typical architecture for HRV classification using a CNN algorithm is shown in **Figure 11**. In order to classify stress using data from the PSD, 10 layers were utilized which included an input layer of size 799 × 1 × 1, a convolutional layer that consisted of 6 filters with size 4 × 1 × 1, batch layer, RELU layer, dropout layer, 3 fully connected layers with batch normalization between them, softmax layer and an output layer. Batch normalization layer normalizes the data, reduces overfitting, and allows each layer to learn independently. RELU layer is essential for effectively updating the data with each iteration. Dropout layer is used to reduce overfitting. Fully connected layers connect the information obtained after being filtered with the output later, in order to classify the data. Softmax layer allows for multiclass classification of the data. CNN produced a 17.3% error rate, which was 7.2 and 32.6% lower than SVM, using comB (combined) feature and LF/HF ratio, respectively. CNN performed better than conventional methods in terms of ER and FAR (false acceptance rate) (24). CNN is really an extension of deep learning models which only use fully connected hidden layers, it's more effective due to its ability to reduce errors through the convolutional layer. Unlike most deep learning models, the convolutional layer allows the model to adapt to the input data more effectively, the activation depth significantly improves due the number of filters, resulting in better classification (43). One of the biggest advantages of CNN is its ability to predict labels with high accuracy using less features than standard deep learning models. Overfitting is the downside to all deep learning models, batch size and epochs allow the model to update the weight and minimize error, but such a method is also prone to overfitting especially if its a smaller dataset. The development of CNN has made remote monitoring of HRV much more effective and simpler. CNN is a powerful algorithm which can be used to extract valuable features from raw ECG signals obtained through a wireless ECG sensor, and classify HRV and stress with a high accuracy of 90.19% (70). The results are biased, most CNN algorithms are very prone to overfitting and memorizing the data, especially

FIGURE 10 | A typical CNN architecture for stress classification using HRV parameters.

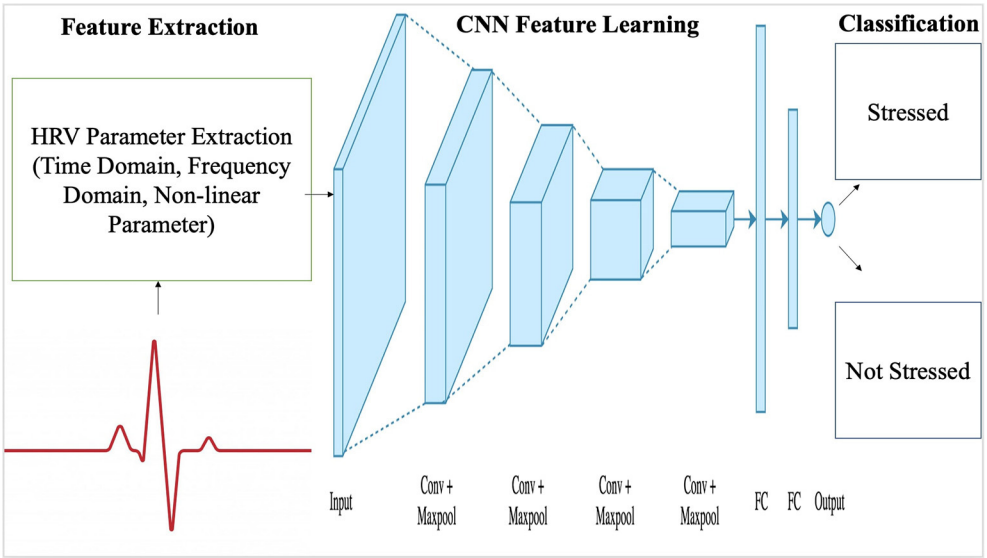

FIGURE 11 | An example of a CNN algorithm process flow chart. Source: Aishwarya (89).

if the data-set is very small. Although it can be combined with wireless sensors to monitor heart rate and classify HRV from a distance, further research should be conducted with 50 people and larger datasets, in order to better verify the significance of developed algorithms for remote monitoring of HRV. The positive outcomes does hint that if researchers continue to improve existing CNN algorithms and the efficacy of analyzing data obtained through wireless sensors, remote monitoring of HRV can make a huge impact on the lives of others who are stressed due to work, suffering from cardiovascular diseases or are incapable of going to a clinician for routine checkups (90). The computational time complexity of convolutional layers is $0(n) = 0(\sum_{l=1}^{d} n_{l-1} \cdot s_l^2 \cdot n_l \cdot m_l^2)$, where l represents the index of the convolution layer, d represents the depth, n_l represents the number of filters in the l-th layer, n_{l-1} describes the number of input channels, s_l indicates the spatial size of the filter and m_l represents the spatial size of the output feature map (91). A typical 1D convolutional layer has a computational complexity of $0(n \cdot k \cdot d)$, further demonstrating the high computational resources and time required for a basic CNN architecture (92). Outside of HRV, there are numerous research conducted to reduce the computational cost of CNN, which typically compromises the output and classification accuracy (93). Inouchi et al. (93) developed a functionally-predefined kernel which significantly reduced the number of training parameters without compromising the accuracy. Further contribution toward similar methods catered toward HRV research can create a significant change within the healthcare system, such as reducing the number tedious hours needed from healthcare professionals and improving patient outcomes while decreasing healthcare costs.

7. CONCLUSION

This article which presented various summaries and reviews of the different applications associated with HRV research emphasized that reduced HRV is associated with increased morbidity and stress. Lower HRV is associated with increased SNS activity, which increases HR and blood pressure, presenting an immediate indication of the threat perceived by the ANS, which reacts to maintain normal function of the body and keep the body in a state of homeostasis. HRV in motion is less efficient in comparison to many other research studies such as stress and myocardial infarction. Numerous studies have indicated the lack of accuracy associated with exercise and drowsiness detection, this aspect of HRV research requires more attention and should be improved, in order to prevent injuries which may occur from performance fatigue near a sports competition or accidents associated with drowsy driving. HRV research will continue to expand due to its relevance in science, health and wellness of the heart. ML algorithms, AI (artificial intelligence) and frequency domain analysis of HRV can cause a huge impact in people's lives in a short period, if it is accurate, thus researchers go with the flow

and improve these processing methods to improve lives/health of patients, prevent possible road accidents and enhance the quality of life.

7.1. Future Direction

HRV is a prominent topic concerning the activity of the heart and the ANS, although research has been steadily increasing, data analysis of HRV in motion is far from where it should be especially concerning drowsiness. Vicente et al. (23) and Georgiou et al. (21) have explained that HRV is hard to detect in motion, whether it involves exercise or drowsy driving, accuracy of HRV detection declines due to motion. Detection method in motion is a concern and should be a priority for improvement with regards to future research involving HRV. Machine learning algorithms, frequency domain analysis have been effective for stress analysis and remote monitoring of cardiovascular diseases through HRV analysis. Expansion in these domains of data analysis could provide effective/efficient results that produce an accurate representation of a person's HRV, which is easy to compute and can analyse a lot of data at once, making the detection process a lot smoother and quicker. Machine learning can be utilized to improve prognosis, since it can better assess medical records through logical algorithms in comparison currents scoring tools, which utilize a generalized thought process. CNN is a great algorithm that can effectively predict pathologies from X-ray images, at a faster rate than radiologists. Recent development also suggests that machine learning algorithms can create an immense impact toward public health, antedating infectious diseases and increasing the chances of preventing a chronic outcome.

AUTHOR CONTRIBUTIONS

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

FUNDING

Financial support from NSERC and Shaftesbury Inc. (CRDPJ537987-18) to conduct the research is highly appreciated. Shaftesbury Inc was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the funding provided by Natural Sciences and Engineering Research Council of Canada, and Ryerson University.

REFERENCES

1. Forte G, Casagrande M. Heart rate variability and cognitive function: a systematic review. *Front Neurosci*. (2019) 13:710. doi: 10.3389/fnins.2019.00710
2. Li K, Rüdiger H, Ziemssen T. Spectral analysis of heart rate variability: time window matters. *Front. Neurol*. (2019) 10:545. doi: 10.3389/fneur.2019.00545
3. Shaffer F, Ginsberg J. An overview of heart rate variability metrics and norms. *Front Public Health*. (2017) 5:258. doi: 10.3389/fpubh.2017.00258
4. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig*. (2018) 15:235. doi: 10.30773/pi.2017.08.17
5. Teschler L. *Monitoring Heart Rate Variability for Better Athletic Ability*. (2020). Available online at: https://www.testandmeasurementtips.com/monitoring-heart-rate-variability-for-better-athletic-ability-faq/
6. Rosenberg WV, Chanwimalueang T, Adjei T, Jaffer U, Goverdovsky V, Mandic DP. Resolving ambiguities in the LF/HF ratio: LF-HF scatter plots for the categorization of mental and physical stress from HRV. *Front Physiol*. (2017) 8:360. doi: 10.3389/fphys.2017.00360
7. Blood JD, Wu J, Chaplin TM, Hommer R, Vazquez L, Rutherford HJ, et al. The variable heart: High frequency and very low frequency correlates of depressive symptoms in children and adolescents. *J Affect Disord*. (2015) 186:119–26. doi: 10.1016/j.jad.2015.06.057
8. Molina GE, Fontana KE, Porto LGG, Junqueira LF. Post-exercise heart-rate recovery correlates to resting heart-rate variability in healthy men. *Clin Auton Res*. (2016) 26:415–21. doi: 10.1007/s10286-016-0378-2
9. Leti T, Bricout VA. Interest of analyses of heart rate variability in the prevention of fatigue states in senior runners. *Auton Neurosci*. (2013) 173:14–21. doi: 10.1016/j.autneu.2012.10.007
10. Walker ED, Brammer A, Cherniack MG, Laden F, Cavallari JM. Cardiovascular and stress responses to short-term noise exposures–A panel study in healthy males. *Environ Res*. (2016) 150:391–7. doi: 10.1016/j.envres.2016.06.016
11. Wang Y, Wei S, Zhang S, Zhang Y, Zhao L, Liu C, et al. Comparison of time-domain, frequency-domain and non-linear analysis for distinguishing congestive heart failure patients from normal sinus rhythm subjects. *Biomed Signal Process Control*. (2018) 42:30–36. doi: 10.1016/j.bspc.2018.01.001
12. Huang JC, Chen CF, Chang CC, Chen SC, Hsieh MC, Hsieh YP, et al. Effects of stroke on changes in heart rate variability during hemodialysis. *BMC Nephrol*. (2017) 18:90. doi: 10.1186/s12882-017-0502-0
13. Pinheiro AdO, Pereira VL Jr., Baltatu OC, Campos LA. Cardiac autonomic dysfunction in elderly women with myocardial infarction. *Curr Med Res Opin*. (2015) 31:1849–54. doi: 10.1185/03007995.2015.1074065
14. Toni G, Murri MB, Piepoli M, Zanetidou S, Cabassi A, Squatrito S, et al. Physical exercise for late-life depression: effects on heart rate variability. *Am J Geriatr Psychiatry*. (2016) 24:989–97. doi: 10.1016/j.jagp.2016.08.005
15. Shi H, Yang L, Zhao L, Su Z, Mao X, Zhang L, et al. Differences of heart rate variability between happiness and sadness emotion states: a pilot study. *J Med Biol Eng*. (2017) 37:527–39. doi: 10.1007/s40846-017-0238-0
16. Ponnusamy A, Marques JL, Reuber M. Comparison of heart rate variability parameters during complex partial seizures and psychogenic nonepileptic seizures. *Epilepsia*. (2012) 53:1314–21. doi: 10.1111/j.1528-1167.2012.03518.x
17. Howells FM, Rauch HL, Ives-Deliperi VL, Horn NR, Stein DJ. Mindfulness based cognitive therapy may improve emotional processing in bipolar disorder: pilot ERP and HRV study. *Metab Brain Dis*. (2014) 29:367–75. doi: 10.1007/s11011-013-9462-7
18. Rios-Aguilar S, Merino JLM, Sánchez AM, Valdivieso AS. Variation of the heartbeat and activity as an indicator of drowsiness at the wheel using a smartwatch. *Int J Interact Multimedia Artif Intell*. (2015) 3:96–100. doi: 10.9781/ijimai.2015.3313
19. Jung SJ, Shin HS, Chung WY. Driver fatigue and drowsiness monitoring system with embedded electrocardiogram sensor on steering wheel. *Intell Transport Syst*. (2014) 8:43–50. doi: 10.1049/iet-its.2012.0032
20. Rahim HA, Dalimi A, Jaafar H. Detecting drowsy driver using pulse sensor. *Jurnal Teknologi*. (2015) 73:400–8. doi: 10.11113/jt.v73.4238
21. Georgiou K, Larentzakis AV, Khamis NN, Alsuhaibani GI, Alaska YA, Giallafos EJ. Can wearable devices accurately measure heart rate variability?

A systematic review. *Folia Med*. (2018) 60:7–20. doi: 10.2478/folmed-2018-0012
22. Gontier C. How to prevent mind-wandering during an EVA? Presentation of a mind-wandering detection method using ECG technology in a Mars-analog environment. *Acta Astronaut*. (2017) 140:105–12. doi: 10.1016/j.actaastro.2017.08.008
23. Vicente J, Laguna P, Bartra A, Bailón R. Drowsiness detection using heart rate variability. *Med Biol Eng Comput*. (2016) 54:927–37. doi: 10.1007/s11517-015-1448-7
24. He J, Li K, Liao X, Zhang P, Jiang N. Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal. *IEEE Access*. (2019) 7:42710–7. doi: 10.1109/ACCESS.2019.2907076
25. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. Boulder, CO (2018) p. 400–8. doi: 10.1145/3242969.3242985
26. Cho D, Ham J, Oh J, Park J, Kim S, Lee NK, et al. Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. *Sensors*. (2017) 17:2435. doi: 10.3390/s17102435
27. Knox J. The heart rate during a simple exercise. *Brit Heart J*. (1940) 2:289. doi: 10.1136/hrt.2.4.289
28. Simonson E. The normal variability of the electrocardiogram as a basis for differentiation between "normal" and "abnormal" in clinical electrocardiography. *Am Heart J*. (1958) 55:80–103. doi: 10.1016/0002-8703(58)90258-8
29. Lynch JJ. Heart rate variability of dogs in classical conditioning. *Psychol Rec*. (1968) 18:101–6. doi: 10.1007/BF03393749
30. Rompelman O, Coenen A, Kitney R. Measurement of heart-rate variability: Part 1–Comparative study of heart-rate variability analysis methods. *Med Biol Eng Comput*. (1977) 15:233. doi: 10.1007/BF02441043
31. Merri M, Farden DC, Mottley JG, Titlebaum EL. Sampling frequency of the electrocardiogram for spectral analysis of the heart rate variability. *IEEE Trans Biomed Eng*. (1990) 37:99–106. doi: 10.1109/10.43621
32. Acharya UR, Joseph KP, Kannathal N, Lim CM, Suri JS. Heart rate variability: a review. *Med Biol Eng Comput*. (2006) 44:1031–51. doi: 10.1007/s11517-006-0119-0
33. Price AD. Heart rate variability and respiratory concomitants of visual and nonvisual "imagery" and cognitive style. *J Res Personality*. (1975) 9:341–55. doi: 10.1016/0092-6566(75)90008-2
34. Cowan MJ, Kogan H, Burr R, Hendershot S, Buchanan L. Power spectral analysis of heart rate variability after biofeedback training. *J Electrocardiol*. (1990) 23:85–94. doi: 10.1016/0022-0736(90)90081-C
35. Ripoli A, Emdin M. Complexity of heart rate, blood pressure and respiration disclosed by pattern fractal analysis. In: *Computers in Cardiology*. Cambridge, MA: IEEE (2000). p. 135–38.
36. Orini M, Bailón R, Enk R, Koelsch S, Mainardi L, Laguna P. A method for continuously assessing the autonomic response to music-induced emotions through HRV analysis. *Med Biol Eng Comput*. (2010) 48:423–33. doi: 10.1007/s11517-010-0592-3
37. Alhitary AE, Hay EWA, Al-bashir AK, et al. Objective detection of chronic stress using physiological parameters. *Med Biol Eng Comput*. (2018) 56:2273–86. doi: 10.1007/s11517-018-1854-8
38. Lima R, de Noronha Osório DF, Gamboa H. Heart rate variability and electrodermal activity in mental stress aloud: predicting the outcome. In: *Biosignals*. Prague: Springer (2019). p. 42–51. doi: 10.5220/0007355200420051
39. Jovic A, Bogunovic N. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. *Artif Intell Med*. (2011) 51:175–86. doi: 10.1016/j.artmed.2010.09.005
40. Balasubramanian K, Kumar RN. Improvising heart attack prediction system using feature selection and data mining methods. *Int J Adv Res Comput Sci*. (2010) 1:356–73 doi: 10.1504/IJDATS.2019.103756
41. Castaldo R, Montesinos L, Melillo P, James C, Pecchia L. Ultra-short term HRV features as surrogates of short term HRV: a case study on mental stress detection in real life. *BMC Med Inform Decis Mak*. (2019) 19:1–13. doi: 10.1186/s12911-019-0742-y
42. Martinez R, Irigoyen E, Arruti A, Martin JI, Muguerza J. A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine. *Comput Methods Prog Biomed*. (2017) 148:81–90. doi: 10.1016/j.cmpb.2017.06.010
43. Hwang B, You J, Vaessen T, Myin-Germeys I, Park C, Zhang BT. Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *Telemed e-Health*. (2018) 24:753–72. doi: 10.1089/tmj.2017.0250
44. Oskooei A, Chau SM, Weiss J, Sridhar A, Martinez MR, Michel B. DeStress: deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (HRV) data. *arXiv preprint arXiv:191113213*. (2019). doi: 10.1007/978-3-030-53352-6_9
45. Muaremi A, Arnrich B, Tröster G. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience 3*. 2:172–83 (2013). doi: 10.1007/s12668-013-0089-2
46. Kanjo E, Al-Husain L, Chamberlain A. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Pers Ubiquit Comput*. (2015) 19:1197–212. doi: 10.1007/s00779-015-0842-3
47. Venkatesan C, Karthigaikumar P, Paul A, Satheeskumaran S, Kumar R. ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications. *IEEE Access*. (2018) 6:9767–73. doi: 10.1109/ACCESS.2018.2794346
48. Stephens J, Moscou-Jackson G, Allen JK. Young adults, technology, and weight loss: a focus group study. *J Obes*. (2015) 2015:1–6. doi: 10.1155/2015/379769
49. Lee JW, Lee SK, Kim CH, Kim KH, Kwon OC. Detection of drowsy driving based on driving information. In: 2014 *International Conference on Information and Communication Technology Convergence (ICTC)*. Busan: IEEE (2014). p. 607–8. doi: 10.1109/ICTC.2014.6983224
50. Sangeetha M, Kalpanadevi S, Rajendiran M, Malathi G. Embedded ECG based real time monitoring and control of driver drowsiness condition. *Int J Sci Technol Soc*. (2015) 3:176. doi: 10.11648/j.ijsts.20150304.17
51. Roy R, Venkatasubramanian K. EKG/ECG based driver alert system for long haul drive. *Indian J Sci Technol*. (2015) 8:8–13. doi: 10.17485/ijst/2015/v8i19/77014
52. Fernandes T, Chec A, Olczak D, Ferreira H. Physiological computing gaming: Use of electrocardiogram as an input for video gaming. In: *Proceedings of the International Conference on Physiological Computing Systems (PhyCS)*. Angers: SCITEPRESS (2015). p. 11–3.
53. Lastre-Dominguez C, Shmaliy YS, Ibarra-Manzano O, Munoz-Minjares J, Morales-Mendoza LJ. ECG signal denoising and features extraction using unbiased FIR smoothing. *BioMed Res Int*. London (2019) 2019. doi: 10.1155/2019/2608547
54. Adochiei F, Edu I, Adochiei N. Comparative filtering methods for noisy ECG signals. In: *2011 E-Health and Bioengineering Conference (EHB)*. IEEE (2011). p. 1–4.
55. Sathyapriya L, Murali L, Manigandan T. Analysis and detection R-peak detection using Modified Pan-Tompkins algorithm. In: *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*. IEEE (2014). p. 483–7. doi: 10.1109/ICACCCT.2014.7019490
56. Wiklendt L, Brookes SJ, Costa M, Travis L, Spencer NJ, Dinning PG. A novel method for electrophysiological analysis of EMG signals using MesaClip. *Front Physiol*. (2020) 11:484. doi: 10.3389/fphys.2020.00484
57. Kumari K, Sahu SS, Sinha RK. R Peak Detection using empirical mode decomposition with Shannon energy envelope. In: *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE (2018). p. 550–4. doi: 10.1109/ICICCT.2018.8473279
58. Malik M, Camm AJ. Heart rate variability and clinical cardiology. *Brit Heart J*. (1994) 71:3. doi: 10.1136/hrt.71.1.3
59. Laborde S, Mosley E, Thayer JF. Heart rate variability and cardiac vagal tone in psychophysiological research-recommendations for experiment planning, data analysis, and data reporting. *Front Psychol*. (2017) 8:213. doi: 10.3389/fpsyg.2017.00213
60. O'Connor P, Neil D, Liu SC, Delbruck T, Pfeiffer M. Real-time classification and sensor fusion with a spiking deep belief network. *Front Neurosci*. (2013) 7:178. doi: 10.3389/fnins.2013.00178
61. Bourdillon N, Schmitt L, Yazdani S, Vesin JM, Millet GP. Minimal window duration for accurate HRV recording in athletes. *Front Neurosci*. (2017) 11:456. doi: 10.3389/fnins.2017.00456

62. Hartmann R, Schmidt FM, Sander C, Hegerl U. Heart rate variability as indicator of clinical state in depression. *Front Psychiatry*. (2019) 9:735. doi: 10.3389/fpsyt.2018.00735
63. Shiro Y, Arai YCP, Matsubara T, Isogai S, Ushida T. Effect of muscle load tasks with maximal isometric contractions on oxygenation of the trapezius muscle and sympathetic nervous activity in females with chronic neck and shoulder pain. *BMC Musculoskelet Disord*. (2012) 13:146. doi: 10.1186/1471-2474-13-146
64. Abad C, Kobal R, Kitamura K, Gil S, Pereira L, Loturco I, et al. Heart rate variability in elite sprinters: effects of gender and body position. *Clin Physiol Funct Imaging*. (2017) 37:442–7. doi: 10.1111/cpf.12331
65. Schneider C, Wiewelhove T, Raeder C, Flatt AA, Hoos O, Hottenrott L, et al. Heart rate variability monitoring during strength and high-intensity interval training overload microcycles. *Front Physiol*. (2019) 10:582. doi: 10.3389/fphys.2019.00582
66. Choi A, Shin H. Quantitative analysis of the effect of an ectopic beat on the heart rate variability in the resting condition. *Front Physiol*. (2018) 9:922. doi: 10.3389/fphys.2018.00922
67. Princip M, Scholz M, Meister-Langraf RE, Barth J, Schnyder U, Znoj H, et al. Can illness perceptions predict lower heart rate variability following acute myocardial infarction? *Front Psychol*. (2016) 7:1801. doi: 10.3389/fpsyg.2016.01801
68. Do Amaral JA, Guida HL, De Abreu LC, Barnabé V, Vanderlei FM, Valenti VE. Effects of auditory stimulation with music of different intensities on heart period. *J Trad Complement Med*. (2016) 6:23–8. doi: 10.1016/j.jtcme.2014.11.032
69. Vickhoff B, Malmgren H, Åström R, Nyberg G, Engvall M, Snygg J, et al. Music structure determines heart rate variability of singers. *Front Psychol*. (2013) 4:334. doi: 10.3389/fpsyg.2013.00334
70. Cho HM, Park H, Dong SY, Youn I. Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network. *Sensors*. (2019) 19:4408. doi: 10.3390/s19204408
71. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. *J Biomed Inform*. (2020) 112:103610. doi: 10.1016/j.jbi.2020.103610
72. Taye GT, Shim EB, Hwang HJ, Lim KM. Machine learning approach to predict ventricular fibrillation based on QRS complex shape. *Front Physiol*. (2019) 10:1193. doi: 10.3389/fphys.2019.01193
73. Arsalan A, Majid M, Butt AR, Anwar SM. Classification of perceived mental stress using a commercially available EEG headband. *IEEE J Biomed Health Inform*. (2019) 23:2257–64. doi: 10.1109/JBHI.2019.2926407
74. Kublanov VS, Dolganov AY, Belo D, Gamboa H. Comparison of machine learning methods for the arterial hypertension diagnostics. *Appl Bionics Biomech*. (2017) 2017:1–13. doi: 10.1155/2017/5985479
75. Ma F, Zhang J, Liang W, Xue J. Automated classification of atrial fibrillation using artificial neural network for wearable devices. *Math Probl Eng*. (2020) 2020:1–6. doi: 10.1155/2020/9159158
76. Persson A, Jonasson H, Fredriksson I, Wiklund U, Ahlström C. Heart rate variability for classification of alert versus sleep deprived drivers in real road driving conditions. *IEEE Trans Intell Transport Syst*. (2020) 21:1–10. doi: 10.1109/TITS.2020.2981941
77. Hu J. Automated detection of driver fatigue based on AdaBoost classifier with EEG signals. *Front Comput Neurosci*. (2017) 11:72. doi: 10.3389/fncom.2017.00072
78. Montantes J. *3 Reasons to Use Random Forest Over a Neural Network-Comparing Machine Learning versus Deep. Towards Data Science*. (2020). Available online at: https://towardsdatascience.com/3-reasons-to-use-random-forest-over-a-neural-network-comparing-machine-learning-versus-deep-f9d65a154d89
79. Posada HF, Bolkhovsky JB. Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity. *Behav Sci*. (2019) 9:45. doi: 10.3390/bs9040045
80. Padmanaban S, Baker J, Greger B. Feature selection methods for robust decoding of finger movements in a non-human primate. *Front Neurosci*. (2018) 12:22. doi: 10.3389/fnins.2018.00022
81. Hong KS, Khan MJ, Hong MJ. Feature extraction and classification methods for hybrid fNIRS-EEG brain-computer interfaces. *Front Hum Neurosci*. (2018) 12:246. doi: 10.3389/fnhum.2018.00246
82. Hohman F, Kahng M, Pienta R, Chau DH. Visual analytics in deep learning: an interrogative survey for the next frontiers. *IEEE Trans Visual Comput Graph*. (2018) 25:2674–93. doi: 10.1109/TVCG.2018.2843369
83. Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, Initiative ADN, et al. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*. (2012) 60:59–70. doi: 10.1016/j.neuroimage.2011.11.066
84. Rukavina S, Gruss S, Hoffmann H, Tan JW, Walter S, Traue HC. Affective computing and the impact of gender and age. *PLoS ONE*. (2016) 11:e0150584. doi: 10.1371/journal.pone.0150584
85. Ishaque S, Rueda A, Nguyen B, Khan N, Krishnan S. Physiological signal analysis and classification of stress from virtual reality video game. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Montreal, QC: IEEE (2020). p. 867–70. doi: 10.1109/EMBC44109.2020.9176110
86. Pathoumvanh S, Airphaiboon S, Hamamoto K. Robustness study of ECG biometric identification in heart rate variability conditions. *IEEE Trans Electric Electron Eng*. (2014) 9:294–301. doi: 10.1002/tee.21970
87. Neto E, Biessmann F, Aurlien H, Nordby H, Eichele T. Regularized linear discriminant analysis of EEG features in dementia patients. *Front Aging Neurosci*. (2016) 8:273. doi: 10.3389/fnagi.2016.00273
88. Terry-Jack M. *Deep Learning, NeuroEvolution & Extreme Learning Machines. Medium*. (2019) Available online at: https://medium.com/@b.terryjack/deep-learning-neuroevolution-extreme-learning-machines-6b448860a72a
89. Aishwarya S. *Introduction to Neural Network: Convolutional Neural Network. Analytics Vidhya*. (2020). Available online at: https://www.analyticsvidhya.com/blog/2020/02/mathematics-behind-convolutional-neural-network/
90. Liang Y, Yin S, Tang Q, Zheng Z, Elgendi M, Chen Z. Deep learning algorithm classifies heartbeat events based on electrocardiogram signals. *Front Physiol*. (2020) 11:569050. doi: 10.3389/fphys.2020.569050
91. He K, Sun J. Convolutional neural networks at constrained time cost. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA (2015). p. 5353-60. doi: 10.1109/CVPR.2015.7299173
92. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *arXiv preprint arXiv:170603762*. (2017).
93. Inouchi Y, Yamaki H, Shinobu M, Tsumura T. Functionally-predefined kernel: a way to reduce CNN computation. In: *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. Victoria, BC: IEEE (2019). p. 1–6. doi: 10.1109/PACRIM47961.2019.8985122

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 Digital Health

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fdgth.2021.637386

An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis

Juan Song[1†], *Jian Zheng*[1†], *Ping Li*[2], *Xiaoyuan Lu*[2], *Guangming Zhu*[1*] and *Peiyi Shen*[1]

[1] School of Computer Science and Technology, Xidian University, Shaanxi, China, [2] Data and Virtual Research Room, Shanghai Broadband Network Center, Shanghai, China

Alzheimer's disease (AD) is an irreversible brain disease that severely damages human thinking and memory. Early diagnosis plays an important part in the prevention and treatment of AD. Neuroimaging-based computer-aided diagnosis (CAD) has shown that deep learning methods using multimodal images are beneficial to guide AD detection. In recent years, many methods based on multimodal feature learning have been proposed to extract and fuse latent representation information from different neuroimaging modalities including magnetic resonance imaging (MRI) and 18-fluorodeoxyglucose positron emission tomography (FDG-PET). However, these methods lack the interpretability required to clearly explain the specific meaning of the extracted information. To make the multimodal fusion process more persuasive, we propose an image fusion method to aid AD diagnosis. Specifically, we fuse the gray matter (GM) tissue area of brain MRI and FDG-PET images by registration and mask coding to obtain a new fused modality called "GM-PET." The resulting single composite image emphasizes the GM area that is critical for AD diagnosis, while retaining both the contour and metabolic characteristics of the subject's brain tissue. In addition, we use the three-dimensional simple convolutional neural network (3D Simple CNN) and 3D Multi-Scale CNN to evaluate the effectiveness of our image fusion method in binary classification and multi-classification tasks. Experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset indicate that the proposed image fusion method achieves better overall performance than unimodal and feature fusion methods, and that it outperforms state-of-the-art methods for AD diagnosis.

Keywords: Alzheimer's disease, multimodal image fusion, MRI, FDG-PET, convolutional neural networks, multi-class classification

OPEN ACCESS

Edited by:
Kezhi Li,
University College London,
United Kingdom

Reviewed by:
Zhibo Wang,
University of Central Florida,
United States
Jun Shi,
Shanghai University, China

***Correspondence:**
Guangming Zhu
gmzhu@xidian.edu.cn

[†]These authors have contributed equally to this work and share the first authorship

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

Received: *10 December 2020*
Accepted: *05 February 2021*
Published: *26 February 2021*

Citation:
Song J, Zheng J, Li P, Lu X, Zhu G and Shen P (2021) An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. Front. Digit. Health 3:637386. doi: 10.3389/fdgth.2021.637386

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive brain disorder and the most common cause of dementia in later life. It causes cognitive deterioration, eventually resulting in inability to carry out activities of daily life. AD not only severely degrades patients' quality of life but also causes additional distress for caregivers (1). At least 50 million people worldwide are likely to suffer from AD or other dementias. Total payments in 2020 for health care, long-term care, and hospice services for people aged 65 and older with dementia are estimated to be $305 billion (2). And the number of AD patients is estimated to be 115 million by 2050. Therefore, accurate early diagnosis and treatment of AD is of great importance.

Currently, the pathogenesis of AD is not fully understood. The academic community generally believes that AD is related to neurofibrillary tangles and extracellular amyloid-β (Aβ) deposition, which cause loss or damage of neurons and synapses (3, 4). In general, the AD diagnostic system classifies a subject into one of three categories: AD, mild cognitive impairment (MCI), and normal control (NC). The main clinical examination methods for AD include neuropsychological examination and neuroimaging examination (5), in which computer-aided diagnosis is of great help in screening at-risk individuals. Psychological auxiliary diagnosis of AD uses the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) to help clinicians determine the severity of dementia. With the rapid development of neuroimaging technology, neuroimaging diagnosis has become an indispensable diagnostic method for AD. In particular, magnetic resonance imaging (MRI) and positron emission tomography (PET) are popular and non-invasive techniques used to capture brain tissue characteristics.

Structural MRI has become a commonly used structural neuroimaging in AD diagnosis because of its high resolution for soft tissue and its ability to present brain anatomical details. Progression of AD results in gross atrophy of the affected regions, including degeneration in the temporal lobe and parietal lobe, as well as parts of the frontal cortex and cingulate gyrus (6). Brain ventricles, which produce cerebrospinal fluid (CSF), become larger in AD patients. And the brain cortex shrivels up, with severe shrinkage occurring particularly in the hippocampus area. MRI, which provides three-dimensional (3D) images of brain tissues, enables clear observation of these structural changes in the patient's brain. Notable results were reported by a number of studies of clinical diagnosis of AD using MRI. Klöppel et al. (7) first segmented the whole brain into gray matter (GM), white matter (WM), and CSF, and used GM voxels as features of MR images to train a support vector machine to discriminate between AD and NC subjects. Owing to the strong relationship of GM with AD diagnosis, compared with WM and CSF (8, 9) only considered spatially normalized GM volumes, called GM tissue densities, for classification. Similarly, Zhu et al. (10) only computed the volume of GM as a feature for each region of the 93 regions of interest in the labeled MR image and used multiple-kernel learning to classify the neuroimaging data. These studies indicate that GM tissue is the most important area for AD classification using MRI (11, 12).

PET imaging has a critical role as a functional technique that enables clinicians to observe activities related to the human brain quickly and precisely, with particular applications in early AD detection (13). As stated in (14), PET images captured via diffusion of radioactive 18-fluorodeoxyglucose (FDG) have been used to obtain sensitive measurements of cerebral metabolic rates of glucose (CMRglc). CMRglc can be used to distinguish AD from other dementias, predict and track decline from NC to AD, and screen at-risk individuals prior to the onset of cognitive symptoms. FDG-PET is particularly useful when changes in physiological and pathological anatomy are difficult to distinguish (15). For instance, the volume of brain structures commonly decreases with age (e.g., in individuals older than 75 years), making it difficult to determine whether a person's brain is in a normal or diseased state only using the brain anatomical changes observed by MRI. In such cases, PET can more effectively detect the disease status of subjects.

Structural MRI can reflect the changes of brain structure, whereas functional PET images can capture the characteristics of brain metabolism to enhance the ability to find lesions (16). Therefore, it has been proposed that multimodal methods combining MRI and PET images could improve the accuracy of AD classification (17–19). Feature fusion strategies are commonly used in multimodal learning tasks, combining high-dimensional semantic features extracted from different unimodal data (20, 21). For example, Shi et al. (22) used two stacked deep polynomial networks (SDPNs) to learn high-level features of MRI and PET images, respectively, which were then fed to another SDPN to fuse the multimodal neuroimaging information. Similarly, Lu et al. (23) used six independent deep neural networks (DNN) to extract corresponding features from different scales of unimodal images (such as those obtained by MRI or PET); the features were then fused by another DNN. Related studies show that a feature fusion strategy can indeed achieve better experimental performance than use of unimodal data alone (24, 25). However, such a method is a "black box," lacking sufficient interpretability to explain the exact reason for better or worse results in a particular case. In addition, deep learning methods based on feature fusion always greatly increase the number of model parameters, as a multi-channel input network is used to extract heterogeneous features from different modalities.

Compared with feature fusion strategies, multimodal medical image fusion is a more intuitive approach that integrates relevant and complementary information from multiple input images into a single fused image in order to facilitate more precise diagnosis and better treatment (26). The fused images have not only richer modal characteristics but also more powerful information representation. Besides, GM is the most important tissue for AD auxiliary diagnosis, which can show the brain's anatomical changes in MRI scans and the overall level of brain metabolism in PET scans. Motivated by these factors, we propose an image fusion method that fuses GM tissue information from MRI and FDG-PET images into a new GM-PET modality. During the fusion process, only the key GM areas are preserved, instead of the full MRI and PET information, to reduce noise and irrelevant information in the fused image and enable the subsequent feature extraction to focus on the crucial characteristics.

The main contributions of this work are two-fold. (1) A novel image fusion method is proposed for AD diagnosis to enhance the information representation ability of neuroimaging modalities by fusing the key GM information from MRI and PET scans into a single composite image. (2) We propose two 3D CNN for AD diagnosis, i.e., 3D Simple CNN and 3D Multi-Scale CNN, to evaluate the performance of different modalities in AD classification tasks. We also prove that the proposed fused modality with its powerful information representation can provide better diagnostic performance and adapt to different CNN.

The rest of this paper is organized as follows. section 2 describes the dataset used and our image fusion method. Our 3D Simple CNN and 3D Multi-Scale CNN are introduced in section

2.3 to extract the features and perform classification based on the neuroimaging data. In section 3, classification experiments for AD vs. NC, MCI vs. NC, AD vs. MCI, and AD vs. MCI vs. NC are conducted to evaluate the effectiveness of our proposed image fusion in an AD diagnostic framework. The discussion and conclusion are presented in sections 4 and 5, respectively.

2. MATERIALS AND METHODS

2.1. Datasets

The data used in the study were acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (https://adni.loni.usc.edu/). ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. ADNI makes all data and samples available for scientists worldwide to promote AD diagnosis and treatment (27, 28). The ADNI researchers have collected and integrated analyses of multimodal data, mainly from North American participants. The dataset contains data from different AD stages. In this study, subjects were selected who had both T1-weighted MRI and FDG-PET scans captured in the same period. MRI scans labeled as MPRAGE were selected as these are considered the best with respect to quality ratings. A total of 381 subjects from the ADNI were selected, comprising 95 AD subjects, 160 MCI subjects, and 126 NC subjects. Clinical information for the selected subjects is shown in **Table 1**.

The MRI and FDG-PET images in ADNI have undergone several processing steps. In detail, the MRI images are processed by the following steps: Gradwarp, B1 non-uniformity, and N3. Gradwarp corrects image geometry distortion caused by the gradient model, and B1 non-uniformity corrects image intensity non-uniformity using B1 calibration scans. Finally, an N3 histogram peak-sharpening algorithm is applied to reduce the non-uniformity of intensity. The need to perform the image pre-processing corrections outlined above varies among manufacturers and system RF coil configurations. We used the fully pre-processed data in our experiments.

In order to obtain more uniform PET data among different systems, the baseline FDG-PET scans are processed by the following steps. (1) Co-Registered dynamic: six 5-min FDG-PET frames are acquired within 30–60 min post-injection, each of which is co-registered to the first extracted frame. The independent frames are co-registered to one another to lessen the effects of patient motion. (2) Averaging: six co-registered frames obtained are averaged. (3) Standardization of image and voxel size: the averaged image is reoriented into a standard 160 × 160 × 96 voxel image grid with 1.5 mm cubic voxels after anterior commissure–posterior commissure correction, followed by intensity normalization using a subject-specific mask so that the average value of voxels within the mask is exactly one. (4) Uniform resolution: the normalized image is filtered with a scanner-specific filter to obtain an image with a uniform isotropic resolution of 8 mm full width at half maximum, in order to smooth the above-mentioned images.

TABLE 1 | Demographic information for subjects. Values are presented as mean ± standard deviation.

Subjects	Number	Male/ Female	Age	MMSE	CDR
NC	126	71/55	75.25 ± 5.82	29.58 ± 0.66	0.02 ± 0.18
MCI	160	108/52	76.97 ± 8.23	26.14 ± 0.81	1.38 ± 2.00
AD	95	54/41	76.52 ± 6.96	18.56 ± 4.20	2.87 ± 3.60

2.2. Proposed Image Fusion

To make the multimodal fusion process more interpretable, we propose fusing MRI and PET scans at the image field. The fused image modality is then fed into a single-channel network for diagnosis of subjects. This approach greatly reduces the number of model parameters compared with the multi-channel input network using feature fusion. Our proposed AD diagnostic framework with multimodal image fusion method is presented in **Figure 1**. It is composed of several parts: image fusion, feature extraction, and classification. First, our image fusion method can obtain a new GM-PET modality from the MRI and PET images. Subsequently, the semantic features are extracted from the GM-PET images. Finally, the classifier consisting of a fully connected (FC) layer and a softmax layer is used to classify subjects from different groups.

The proposed multimodal image fusion can merge complementary information from different modality images so that the composite modality conveys a better description of the information than the individual input images. As depicted in **Figure 2**, our proposed image fusion method only extracts the GM area that is critical for AD diagnosis from FDG-PET, using the MRI scan as an anatomical mask. The GM-PET modality contains both structural MRI information and functional PET information. The details of our image fusion method include the following steps.

(a) Skull-stripping is performed on structural MRI scans using the "watershed" module in FreeSurfer 6.0 (29), as shown in **Figure 2A**. The watershed segmentation algorithm can strip skull and other outer non-brain tissue to produce the brain volume with much less noise and irrelevant information. As expected, the result, called SS-MRI, preserves only the intracranial tissue structure and removes areas of irrelevant anatomical organs.

(b) As shown in **Figure 2B**, SS-MRI is affine transformed to MNI152 space (30), a universal brain atlas template, using the FLIRT (FMRIB's Linear Image Registration Tool) module (31) in the FSL package. FLIRT is a fully automated robust and accurate tool for intra- and inter-modal brain image registration by linear affine (31, 32). The registration aims to remove any spatial discrepancies between subjects in the scanner and minimize translations and rotations from a standard orientation. This helps to improve the precision of the subsequent tissue segmentation. This registered MNI-MRI is used as the input modality to unimodal AD classification tasks.

(c) The GM area is segmented from the MRI scan using the FAST (FMRIB's Automated Segmentation Tool) module (33) in the FSL package. FAST segments a 3D brain image

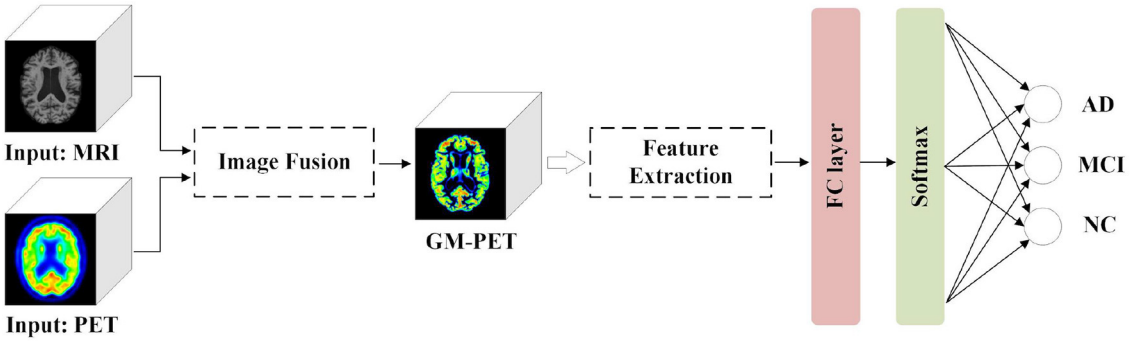

FIGURE 1 | Proposed AD diagnostic framework with multimodal image fusion method.

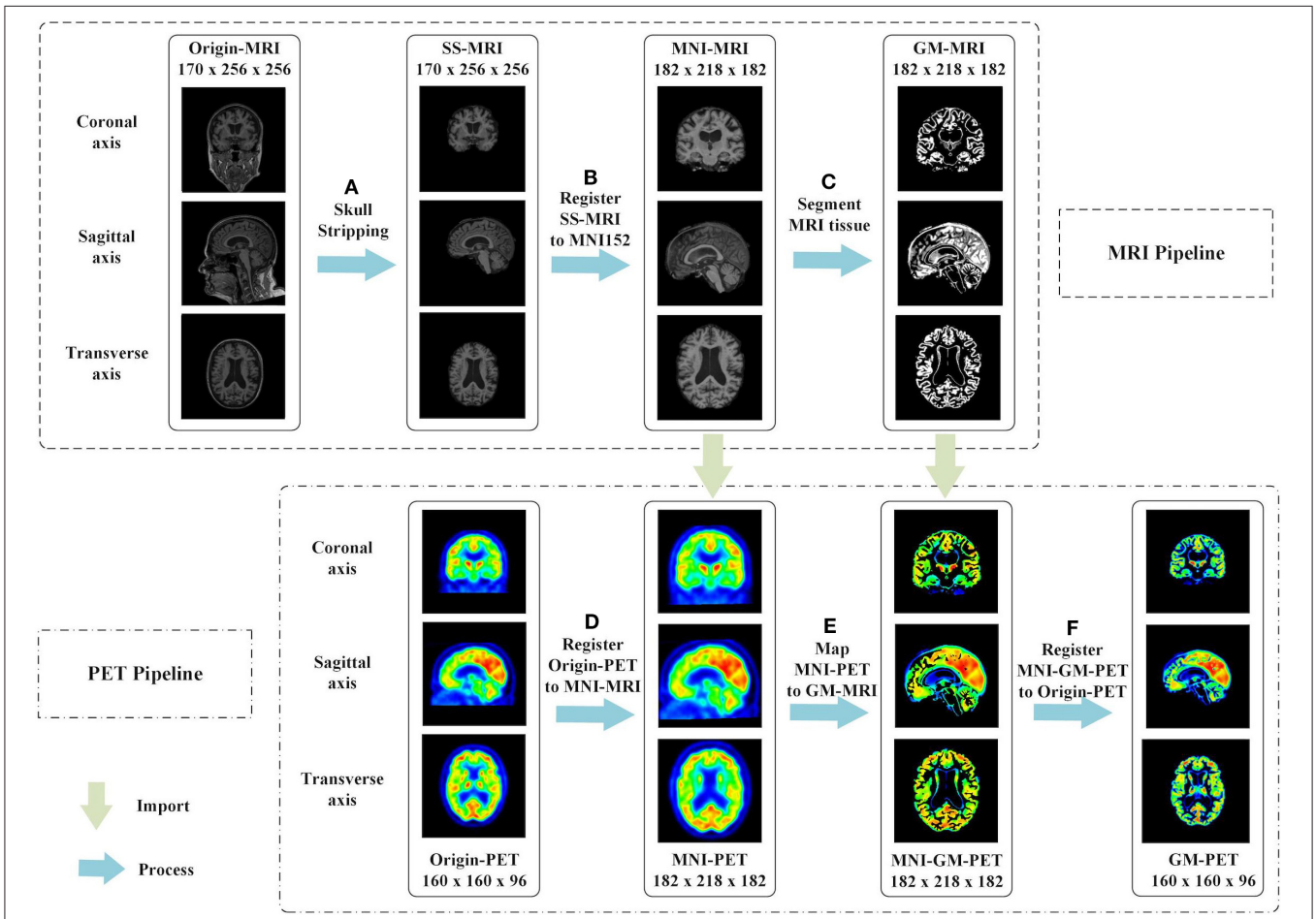

FIGURE 2 | Proposed multimodal image fusion method. In the MRI pipeline, we executed the following steps in sequence: **(A)** skull-stripping, **(B)** registration of SS-MRI to MNI152, and **(C)** segmentation of MRI tissue. The phased output of the MRI pipeline guided the subsequent processing of PET images, as shown by the green arrows. In the PET pipeline, we performed the following steps: **(D)** registration of Origin-PET to MNI-MRI, **(E)** mapping MNI-PET to GM-MRI, and **(F)** registration of MNI-GM-PET to Origin-PET.

into different tissue types, while correcting for spatial intensity variations (also known as bias field or RF inhomogeneities). The underlying method is based on a hidden Markov random field model and an associated expectation-maximization algorithm. The whole automated process can produce a bias-field-corrected input image and probabilistic and/or partial volume tissue segmentation. It is robust and reliable compared with most finite mixture model-based methods, which are sensitive to noise. As

shown in **Figure 2C**, the segmentation output of GM tissue is called GM-MRI.

(d) MNI-PET is obtained by co-registering the FDG-PET image to its respective MNI-MRI image using the FSL FLIRT module, as shown in **Figure 2D**. This gives the FDG-PET image the same spatial orientation, image size (for example, 182 × 218 × 182), and voxel dimensions (for example, 1.0 × 1.0 × 1.0 mm) as the MNI-MRI. After co-registration, the MNI-PET and MNI-MRI obtained are in the same sample space.

(e) The GM-MRI obtained in step (c) is used as an anatomical mask to cover the full MNI-PET image. MNI-GM-PET is obtained by a mapping operation, as illustrated in **Figure 2E**. So far, we have obtained the anatomical structure of GM on FDG-PET images. Nevertheless, compared with Origin-PET from coronal-axis and transverse-axis views, the mapped grayscale values in MNI-GM-PET images change significantly after MNI152 spatial registration; thus, they cannot reflect the true metabolic information as the Origin-PET does.

(f) In order to solve the grayscale deviation problem mentioned above, MNI-GM-PET is co-registered to the corresponding Origin-PET image, using the FSL FLIRT module, to obtain the GM-PET image, as shown in **Figure 2F**. On the one hand, this registration operation eliminates the deviation caused by affine transformation and preserves the true grayscale distribution of the original PET image; on the other hand, it ensures that the GM-PET has the same spatial size as the Origin-PET, that is, the MNI-GM-PET size of 182 × 218 × 182 is reduced to the original PET size of 160 × 160 × 96. This resolution reduction could also save computational time and memory costs.

2.3. Networks

At present, CNN is attracting increasing attention owing to its significant advantages in medical image classification tasks. In two-dimensional (2D) CNN approaches, where the 3D medical image is processed slice-by-slice, the anatomical context in directions orthogonal to the 2D plane is completely discarded. As discussed recently by (34), 3D CNN can greatly improve performance by considering the 3D data as a whole input, although the computational complexity and memory cost are increased owing to the larger number of parameters. To evaluate the effectiveness of the fused GM-PET modality in different CNNs, this paper introduces the 3D Simple CNN and 3D Multi-Scale CNN, designed by observing the characteristics of AD classification tasks, which will be explained in detail below.

2.3.1. 3D Simple CNN

Considering the tradeoffs between the feature capture capabilities of 3D CNN and the potential overfitting risk caused by a small dataset, we propose a 3D Simple CNN to capture AD features from medical images. As shown in **Figure 3**, the 3D Simple CNN contains 11 layers, of which there are only four convolutional layers. Compared with deeper networks, the 3D Simple CNN has far fewer parameters and can better alleviate overfitting problems.

Specifically, the base building block, called Conv-block(s, n), consists of three serial operations: Conv3D(s, n), which stands for 3D convolution with n filters of $s \times s \times s$ size, batch normalization (35), and a rectifier linear unit (ReLU). In this architecture, the "Feature Extraction" module is mainly composed of four Conv-blocks with parameters (3,8), (3,16), (3,32), and (3,64). That is, the convolution kernel sizes are (3, 3, 3), and the number of channels doubles in turn. There is also a 3D max-pooling layer with a pooling size of (2, 2, 2) between every two Conv-blocks. Besides, we add a global average pooling (GAP) layer and a dropout layer with a rate of 0.6 to avoid overfitting. After the Feature Extraction module, we connect an FC layer and a softmax layer for AD classification. In general, the 3D Simple CNN can be regarded as a baseline network for evaluating our image fusion method because of its plain structural composition.

2.3.2. 3D Multi-Scale CNN

Numerous UNet-based networks have been proven effective in biomedical image recognition tasks (36–38), as the U-shaped network architecture with skip connections can obtain both relevant context information and precise location information. Motivated by the observation that features both from low-level image volumes and high-level semantic information can be obtained at different resolution scales, a 3D Multi-Scale CNN is proposed for AD classification, as shown in **Figure 4**.

The Feature Extraction module is used to extract and merge multi-scale features, and a classifier module consisting of an FC layer and a softmax layer predicts the group labels. The Feature Extraction module consists of seven convolutional layers (Conv1–Conv7) where the first four convolutional layers generate feature maps in a coarse-to-fine manner, and the last two layers (Conv6 and Conv7) are obtained by up-sampling the combined output of the "skip connection." These convolutional layers are designed using a conventional CNN structure with kernel sizes of (3, 3, 3) and channel numbers as shown in **Figure 4**. Taking into account the overfitting problem, we properly reduce the channel numbers of convolutional layers. Detailed image features are often related to shallow layers, whereas semantically strong features are often associated with deep layers. It is desirable to obtain both types of features for AD classification by integrating information from different scales. Hence, the skip connection is used to combine features from both shallow and deep convolutional layers. More specifically, the down-sampled outputs of convolutional layers 1 and 2 are combined with the outputs of convolutional layers 7 and 6, respectively. Besides, the outputs of convolutional layers 4 and 5 are concatenated. Owing to the limitations of GPU memory when using 3D scans as inputs, three scales are used here. For each scale feature, we apply a GAP layer and a dropout layer to retain multi-resolution features, after which the outputs are concatenated to feed the following classifier. It is expected that multi-scale features with different levels of information will contribute to the diagnosis of AD.

3. EXPERIMENT AND RESULTS

3.1. Pre-processing

As inputs to CNN, 3D data with a generally high resolution would consume more computing resources during network training. Therefore, we process the input data using cropping and

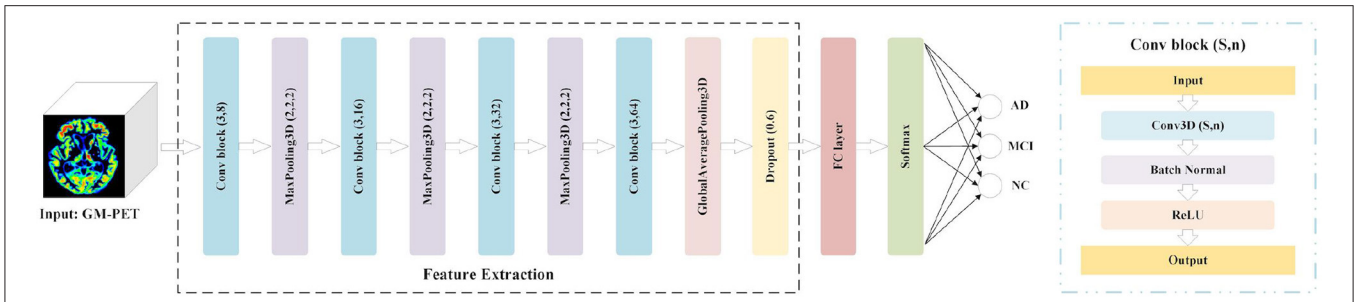

FIGURE 3 | 3D Simple CNN architecture for AD classification.

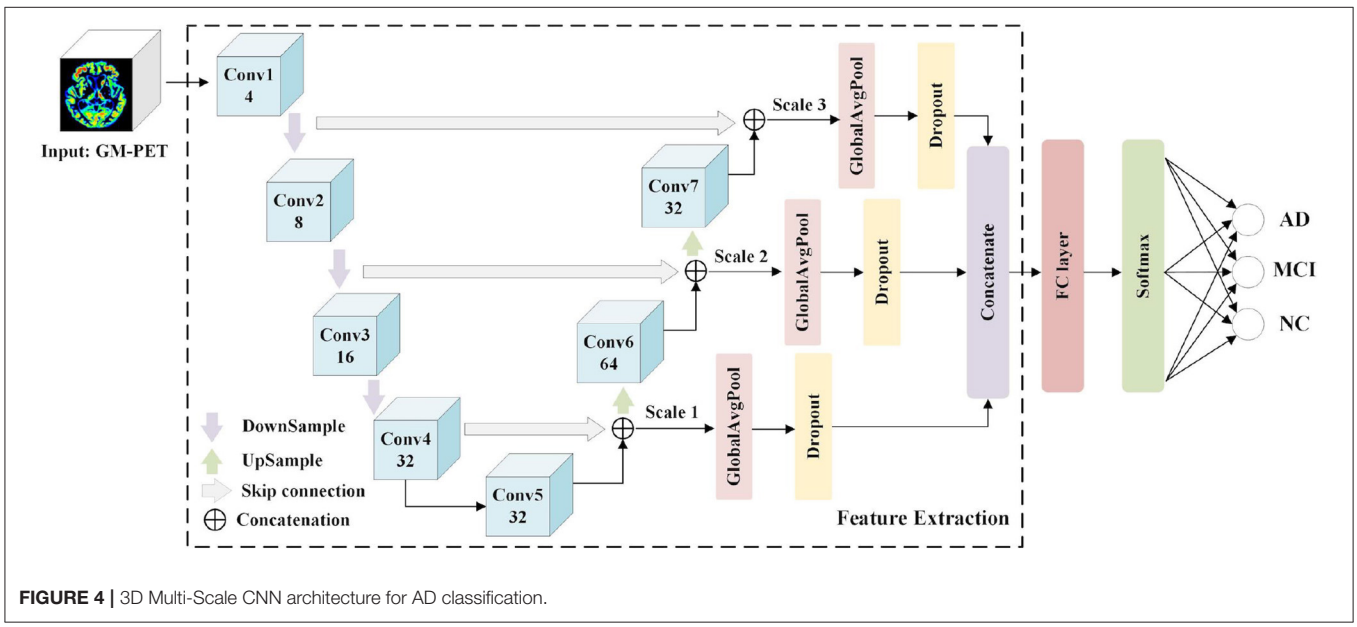

FIGURE 4 | 3D Multi-Scale CNN architecture for AD classification.

sampling operations to speed up the calculation of singleton data. (1) Cropping: As shown in **Figure 2**, there are many background areas with a pixel value of 0 outside the brain tissue area in each modality image. Without affecting the brain tissue regions, we appropriately reduce these meaningless background areas to decrease the size of the input data. Specifically, MRI is cropped from 182 × 218 × 182 to 176 × 208 × 176. In addition, PET and GM-PET are both cropped from 160 × 160 × 96 to 112 × 128 × 96. (2) Sampling: Each sample is divided into two by taking every other slice along the transverse axis. Concretely, the sizes of the MRI, PET, and GM-PET images become 176 × 208 × 88, 112 × 128 × 48, and 112 × 128 × 48, respectively. This can double the number of samples while reducing the resolution, which is conducive to better iteration and optimization of the network model.

3.2. Experimental Setup

In this paper, the networks involved are implemented in the Tensorflow (39) deep learning framework. We execute four classification tasks, i.e., AD vs. NC, AD vs. MCI, MCI vs. NC, and AD vs. MCI vs. NC, whereas previous studies such as (40) and (41) only classified AD vs. NC, which are the easiest groups to distinguish. We conduct comparative experiments on unimodal and multimodal data. For the network optimizer, Adam with an initial learning rate of 1e-4 is used to update the weights during training. The binary cross-entropy is applied as the loss function in the binary-classification task, whereas the categorical cross-entropy is used in the three-classification task.

We adopt a 10-fold cross-validation strategy to calculate the measures, so as to obtain a fairer performance comparison. We randomly divide the subjects in the dataset into 10 subsets, with one subset used as the test set, another subset used as the validation set, and the remaining eight subsets used as the training set. We train each experiment during 500 epochs and use two strategies to update the learning rate. (1) When the loss in the validation set does not decrease within 30 epochs, the learning rate drops to one-tenth of the current level. (2) When the accuracy in the validation set does not increase within 20 epochs, the learning rate is reduced by half. At the same time, an early stopping strategy is applied. That is, the training is stopped if the loss on validation does not decrease within 50 epochs. The classification accuracy (ACC), sensitivity (SEN), and specificity (SPE) are selected as the evaluation measures. We report the results as the mean ± SD (standard deviation) of the 10-fold tests.

We aim to comprehensively evaluate the effectiveness of our image fusion method in the proposed diagnostic framework for AD classification tasks. In addition to considering other unimodal scans (for example, MRI and PET) as inputs, we present an AD diagnostic framework with the feature fusion method as a benchmark. As shown in **Figure 5**, the Feature Extraction module is used to obtain semantic information from the 3D volumes of MRI and PET images, respectively. After the extracted features are concatenated, three FC layers with unit numbers of 64, 32, and 16, respectively, perform the correlation fusion. Moreover, a GAP layer and a dropout layer are applied to avoid overfitting. Finally, the classification module, which consists of an FC layer and a softmax layer, predict the group labels.

3.3. Performance

3.3.1. Results for AD vs. NC

In the classification of AD vs. NC, **Table 2** shows the results of unimodal and multimodal modalities with different networks. The multi-modality-based methods such as the feature fusion method and the proposed image fusion method achieve better performance, because they successfully fuse MRI and PET information. Between the two multimodal methods, our image fusion method has better overall indicators. With the 3D Simple CNN, our image fusion method obtained the best classification accuracy of 94.11 ± 6.0% and specificity of 95.04 ± 5.7%, and the second best sensitivity of 92.22 ± 6.7%. The feature fusion method achieved the best sensitivity of 94.44 ± 7.9% but showed lower accuracy and specificity. With the 3D Multi-Scale CNN, the proposed image fusion method for AD diagnosis achieved the best classification accuracy of 94.11 ± 4.0%, sensitivity of 93.33 ± 7.8%, and specificity of 94.27 ± 6.3%. Moreover, it showed improvements in classification accuracy, sensitivity, and specificity over the unimodal methods of at least 4.75, 6.27, and 3.46%, respectively. Overall, our image fusion method achieved the overall best performance in the AD vs. NC classification task.

3.3.2. Results for MCI vs. NC

Table 3 shows the results for different modalities in the classification of MCI vs. NC with different networks. The proposed image fusion method showed significant performance superiority. With the 3D Simple CNN, our image fusion method achieved the best classification accuracy of 88.48 ± 6.5%, sensitivity of 93.44 ± 6.5%, and specificity of 82.18 ± 12.3%. It also showed improvements in classification accuracy, sensitivity, and specificity over the feature fusion method of at least 6.11, 1.25, and 11.62%, respectively, indicating that the proposed image fusion method fuses multimodal information in a more effective way. When applying the 3D Multi-Scale CNN, our image fusion method still achieved the best accuracy of 85.00 ± 9.4% and specificity of 85.60 ± 11.7%, and

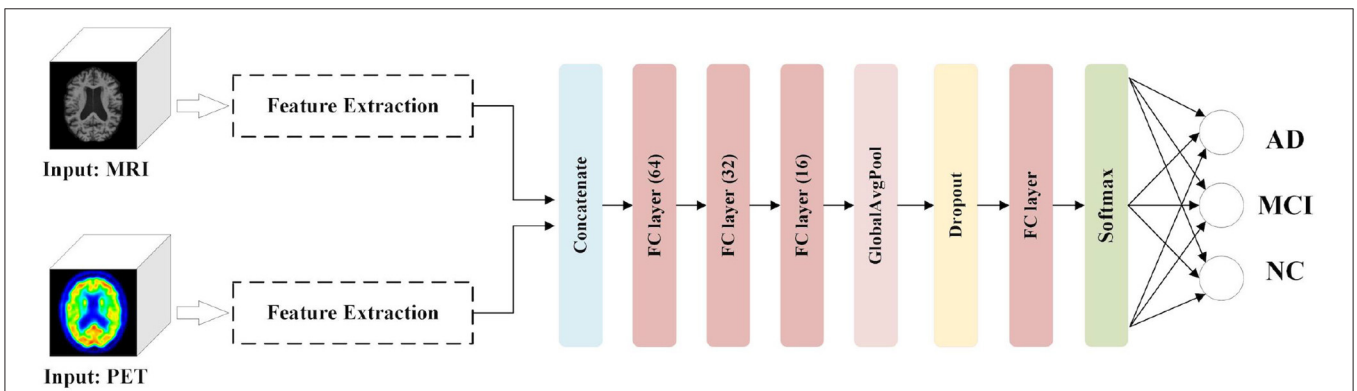

FIGURE 5 | AD diagnostic framework with multimodal feature fusion method.

TABLE 2 | Results of different modalities with different networks for AD vs. NC (UNIT:%).

Network	Modalities	ACC	SEN	SPE
3D Simple CNN	Unimodal MRI	89.80 ± 4.7	86.31 ± 12.0	91.97 ± 5.5
	Unimodal PET	92.10 ± 5.8	89.13 ± 9.7	94.27 ± 4.1
	Feature fusion	93.22 ± 3.8	**94.44 ± 7.9**	91.62 ± 7.5
	Proposed image fusion	**94.11 ± 6.0**	92.22 ± 6.7	**95.04 ± 5.7**
3D Multi-Scale CNN	Unimodal MRI	88.88 ± 6.8	86.11 ± 13.9	90.43 ± 4.5
	Unimodal PET	89.36 ± 9.1	87.06 ± 16.3	90.81 ± 7.5
	Feature fusion	93.66 ± 5.3	93.33 ± 9.4	93.50 ± 6.3
	Proposed image fusion	**94.11 ± 4.0**	**93.33 ± 7.8**	**94.27 ± 6.3**

Bold value mean the best indicator value under the same conditions.

the second best sensitivity of 84.69 ± 12.5%. In terms of specificity, our method far exceeded other methods by at least 11.33%. Generally speaking, the proposed image fusion method achieved the overall best performance in the MCI vs. NC classification task.

3.3.3. Results for AD vs. MCI

In the classification of AD vs. MCI, **Table 4** shows the results of unimodal and multimodal modalities with different networks. With the 3D Simple CNN, our image fusion method for AD diagnosis achieved the best classification accuracy of 84.83 ± 7.8% and specificity of 94.69 ± 6.3%, and the second best sensitivity of 68.29 ± 19.8%. Moreover, the proposed image fusion method showed improvements in classification accuracy, sensitivity, and specificity over the unimodal methods by at least 6.53, 10.83, and 5.00%, respectively. With the 3D Multi-Scale CNN, our image fusion method obtained the best classification accuracy of 80.80 ± 5.9% and sensitivity of 71.19 ± 14.6%, and the second best specificity of 85.94 ± 11.8%. Compared with the feature fusion method, which achieved the best specificity, the proposed image fusion method showed improvements in classification accuracy and sensitivity of 0.33 and 17.78%, respectively. On the whole, our method outperformed the other methods and showed the best overall performance in the AD vs. MCI classification task.

3.3.4. Results for AD vs. MCI vs. NC

Table 5 shows the results of different modalities for the classification of AD vs. MCI vs. NC with the 3D Simple CNN and 3D Multi-Scale CNN. As MCI is a transitional state between AD and NC, many confounding factors are introduced in the multi-class task. Clearly, the classification task of AD vs. MCI vs. NC is more difficult than the above binary-classification tasks. In this case, our image fusion method still showed the best performance on all evaluation indices, whereas the unimodal and feature fusion methods were particularly lacking in power for the three-classification task. With the 3D Simple CNN, the best classification accuracy, sensitivity, and specificity were 74.54 ± 6.4, 59.41 ± 8.2, and 85.41 ± 4.2%, respectively. Compared with other methods, our image fusion method showed improvements in classification accuracy, sensitivity, and specificity of at least 9.06, 10.73, and 6.27%, respectively. With the 3D Multi-Scale CNN, our image fusion method achieved the best classification accuracy of 71.52 ± 5.0%, sensitivity of 55.67 ± 6.2%, and specificity of 83.40 ± 3.3%. Furthermore, our image fusion method showed improvements in classification accuracy, sensitivity, and specificity over the other methods of at least 3.37, 4.03, and 2.37%, respectively. Clearly, our image fusion method showed significant advantages in the multi-class task.

3.3.5. Comparisons With State-of-the-Art Methods

The proposed image fusion method was evaluated and compared with the state-of-the-art multimodal approaches for each task-specific classification (**Table 6**). The results indicate that our method (Image Fusion + 3D Simple CNN) achieved the highest accuracy and outperformed other multimodal methods for each AD diagnostic task. Although our multimodal image fusion

TABLE 3 | Results of different modalities with different networks for MCI vs. NC (UNIT:%).

Network	Modalities	ACC	SEN	SPE
3D Simple CNN	Unimodal MRI	79.46 ± 9.4	87.50 ± 16.1	69.15 ± 10.7
	Unimodal PET	72.00 ± 7.8	72.81 ± 10.5	70.56 ± 12.2
	Feature fusion	82.37 ± 9.0	92.19 ± 13.1	69.74 ± 18.0
	Proposed image fusion	**88.48 ± 6.5**	**93.44 ± 6.5**	**82.18 ± 12.3**
3D Multi-Scale CNN	Unimodal MRI	76.01 ± 8.8	77.50 ± 13.4	74.27 ± 9.7
	Unimodal PET	68.55 ± 5.4	65.94 ± 13.5	70.64 ± 14.8
	Feature fusion	83.17 ± 6.5	**90.63 ± 15.7**	73.55 ± 16.7
	Proposed image fusion	**85.00 ± 9.4**	84.69 ± 12.5	**85.60 ± 11.7**

Bold value mean the best indicator value under the same conditions.

TABLE 4 | Results of different modalities with different networks for AD vs. MCI (UNIT:%).

Network	Modalities	ACC	SEN	SPE
3D Simple CNN	Unimodal MRI	72.47 ± 7.8	46.59 ± 18.8	87.50 ± 12.1
	Unimodal PET	78.30 ± 10.3	57.46 ± 20.1	89.69 ± 10.9
	Feature fusion	81.00 ± 8.1	**68.33 ± 15.3**	88.75 ± 9.2
	Proposed image fusion	**84.83 ± 7.8**	68.29 ± 19.8	**94.69 ± 6.3**
3D Multi-Scale CNN	Unimodal MRI	68.40 ± 8.4	52.70 ± 19.7	77.50 ± 11.9
	Unimodal PET	73.07 ± 15.3	61.90 ± 27.6	79.38 ± 16.9
	Feature fusion	80.47 ± 9.4	53.41 ± 25.1	**95.94 ± 5.1**
	Proposed image fusion	**80.80 ± 5.9**	**71.19 ± 14.6**	85.94 ± 11.8

Bold value mean the best indicator value under the same conditions.

TABLE 5 | Results of different modalities with different networks for AD vs. MCI vs. NC (UNIT:%).

Network	Modalities	ACC	SEN	SPE
3D Simple CNN	Unimodal MRI	64.00 ± 8.6	47.10 ± 9.5	78.08 ± 6.5
	Unimodal PET	60.65 ± 9.7	43.50 ± 10.6	75.49 ± 7.3
	Feature fusion	65.48 ± 5.9	48.68 ± 6.7	79.14 ± 4.3
	Proposed image fusion	**74.54 ± 6.4**	**59.41 ± 8.2**	**85.41 ± 4.2**
3D Multi-Scale CNN	Unimodal MRI	66.24 ± 5.9	49.56 ± 6.6	79.72 ± 4.3
	Unimodal PET	59.98 ± 7.1	42.83 ± 7.0	74.98 ± 5.9
	Feature fusion	68.15 ± 9.4	51.64 ± 10.5	81.03 ± 6.9
	Proposed image fusion	**71.52 ± 5.0**	**55.67 ± 6.2**	**83.40 ± 3.3**

Bold value mean the best indicator value under the same conditions.

TABLE 6 | Comparative performance of our classifiers vs. competitors. Numbers in parentheses denote the numbers of AD/MCI/NC subjects in the dataset used.

Approach	Dataset	Accuracy (%)			
		AD vs. NC	MCI vs. NC	AD vs. MCI	AD vs. MCI vs. NC
(42)	MRI+PET (85/169/77)	91.4	82.1	–	53.79
(20)	MRI+PET (51/99/52)	91.4	77.4	70.1	–
(21)	MRI+PET+CSF+Genetic (37/75/35)	91.8	79.5	–	60.2
(23)	MRI+PET (238/217/360)	84.59	85.96	–	–
(24)	MRI+PET (93/204/100)	93.26	74.34	–	–
(10)	MRI+PET+CSF (210/541/160)	88.02	84.14	–	–
(43)	MRI+PET (160/187/160)	92.51	82.53	–	–
(19)	fMRI+SNP (37/37/35)	81.0	80.0	–	–
Our Method (Image Fusion+3D Simple CNN)	MRI+PET (95/160/126)	**94.11**	**88.48**	**84.83**	**74.54**

Bold value mean the best indicator value under the same conditions.

method is time-consuming during the pre-processing steps, the network parameters are greatly reduced because only the composite image is fed into the classification network instead of a set of images of different modalities. In other words, the computation complexity and the memory cost of the proposed image fusion method are no higher than those of competing methods.

3.4. Visualization

To further illustrate the plausibility of our image fusion method, we visualized origin images and the corresponding features in different modalities for different subject groups, as shown in **Figure 6**. The picture on the left in each cell is a slice of the subject in different modalities. From the MRI and PET modality slices, we observed that the AD subject had the most obvious brain tissue loss and decrease in metabolism, respectively, followed by the MCI subject, whereas the NC subject had a healthy brain imaging scan. From the GM-PET slices, we observed that the GM area was delineated while maintaining the same pattern as that of the PET modality. GM-PET well-inherited the ability of MRI to express atrophy of brain tissue and the ability of PET to observe metabolic levels. As only the GM region was retained, there was no noise information around the brain tissue in the GM-PET images; in particular, the irrelevant skull area was cleanly removed. Based on the richness of the information expressed by the images, there is no doubt that our proposed image fusion method achieved better results.

It was worth investigating whether the multimodal GM-PET provided the feature extraction module of the CNN with ample information. We applied 3D Grad-CAM technology (44) to visualize the region of interest in the second convolutional layer of the 3D Simple CNN, shown as the right picture of each cell in **Figure 6**. The highlighted areas in the output images of Grad-CAM represent the key areas on which the convolutional layer focuses. In the outputs of the MRI slices, the focus was on the contour and edge texture areas, as outlined by the red circles. In

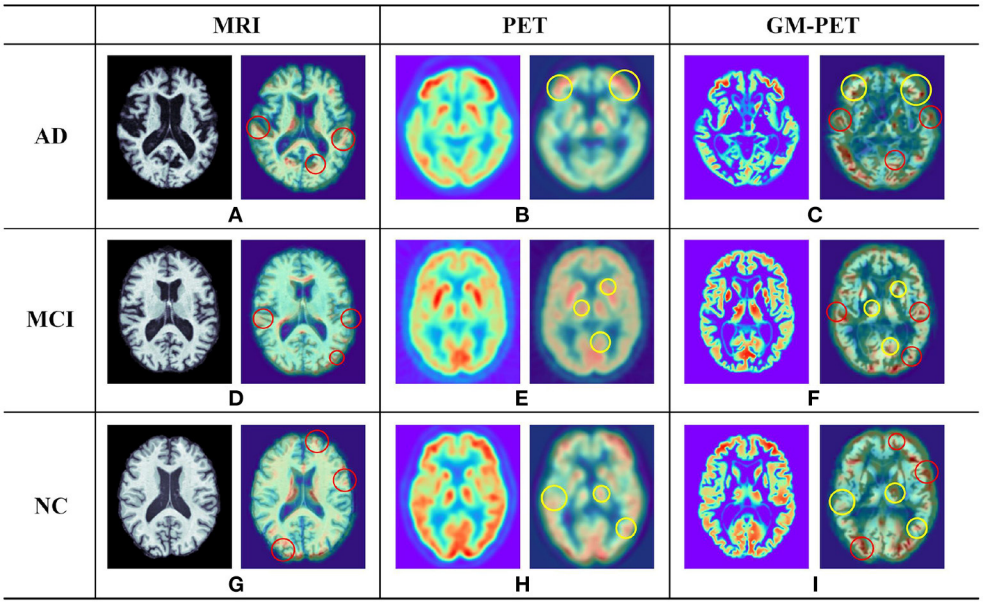

FIGURE 6 | Examples of different modality images for AD, MCI, and NC subjects. In each of the nine cells **(A–I)**, the picture on the left is a subject slice and the picture on the right is the Grad-CAM result for that slice. The red circle in the 3D Grad-CAM results outlines the contour areas of common interest in the MRI and GM-PET images, while the yellow circle outlines the metabolic characteristic areas of common interest in the PET and GM-PET images.

the outputs of the PET slices, the areas of interest were highly consistent with the areas of high metabolic levels, as represented by the yellow circles. As expected, the convolutional layer on GM-PET considered both contour and metabolic information at the same time. Namely, the GM-PET modality provides more abundant characteristics for AD diagnosis.

4. DISCUSSION

As multimodal data can provide more comprehensive pathological information, we propose an image fusion method to effectively merge the multimodal neuroimaging information from MRI and PET scans for AD diagnosis. Based on the observation that GM is the tissue area of most interest in AD diagnostic researches (10, 11, 45), the proposed fusion method extracts and fuses the GM tissue of brain MRI and FDG-PET in the image field so as to obtain a fused GM-PET modality. As can be seen from the image fusion flow, shown in **Figure 2**, the GM-PET image not only reserves the subject's brain structure information from MRI but also retains the corresponding metabolic information from PET. With the 3D Grad-CAM technology, we observe that the convolutional layer that extracts the GM-PET features can capture both contour and metabolic information, indicating that the GM-PET modality can indeed provide richer modality information for classification tasks. Moreover, our proposed image fusion method, through its registration operation, better solves the heterogeneous features alignment problem between multimodal images, compared with methods based on multimodal feature learning.

In addition, the 3D Simple CNN and 3D Multi-Scale CNN are presented to perform four AD classification tasks, comprising three binary-classification tasks, i.e., AD vs. NC, AD vs. MCI and MCI vs. NC, and one multi-classification task, AD vs. MCI vs. NC. The 3D Simple CNN, with a plain structure, was proposed first as a baseline network. Then we proposed a 3D Multi-Scale CNN network that combines information from different scale features while capturing context information and location information. In order to prevent over-fitting, we designed these two networks using the following strategies: 1) Use fewer convolutional layers; (2) reduce the number of channels of the convolutional layer; (3) use GAP and dropout layers to reduce redundant information. Furthermore, the proposed AD diagnostic framework uses a single-input network instead of the multiple-input network used in feature fusion methods, as our image fusion method fuses multimodal image scans into a single composite image. Therefore, our image fusion method can greatly reduce the number of CNN parameters.

Extensive experiments and analyses were carried out to evaluate the performance of our proposed image fusion method. According to the classification results shown in **Tables 2**–**5**, the multimodal methods, including feature fusion and the proposed image fusion method, achieved better performance than the unimodal methods, as the multimodal methods contained abundant and complementary information. Our image fusion method outperformed the feature fusion method, especially in the complex three-classification task. Moreover, both the 3D Simple CNN and 3D Multi-Scale CNN produced

consistent results indicating that our image fusion method had the best overall performance, with great adaptability to different classification networks. And our image fusion method also achieved better performance compared with the state-of-the-art multimodal-learning-based methods. Although the proposed image fusion method always showed the best accuracy, sometimes its performance was not optimal in terms of sensitivity and specificity. In order to solve this problem, we will further focus on WM and CSF tissues and combine their information with the existing GM information to provide better support for AD auxiliary diagnosis in the future.

5. CONCLUSION

We propose an image fusion method to combine MRI and PET scans into a composite GM-PET modality for AD diagnosis. The GM-PET modality contains both brain anatomic and metabolic information and eliminates image noise subtly so that the observer can easily focus on the key characteristics. To further evaluate the applicability of the proposed image fusion method, 3D Grad-CAM technology was used to visualize the area of interest of the CNN in each modality, showing that both the structural and functional characteristics of brain scans were included in the GM-PET modality. A series of evaluations based on the 3D Simple CNN and 3D Multi-Scale CNN confirmed the superiority of the proposed image fusion method. In terms of experimental performance, our proposed image fusion method not only overwhelmingly surpassed the unimodal methods but also outperformed the feature fusion method. Besides, the image fusion method showed better performance than other competing multimodal learning methods described in the literature. Therefore, our image fusion method is an intuitive and effective approach for fusing multimodal information in AD classification tasks.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JS wrote the main part of the manuscript. JZ proposed the key image fusion approach. PL and XL carried out the experiments and analyzed the results. GZ and PS built the AD diagnostic framework based on 3D CNN. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key R&D Program of China under Grant (No. 2019YFB1311600) and the Shanghai Science and Technology Committee (Nos. 18411952100 and 17411953500).

REFERENCES

1. Carrion C, Folkvord F, Anastasiadou D, Aymerich M. Cognitive therapy for dementia patients: a systematic review. *Dement Geriatr Cogn Disord*. (2018) 46:1–26. doi: 10.1159/000490851
2. Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimers Dement*. (2020) 16:391–460. doi: 10.1002/alz.12068
3. Theofilas P, Ehrenberg AJ, Nguy A, Thackrey JM, Dunlop S, Mejia MB, et al. Probing the correlation of neuronal loss, neurofibrillary tangles, and cell death markers across the Alzheimer's disease Braak stages: a quantitative study in humans. *Neurobiol Aging*. (2018) 61:1–12. doi: 10.1016/j.neurobiolaging.2017.09.007
4. Wang C, Saar V, Leung KL, Chen L, Wong G. Human amyloid β peptide and tau co-expression impairs behavior and causes specific gene expression changes in Caenorhabditis elegans. *Neurobiol Dis*. (2018) 109:88–101. doi: 10.1016/j.nbd.2017.10.003
5. Dai Z. Applications, opportunities and challenges of molecular probes in the diagnosis and treatment of major diseases. *Chin Sci Bull*. (2017) 62:25–35. doi: 10.1360/N972016-00405
6. Wenk GL. Neuropathologic changes in Alzheimer's disease. *J Clin Psychiatry*. (2003) 64(Suppl 9):7–10. Available online at: https://www.psychiatrist.com/JCP/article/Pages/neuropathologic-changes-alzheimers-disease.aspx
7. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain*. (2008) 131:681–9. doi: 10.1093/brain/awm319
8. Liu M, Zhang D, Shen D. Ensemble sparse classification of Alzheimer's disease. *Neuroimage*. (2012) 60:1106–16. doi: 10.1016/j.neuroimage.2012.01.055
9. Suk HI, Lee SW, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage*. (2014) 101:569–82. doi: 10.1016/j.neuroimage.2014.06.077
10. Zhu Q, Yuan N, Huang J, Hao X, Zhang D. Multi-modal AD classification via self-paced latent correlation analysis. *Neurocomputing*. (2019) 355:143–54. doi: 10.1016/j.neucom.2019.04.066
11. Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: *Proceedings of the International Conference on Imaging Systems and Techniques*. Beijing: IEEE (2017). p. 1–6. doi: 10.1109/IST.2017.8261460
12. Ge C, Qu Q, Gu IYH, Jakola AS. Multi-stream multi-scale deep convolutional networks for Alzheimer's disease detection using MR images. *Neurocomputing*. (2019) 350:60–9. doi: 10.1016/j.neucom.2019.04.023
13. Noble JM, Scarmeas N. Application of PET imaging to diagnosis of Alzheimer's disease and mild cognitive impairment. *Int Rev Neurobiol*. (2009) 84:133–49. doi: 10.1016/S0074-7742(09)00407-3
14. Mosconi L, Berti V, Glodzik L, Pupi A, De Santi S, de Leon MJ. Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. *J Alzheimers Dis*. (2010) 20:843–54. doi: 10.3233/JAD-2010-091504
15. Camus V, Payoux P, Barré L, Desgranges B, Voisin T, Tauber C, et al. Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *Eur J Nucl Med Mol Imaging*. (2012) 39:621–31. doi: 10.1007/s00259-011-2021-8
16. Riederer I, Bohn KP, Preibisch C, Wiedemann E, Zimmer C, Alexopoulos P, et al. Alzheimer disease and mild cognitive impairment: integrated pulsed arterial spin-labeling MRI and18F-FDG PET. *Radiology*. (2018) 288:198–206. doi: 10.1148/radiol.2018170575
17. Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*. (2011) 55:856–67. doi: 10.1016/j.neuroimage.2011.01.008
18. Li Y, Meng F, Shi J. Learning using privileged information improves neuroimaging-based CAD of Alzheimer's disease: a comparative study. *Med Biol Eng Comput*. (2019) 57:1605–16. doi: 10.1007/s11517-019-01974-3

19. Bi XA, Hu X, Wu H, Wang Y. Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. *IEEE J Biomed Health Inform*. (2020) 24:2973–83. doi: 10.1109/JBHI.2020.2973324
20. Li F, Tran L, Thung KH, Ji S, Shen D, Li J. A robust deep model for improved classification of AD/MCI patients. *IEEE J Biomed Health Inform*. (2015) 19:1610–6. doi: 10.1109/JBHI.2015.2429556
21. Tong T, Gray K, Gao Q, Chen L, Rueckert D. Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recogn*. (2017) 63:171–81. doi: 10.1016/j.patcog.2016.10.009
22. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J Biomed Health Inform*. (2018) 22:173–83. doi: 10.1109/JBHI.2017.2655720
23. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF, Weiner M, et al. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci Rep*. (2018) 8:1–13. doi: 10.1038/s41598-018-22871-z
24. Liu M, Cheng D, Wang K, Wang Y. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics*. (2018) 16:295–308. doi: 10.1007/s12021-018-9370-4
25. Punjabi A, Martersteck A, Wang Y, Parrish TB, Katsaggelos AK. Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks. *PLoS ONE*. (2019) 14:e0225759. doi: 10.1371/journal.pone.0225759
26. Rajalingam B, Priya R, Bhavani R. Multimodal medical image fusion using hybrid fusion techniques for neoplastic and Alzheimer's disease analysis. *J Comput Theor Nanosci*. (2019) 16:1320–1331. doi: 10.1166/jctn.2019.8038
27. Jack Jr CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging*. (2008) 27:685–91. doi: 10.1002/jmri.21049
28. Liu M, Zhang J, Yap PT, Shen D. View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. *Med Image Anal*. (2017) 36:123–34. doi: 10.1016/j.media.2016.11.002
29. Bartos A, Gregus D, Ibrahim I, Tintěra J. Brain volumes and their ratios in Alzheimer's disease on magnetic resonance imaging segmented using Freesurfer 6.0. *Psychiatry Res Neuroimaging*. (2019) 287:70–4. doi: 10.1016/j.pscychresns.2019.01.014
30. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*. (2011) 54:313–27. doi: 10.1016/j.neuroimage.2010.07.033
31. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. (2002) 17:825–41. doi: 10.1006/nimg.2002.1132
32. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal*. (2001) 5:143–56. doi: 10.1016/S1361-8415(01)00036-6
33. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*. (2001) 20:45–57. doi: 10.1109/42.906424
34. Milletari F, Ahmadi SA, Kroll C, Plate A, Rozanski V, Maiostre J, et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Comput Vis Image Und*. (2017) 164:92–102. doi: 10.1016/j.cviu.2017.04.002
35. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *32nd International Conference on Machine Learning*. Lille: JMLR (2015). p. 448–56.
36. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28
37. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans Med Imaging*. (2018) 37:2663–74. doi: 10.1109/TMI.2018.2845918
38. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2020) 18:203–11. doi: 10.1038/s41592-020-01008-z
39. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation*. Savannah, GA: USENIX Association (2016). p. 265–83.
40. Sarraf S, DeSouza D, Anderson J, Tofighi G. DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*. (2016) 070441. doi: 10.1101/070441
41. Cheng D, Liu M. CNNs based multi-modality classification for AD diagnosis. In: *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*. Shanghai: IEEE (2018). p. 1–5. doi: 10.1109/CISP-BMEI.2017.8302281
42. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans Biomed Eng*. (2015) 62:1132–40. doi: 10.1109/TBME.2014.2372011
43. Shao W, Peng Y, Zu C, Wang M, Zhang D. Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease. *Comput Med Imaging Graph*. (2020) 80:101663. doi: 10.1016/j.compmedimag.2019.101663
44. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the International Conference on Computer Vision*. Venice: IEEE (2017). p. 618–26. doi: 10.1109/ICCV.2017.74
45. Zhou T, Thung KH, Liu M, Shi F, Zhang C, Shen D. Multi-modal neuroimaging data fusion via latent space learning for Alzheimer's disease diagnosis. In: *Proceedings of the International Workshop on Predictive Intelligence in Medicine*. Cham: Springer (2018). p. 76–84. doi: 10.1007/978-3-030-00320-3_10

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 Physiology

ORIGINAL RESEARCH
published: 02 March 2021
doi: 10.3389/fphys.2021.643202

Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition

Hua Zhang[1,2], *Ruoyun Gou*[1], *Jili Shang*[1], *Fangyao Shen*[1], *Yifan Wu*[1,3]* *and Guojun Dai*[1]

[1] School of Computer Science and Technology, HangZhou Dianzi University, Hangzhou, China, [2] Key Laboratory of Network Multimedia Technology of Zhejiang Province, Zhejiang University, Hangzhou, China, [3] Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, HangzhouDianzi University, Hangzhou, China

Speech emotion recognition (SER) is a difficult and challenging task because of the affective variances between different speakers. The performances of SER are extremely reliant on the extracted features from speech signals. To establish an effective features extracting and classification model is still a challenging task. In this paper, we propose a new method for SER based on Deep Convolution Neural Network (DCNN) and Bidirectional Long Short-Term Memory with Attention (BLSTMwA) model (DCNN-BLSTMwA). We first preprocess the speech samples by data enhancement and datasets balancing. Secondly, we extract three-channel of log Mel-spectrograms (static, delta, and delta-delta) as DCNN input. Then the DCNN model pre-trained on ImageNet dataset is applied to generate the segment-level features. We stack these features of a sentence into utterance-level features. Next, we adopt BLSTM to learn the high-level emotional features for temporal summarization, followed by an attention layer which can focus on emotionally relevant features. Finally, the learned high-level emotional features are fed into the Deep Neural Network (DNN) to predict the final emotion. Experiments on EMO-DB and IEMOCAP database obtain the unweighted average recall (UAR) of 87.86 and 68.50%, respectively, which are better than most popular SER methods and demonstrate the effectiveness of our propose method.

Keywords: speech emotion recognition, deep convolutional neural network, attention mechanism, long short-term memory, deep neural network

OPEN ACCESS

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Maximilian Schmitt,
University of Augsburg, Germany
Zhehuai Chen,
Google, United States

***Correspondence:**
Yifan Wu
yfwu@hdu.edu.cn

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *17 December 2020*
Accepted: *01 February 2021*
Published: *02 March 2021*

Citation:
Zhang H, Gou R, Shang J, Shen F, Wu Y and Dai G (2021) Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition. Front. Physiol. 12:643202. doi: 10.3389/fphys.2021.643202

1. INTRODUCTION

As the most natural and convenient medium in human communication, speech signals not only contain the linguistic information like semantic and language type, but also contain rich non-linguistic information, such as facial expression, speech emotion, and so on. In recent years, with the continuous development of artificial intelligence, speech emotion recognition (SER) plays a crucial role in human-machine interactions (Ayadi et al., 2011). More and more researchers are attracted by the study that computer automatically recognize speech emotions of people. Speech emotion recognition has become an attractive research topic in many fields, such as speaker's semantic and culture, but also contain a wealth of paralinguistic information, such as emotion.

Speech emotion recognition is under great challenges. Firstly, there are too few datasets in the speech field as it is difficult and time-consuming to build high-quality speech emotion database. Secondly, different data in the database has different speakers whose gender, age, language, and

culture etc are different. Finally, the emotions in speech are often based on sentences rather than just certain words. So how to use LLDs and sentence-level features to improve the accuracy of emotion recognition is a difficult point in current research. The traditional speech emotion recognition methods usually contain three steps (Deng et al., 2014). The first step is data preprocessing, including data normalization, speech segmentation, and other operations. Next step is feature extraction from the speech signals using some machine learning algorithms. These features are usually called Low-Level Descriptors (LLDs), such as Fundamental Frequency(F0) (Origlia et al., 2010), Formant (Deng et al., 2013), Mel Frequency Cepstrum Coefficient (MFCC) (Milton et al., 2014), etc. Finally, appropriate classifiers are selected for speech emotion classification, including Support Vector Machine (SVM) (Chen et al., 2012), Gaussian Mixture Model (GMM) (Bhaykar et al., 2013), Hidden Markov model (Schuller et al., 2003), etc. However, a major disadvantage of those methods lies in the involved traditional machine learning technology which requires prior knowledge of all necessary features (such as fundamental frequency, energy, etc.) affecting emotion recognition. And the extraction process may lose some important information.

To address this problem, the deep learning techniques provide reasonable solutions in feature extraction for SER. One of the most popular deep learning methods is DNN, which have shown excellent performances in extracting discriminative features especially in image classification. For speech emotion recognition, using deep learning technology can automatically extract deep speech emotional features and learn the correlation between features. It has shown better performance compared with the traditional methods.

Originally, Han et al. (2014) proposed a DNN and ELM model in 2014, which adopted the highest energy fragments to train DNN model and extract effective emotional information. In 2014, Mao et al. (2014) first used convolutional neural network (CNN) to learn the emotional salient features of SER, and demonstrated the feasibility of CNN model on several benchmark data sets. Lee and Tashev (2015) used bidirectional long short-term memory (BLSTM, a special type of RNN) to extract high-level emotional representation which contained its temporal dynamics information. In 2016, Trigeorgis et al. (2016) proposed a convolutional RNN (CRNN) network which used the raw speech data to predict emotional changes.

Although DNN has achieved great success in SER, there are still some problems. First, the speech signal is quite different due to the variance of speaker's style, content, and environment. Second, DNN learned high-level feature representations from Low-Level Descriptors (LLDs) which cannot sufficiently extract emotional features. Then researchers began to use spectrograms to represent speech signals. The horizontal axis of spectrogram represents the information in time domain and the vertical axis represents the frequency information, making it a decent speech representation that retains the important emotional features of speech. Then CNN is used to automatically extract emotional features from spectrograms which has achieved superior performance in the field of SER.

In 2017, Badshah et al. (2017) used spectrograms and DCNN model to extract features related to speech emotion. They demonstrated the effectiveness of the method and achieved a good result of 84.3% on Berlin Emo-DB. Zhang et al. (2017) proposed a new method which directly to use three channels of log Mel-spectrograms as the pre-trained DCNN's input. Then, they used pyramid matching algorithm (DTPM) to normalize the segment-level features with unequal length. They verified the effectiveness of pre-trained DCNN model with 3-D log Mels on four speech databases. In 2018, Zheng et al. (2018) proposed a new SER model combine with convolutional neural network (CNN) and random forest (RF). They adopted CNN to extract the emotional features from spectrograms, and then used RF for classification. The satisfactory results proved that their model was robust and reasonable.

While spectrogram can retain emotional features well, there is an important and common problem in the above researches that the emotion labels of segments after speech segmentation are marked at the utterance-level. However, not all segments in an utterance contain emotional feature, such as silent frames and emotion irrelevant frames. Therefore, it is important to reduce the influence of these irrelevant segments. Attention mechanism can increase relatively high weights to emotion-related features, emphasizing the importance of these features, and reduce the influence of irrelevant features. It can help the network automatically focus on the emotion relevant segments and obtain discriminative features with utterance-level for SER.

Attention mechanism is adapted for speech emotion recognition work well (Mirsamadi et al., 2017). Zhao et al. (2018) proposed a new method combining Fully Convolutional Networks (FCNs) and attention-based RNNs for speech emotion recognition. The experimental results showed the high performance of the proposed method in IEMOCAP (Busso et al., 2008) and CHEAVD (Li et al., 2017) dataset. Mu et al. (2017) used distributed convolutional neural network (CNN) to automatically learn the emotion features from the raw speech spectrum, and they used bidirectional BRNN to obtain the time information from the CNN output. Finally, the output sequence of BRNN was weighted by attention mechanism algorithm to focus on the useful part of emotion. The weighted accuracy (WA) and unweighted accuracy (UA) of 64.08 and 56.41% were obtained from the IEMOCAP dataset, respectively. Lee et al. (2018) proposed a model combining the convolutional neural network with the attention mechanism and the text data. The promising experimental result in the CMU-MOSEI database proved the effectiveness of the combination of the two modalities.

Inspired by Zhang et al. (2017) and Zhao et al. (2018), in this paper, we propose a novel method based on DCNN and Bidirectional Long Short-Term Memory with attention model (DCNN-BLSTMwA). As illustrated in **Figure 1**, we first conduct data enhancement operation by adjusting different speech playing speed on the original speech data and use balancing datasets weight method to solve the problem of unbalanced emotion data distribution. Secondly, log Mel-Spectrograms (static, delta, delta-delta) of three channels are extracted as the DCNN input. And we initialize parameters by using pre-trained

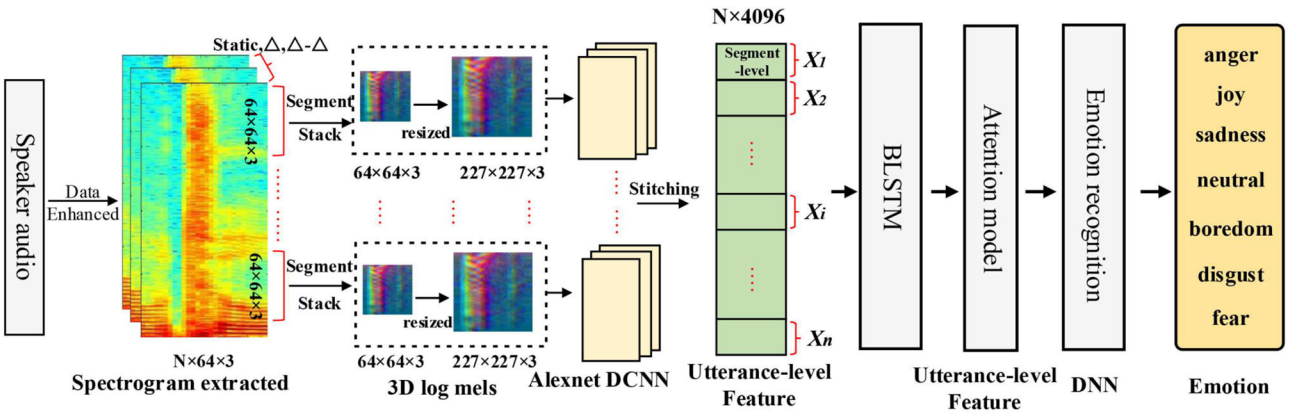

FIGURE 1 | The overview of our proposed method using ADCNN-BLSTM model for SER.

model on ImageNet dataset. Then, we fine-tune the DCNN with our speech data to extract segment-level features and all the segment-level features of a sentence are combined into an utterance-level feature as the input of BLSTM-Attention model. Next, the BLSTM further captures time-frequency relationship of utterance-level features, and the attention model is used to make the emotion features more prominent. After attention layer, we have extracted high-level utterance-level features. Finally, we adopt fully-connected DNN classifier for emotion classification. Abundant experiments on the Berlin Emotional database (EMO-DB) and the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) demonstrate the stable and robust performance of our propose method. The main contributions of our paper can be summarized as follows:

(1) To solve the problem of the small number of training samples for DCNN network training, we use data enhancement and speech segmentation to expand the number of samples. Firstly, we propose a data enhancement method based on overlapping window segmentation, which is not tried in the current DCNN method based on spectrogram and pre training. Secondly, for the preprocessing of overlapping window segmentation, we use BLSTM to enhance the time dimension correlation of DCNN speech data, and add attention mechanism to improve the speech segment Feature extraction, which has not been tried by the existing methods combining attention mechanism. Besides, we prove that the pre-trained DCNN model can reduce the influence of small sample to train deep network and improve the accuracy of speech emotion recognition.
(2) We demonstrate that the three channels of log Mel-spectrograms (3-D log-Mels) as DCNN input is suitable for affective feature extraction which achieves better performance than LLDs. It is natural and will not lose the important emotional features. Besides, we investigate the effects of different number of channels in Mel-spectrograms.
(3) To solve the impact of silent frames and emotion irrelevant frames, an additional attention model is adopted to automatically focus on emotion relevant information. The propose DCNN-BLSTMwA model produce discriminative utterance-level features and the experimental results manifest that this method outperforms the baseline (DNN+ELM) by 16.30% for EMO-DB and 17.26% for IEMOCAP respectively.

The rest of this paper is distributed as follows. Section 2 describes our method and the structure of DCNN-BLSTMwA. The experimental process and the details of the parameter setting are reviewed in section 3. Section 4 analyzes and describes the experimental results. Conclusions are provided in section 5, followed by the future work.

2. PROPOSED METHODOLOGY

In this section, we introduce our new method DCNN-BLSTMwA for speech emotion recognition. Firstly, speech samples need to be preprocessed to reduce individual differences. Secondly, we generate the input of DCNNs from the speech signals, the three-channel log Mel-spectrograms (static, deltas, and delta-deltas). And we describe the process of pre-trained and fine-tuning. Then we introduce the structure of DCNN-BLSTMwA which is used to extract emotion features at utterance-level. Finally, we use a three layers fully-connected DNN modal for emotion classification by utterance-level features, see section 2.5 for more details of the DNN.

2.1. Preprocessing

The speech emotion database is usually composed of multiple speakers (Neumann and Vu, 2017), whose speech exist differences and variations due to age, gender, cultural etc. Therefore, it is necessary to do speech preprocessing before extracting emotion features. Firstly, zero mean and unit variance are calculated for speech standardization and reducing the impact of individual differences problem. Then, we enhance the speech data according to different speech speed and sampling frequency due to DCNN's training requires a large amount of labeled data and data enhancement can make up for small

data samples. Changing the original speed of the speech to a certain extent may change the emotion in the speech, such as speeding it up by 1.5 times or even 2 times. However, we controlled the change of voice speed within the interval of 0.8–1.2, and performed a manual secondary check on the data set after data enhancement, which is equivalent to manually labeling the speech data, proving that this will not change the sample label. Finally, we use data balancing method to make the training data be balanced relatively. Because the number of samples in each class of database is different, there exist the phenomenon of data imbalance influencing the DCNN's training effect. More detail process of data enhancement and balancing datasets will be introduced in section 3.2.

2.2. Log Mel-Spectrograms

In recent years, CNN has showed excellent performance in speech emotion recognition (Kim et al., 2017; Weißkirchen et al., 2017). Different from the CRNN model proposed in Trigeorgis et al. (2016), the input of DCNN model is fixed, that is, it should be appropriately calculated from 1-D speech signals. Abdel-Hamid et al. (2014) used the extracted log Mel-spectrograms and organized it into a 2-D array as the CNN input. Chan and Lane (2015) found that 2-D convolution is superior to 1-D convolution in the case of limited data. Motivated by research (Zheng et al., 2018), we use three-channel log Mel-spectrograms as DCNN input. The process of generating three-channel log Mel-spectrograms as follows. Firstly, a pre-emphasis is performed on the speech data to amplify the high frequency part. Then, the hamming window of 25 ms is used to divide it into smaller frames for each speech sentence, the shift is 10 ms. After that, STFT is used to generate the whole log Mel-spectrogram of an utterance. In this paper, we adopt 64 Mel-filter banks from 20 to 8,000 Hz. Then, a context window of 64 frames is adopted to extract the static Mel-spectrogram. The frame shift size of 32 frames is used to generate overlapping segments of Mel spectrogram. The overlapping segments is the key of speech segmentation. As a result, the static Mel-spectrogram is obtained at the size of 64×64. The first 64 represent the number of Mel-filter banks and the other is represent 64 frames of segment window. The length of a segment is 64 frames, that is 655 ms (10 ms × 63 + 25 ms). Some studies have proved that over 250ms can express an emotion enough (Wöllmer et al., 2013), so the segmentation length of this paper is reasonable. In section 3.3, we will compare the relationship between context window size and the recognition accuracy to find the best effects of the segment length. After generating static Mel-spectrogram, the first and second temporal derivatives are employed to obtain other two channers of Mel-spectrograms. We calculated the first and second order regression coefficients along the timeline as the delta and delta-delta coefficients of the Mel-spectrograms. As shown by (1), m_i are log-Mels, and P_i are outputs. m_i^d are the deltas features of the log-Mels, we use the formula is given by (2). A popular choice for N is 2. And the delta-deltas features $m^d d_i$ are calculated by taking the time derivative of the deltas, as shown in (3). In this way, three-channel of Mel-spectrograms are extracted. The three-channel of Mel-spectrograms as the DCNN input can be expressed as X, $X \in R^{F\times T\times C}$ where F is the number of Mel-filter banks represent the frequency dimension, T is the segment length parallel with the frame number in a segment window, and C is the number of channels. In this paper, we generate three channers of Mel-spectrogram, so the C is 3. The size of X is 64 × 64×3, it is similar with a colorful image.

$$m_i = \log(P_i) \tag{1}$$

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

$$m^d d_i = \frac{\sum_{n=1}^{N} n(m_{i+n}^d - m_{i-n}^d)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

2.3. Pre-training and Finetuning

In this section, we introduce pre-training and finetuning technology. The technology of initialize parameters with pre-trained model is transfer learning which is wildly used in field of image classification (Krizhevsky et al., 2012) and speech recognition (Dahl et al., 2011). Although speech task and image task are two different fields, they get the same network input after preprocessing. The input of DCNN in our model is the spectrum map, which is also a picture. And we proved that the transfer learning is effective through experiments as shown in **Table 4**. As the number of samples in database is relatively small and with the network become deeper, small-scale samples could easy to cause overfitting. And it can also easy to fall into local solution (Yanai and Kawano, 2015). Firstly, the pre-trained model adopted the initialization weight parameters trained by natural scene image database (ImageNet) of 1,000 classes on the annual competition which is now known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) whose recognition accuracy is higher than 95%. Then we finetune the weight by using our speech emotion database. This process can accelerate the speed of network convergence and better fitting the network with small number of samples. More details about fine-tuning of pre-trained DCNN are given in section 3.2.

2.4. Architecture of DCNN-BLSTMwA

In this section, we introduce the architecture of DCNN-BLSTMwA model to analyze 3-D log Mel-spectrograms for SER. First, the deep emotion features are extracted from the 3-D log spectrograms by DCNN. Next, we stack the segment-level 3-D DCNN sequence features of a sentence into the utterance-level features. Then we input these utterance-level features into the Bi-directional LSTM (long short-term memory) to extract higher level features with the long-time information. In this way, the high-level features in the two dimensions are obtained. After BLSTM, an attention layer is devoted to highlighting emotion features and reducing the distractions of unrelated segments. Finally, DNN model is adopted to classify the utterance-level features for SER.

2.4.1. DCNN Model

As shown in **Figure 2**, in this paper we use the classic Alexnet network (Abdel-Hamid et al., 2014) as the DCNN model. The

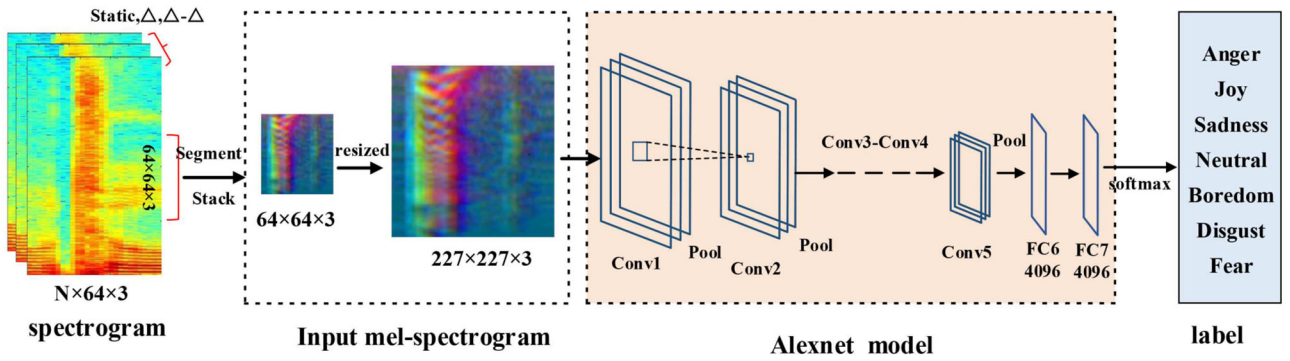

FIGURE 2 | The structure of the DCNN model and the training and fine-tuning process.

DCNN model includes five convolution layers, three max-pooling layers, and two fully connected layers. The size of the convolution kernel of the first layer is 11 × 11 × 96, and the step size is 4 × 4. After the convolution layer of the c1, c2, and c5, there is a max-pooling layer. The pooling size of all pooling layers is 3 × 3, and the step size is 2 × 2. The size of the second convolution kernel is 5 × 5 × 256, while the second and third convolution layer is 3 × 3 × 384. The last convolution kernel of c5 is 3 × 3 × 256. The step size of all convolution layers is 1 × 1. The fully connected layer contains 4,096 linear units, and the output is the segment-level emotional features of the 4,096-dimensional. And the activation function we use Relu. After the last fully connected layer, a dropout layer is followed to minimize the influence of overfitting. Because the input of Alexnet is the fixed size of 227 × 227 pixels, the Mel-spectrograms is 64 × 64 × 3 obtained in section 2.2 this paper. So, we need to reshape the size of DCNN input into 227 × 227 × 3. In this paper, we adopt linear interpolation method to modify the Mel-spectrograms the size, and then input them into DCNN model. More details about parameters setting and fine-tuning of pre-trained DCNN are given in section 3.2.

2.4.2. BLSTM Layer

After extracting and segment-level emotion features with the DCNN model, we stack the segment-level feature sequences into utterance-level with the same length of a sentence. Then we input utterance-level features into BLSTM to extract higher level features for temporal summarization. The structure of combining DCNN and BLSTM can have better performance because DCNN can extract spectral features and BLSTM can extract temporal features from log Mel-spectrograms. These two parts of features are complementary features. Each direction of BLSTM contains 128 cells, after BLSTM we get the 256-dimensional high-level feature representation. We define it as Y, $Y = \{y_1, y_2, \ldots, y_i, \ldots, y_n\}$, where y_i is a feature representation, t is the dimension of BLSTM.

2.4.3. Attention Layer

Attention layer: In a speech, not all the segments are related to emotion such as silent frame and pause segments. These irrelevant features will affect the training and final recognition performance. Attention mechanism can reduce the influence of this problem. Initially, the attention mechanism was applied to image recognition and machine translation. When mimicking human to listen a speech, people often focus on certain strong tones which is more contribute to emotion expression. Therefore, in this paper, attention layer is adopted to focus on emotion features, and weaken the irrelevant ones (e.g., silent frame). Rather than simple operations like max-pooling or average-pooling, attention layer can help to produce the discriminative utterance-level feature representation for final speech emotion classification.

$$\alpha_i = \frac{exp(\mu^T y_i)}{\sum_{j=1}^{J} exp(\mu^T y_j)} \tag{4}$$

$$Z = \sum_{i=1}^{I} \alpha_i y_i \tag{5}$$

As shown in **Figure 3**, the output of the bidirectional LSTM is Y. First, we calculate attention weight α_i. α_i is obtained from a softmax function as the Equation (4). In evaluate (4), the weight μ is obtained by the process of training. Then, we calculate the utterance-level feature representations Z, where Z is got by performing a weighted sum on Y. As shown in Equation (5), we finally produce the higher utterance-level feature Z for SER.

2.5. DNN Classification

In this section, we introduce the architecture of classification for SER. We gain the final utterance-level features (Z), DNN is used to emotion classification. First, DNN classification model was constructed, and then the utterance-level features were used for DNN training. Finally, we produce the output of DNN as the result of SER. DNN has three layers, the first layer contains 512 cells with a dropout layer. And the second and third layer contains 256 cells, the output is one of the 7 classes emotion. In this way, we achieve the speech emotion recognition by proposed method.

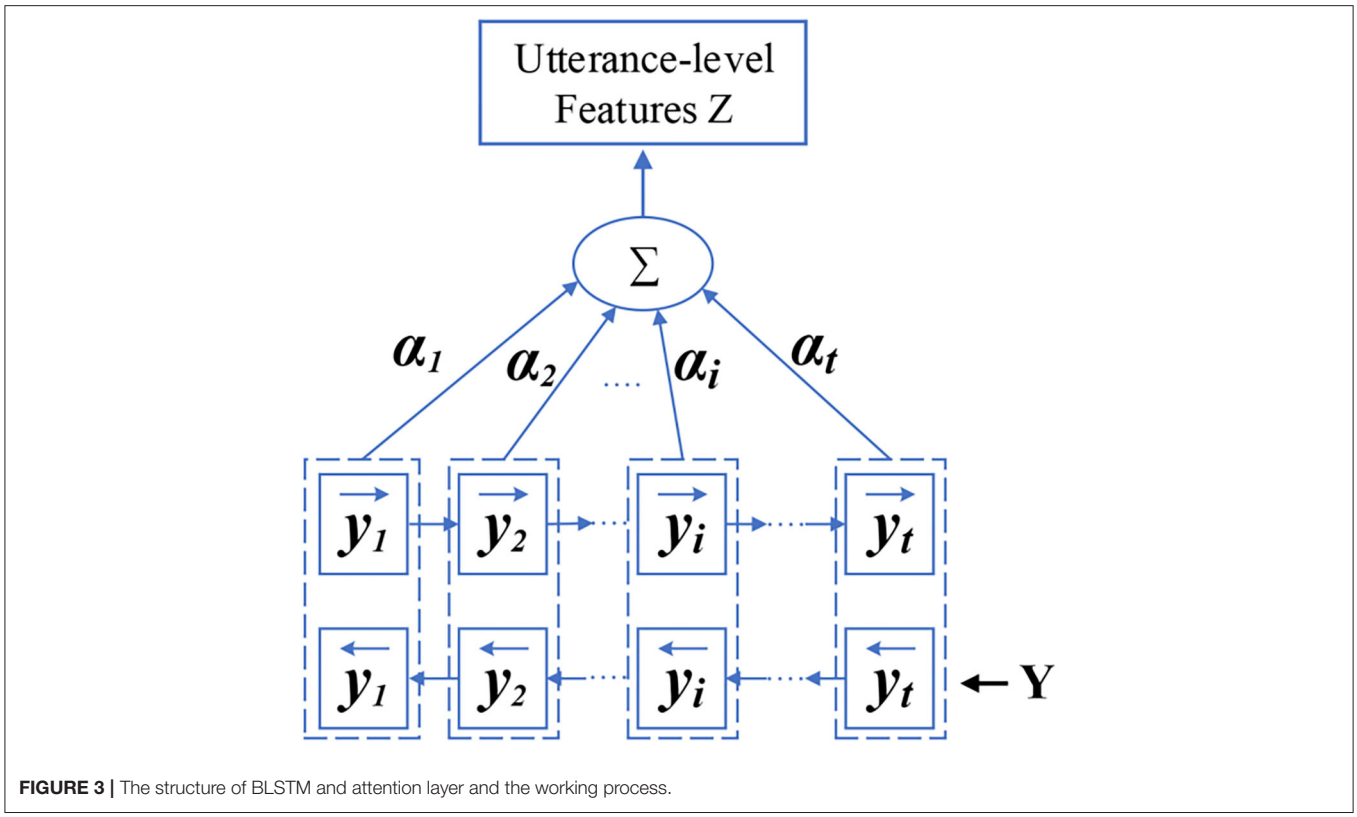

FIGURE 3 | The structure of BLSTM and attention layer and the working process.

3. EXPERIMENTS

3.1. Datasets

In this paper, in order to prove and evaluate the performance of our propose method, we perform abundant speech emotion recognition experiments on Berlin Emotional database (EMO-DB) (Burkhardt et al., 2005) and the Interactive. Emotional Dyadic Motion Capture database (IEMOCAP) (Busso et al., 2008). EMO-DB corpus contains 535 emotional utterance, including seven different emotions: anger, joy, sadness, neutral, boredom, disgust, and fear. The process of this database is that ten professional native German-speaking actors (five men and five women) is asked to imitate the six or seven emotions, and utter 10 sentences in the tone of this emotion. The 10 sentences are five long sentences and five short sentences respectively, which are commonly used in daily communication. The recordings of this database were performed in an extremely quiet room with high-quality equipment at a sampling rate of 16 kHz, 16-bit resolution, and mono channel. The average length of speech file is about 3 s. Twenty participants are required to score the labels, and assess the quality of collected the recordings. IEMOCAP corpus totally contains 10,039 utterances and consists of five sessions, each of which collected the recordings from a pair of actors in scripted and improvised scene (one male and one female). Each utterance is labeled by 3 annotators. If their marks are inconsistent with one another, the data is invalid. The average length of speech file is about 4.5 s at the sample rate of 16 kHz. In this paper, we only use the improvised speech data and use utterances at the emotion labels between four emotion categories, i.e., angry, sad, happy, and neutral. Because the improvised data is more natural and help to the task of SER.

3.2. Experiment Setup

There are only 535 speech samples in Berlin emotional database (Burkhardt et al., 2005), and the number of each emotion category is different. And in IEMOCAP database, there are 3,784 speech samples. Although the number of speech samples is enough in IEMOCAP, the number of each emotion category is unbalanced. The problem of the unbalanced data distribution may influence the training effect of DCNN model. The data distribution of EMO-DB and IEMOCAP as shown in **Figure 4**. It is difficult to train the DCNN model in the case of small amount of data and unbalanced data distribution.

To solve the problem of small samples, we employ the method of data enhancement to expand the speech samples. The details of process: according to the sampling frequency and the playing speed of speech, we conduct data enhancement by adjusting the speed at 0.8, 0.9, 1.0, 1.1, and 1.2 times of the raw speech, respectively. The enhanced data will not lose emotional information, so it can not affect the recognition effect. After data enhancement, we obtain 4 times more than raw speech, and obtain 2,675 sentences at last. To some degree, we solved the problem of small amount of data.

To solve the problem of unbalanced data distribution, we calculate the weight of each category at whole database. The details of process: according to the proportion of each category in all samples, we calculate the weight of the category. Then

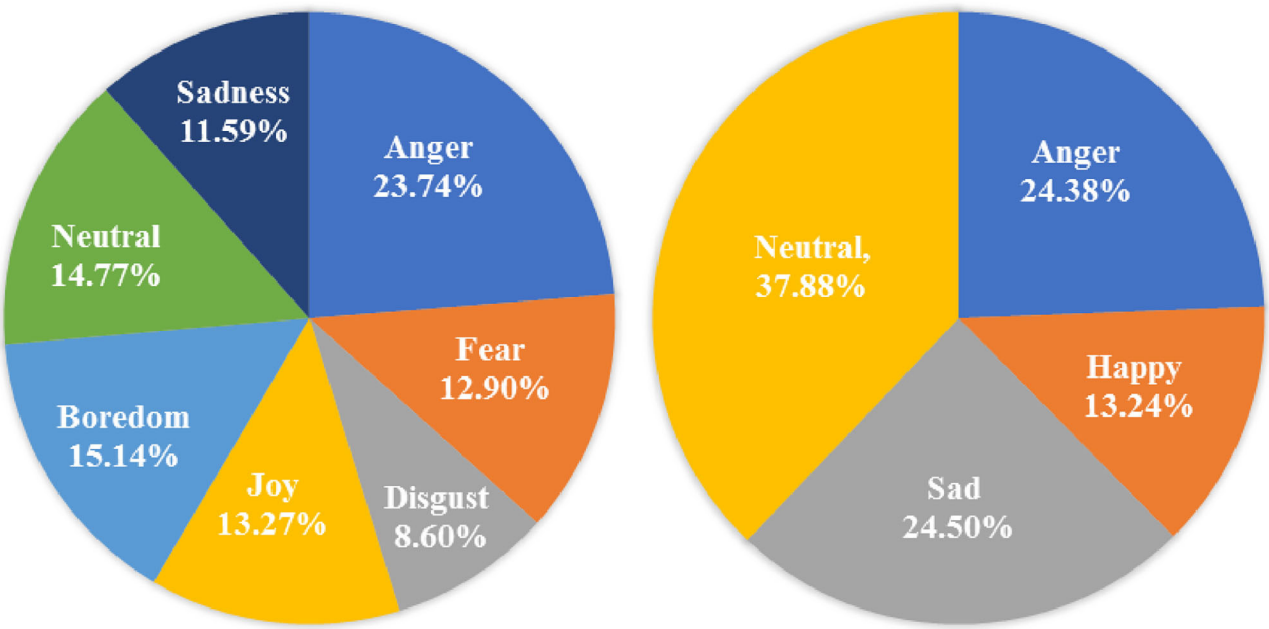

FIGURE 4 | Data distribution of EMO-DB database and IEMOCAP.

TABLE 1 | The unweighted average recall (UAR) (%) of the different number of channers (The value of C) in log Mel-Spectrograms.

The value of C	C=1	C=2	C=3
EMO-DB	80.37 ± 6.17	85.05 ± 8.75	87.86 ± 6.92
IEMOCAP	62.38 ± 4.58	66.25 ± 6.65	68.50 ± 6.20

TABLE 2 | The unweighted average recall (UAR) (%) of different multiples of samples and the effects of data enhancement on EMO-DB database.

Times (speech speed)	1 (1.0)	3 (0.9,1.0,1.1)	5 (0.8,0.9,1.0,1.1,1.2)
Average accuracy	80.92 ± 6.38	84.72 ± 7.76	87.86 ± 6.92

In parentheses is the different speech speed compare with raw samples.

TABLE 3 | The unweighted average recall (UAR)(%) for SER with or without attention model.

Model (architecture)	Without attention (DCNN-BLSTM)	With attention (DCNN-BLSTMwA)
EMO-DB	80.17 ± 6.57	87.86 ± 6.92
IEMOCAP	65.14 ± 4.94	68.50 ± 6.20

TABLE 4 | The unweighted average recall (UAR)(%) for SER with or without pre-training.

Model	Without Pre-training	With pre-training
EMO-DB	81.31 ± 4.89	87.86 ± 6.92
IEMOCAP	64.04 ± 5.24	68.50 ± 6.20

in the process of DCNN training, network parameters are adjusted according to the weight of each category. For example, the number of Disgust is minimal, it's weight will be largest. And in the process of training, the influence of this class on network parameters will increase accordingly. The smaller the number of categories, the greater the weight and the greater the impact on parameters. Indirectly, the problem of unbalanced data distribution can be eased.

In this way, we reduce the impact of these two issues after data enhancing and balancing weight. Next, we introduce the details of speech segmentation and DCNN training.

The length of each utterance in database is different, in order to better fine-tuning of the DCNN network, we split each speech into equal-length segments. The length of each segment is set as 3 s. If the segment is larger than 3 s, we cut off the redundant part. Otherwise, zeros padding is used for smaller than 3 s. In this paper, we use the PyTorch framework to implement our method. We use the *librosa* — *toolkit* to extract 3-D log Mel-spectrograms (static, delta, and delta-delta). Then we stack the three-channer log Mel-spectrograms, and speech segmentation method is used to obtain segment-level speech data as described in section 2.2. The label of each speech segment is consistent with sample label. Since we normalize the sentences at the equal-length of 3 s, each

sentence has the same number of segment-level features, which can be input into DCNN model for training more reasonable.

The training process and details are as follows: First, the network initial parameters are copied from Alexnet training with ImageNet. And then we finetune the DCNN model by using our speech segments, which contain the three-channel log Mel-spectrograms. As shown in **Figure 2**, the size of spectrogram is $227 \times 227 \times 3$ as DCNN input, and a softmax layer is used to predict emotion categories at training. After finetuning the DCNN model, we take the output of its FC7 layer as segment-level emotional feature Xi. After that, all X_i of an utterance are stacked together to form the utterance-level feature X. $X = \{X_1, X_2, \ldots, X_i, \ldots, X_n\}$, where X_i is the segment-level features, and n is the number of segments, and every utterance has the same n. Then we input X into BLSTM-Attention model to further extract higher level features. The parameters of the model are optimized by minimizing the cross-entropy objective function. The batch-size is set to 128, the epoch = 20, and the initial learning rate is set to 10^{-4}, using Adam optimizer with Nestorov momentum, and the momentum was set to 0.9.

The utterance-level features X were input into the BLSTM model to extract the features of temporal information, and the output Y is input into an attention layer to highlight emotional feature. Finally, the utterance-level features were classified by the DNN model with three linear layers. The parameters of DNN are also optimized by minimizing the cross-entropy objective function. The batch-size is set to 16, the epoch is 30, and the initial learning rate is set to 5×10^{-6}, using Adam optimizer with Nestorov momentum, and the momentum was set to 0.9. To get the results more reliable, we performed the "Leave-one-Speaker-out" (LOSO) cross-validation on EMO-DB, eight people are selected as training data, one as validation data, and the last one as test data in each experiment. For IEMOCAP, evaluation is performed in 5-folds, four sessions are selected as training data, one as test data. For each experiment, we test three times and take the average accuracy. Finally, we calculate the unweighted average recall (UAR) of all speakers or sessions as our final experiment results. For each speaker, we test three times and take the average accuracy as the speaker's result. Then, we calculate unweighted average recall (UAR) of all speakers as our final experiment results.

3.3. Experiment Results

(1) Firstly, we tested the effects of the number of channels in log Mel-Spectrograms. We used DCNN-BLSTMwA model to investigate the effects. Specifically, we extracted 3-D log

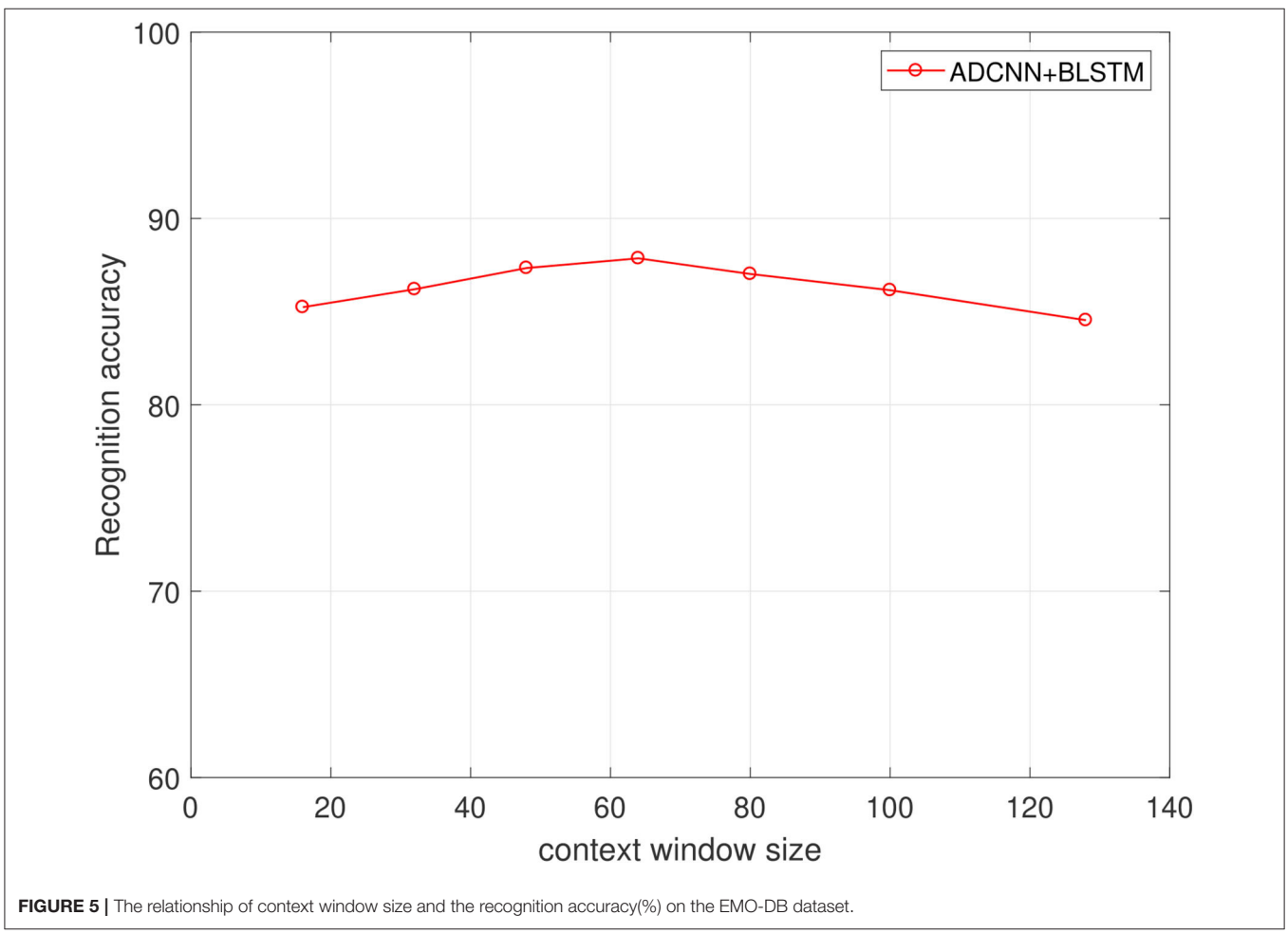

FIGURE 5 | The relationship of context window size and the recognition accuracy(%) on the EMO-DB dataset.

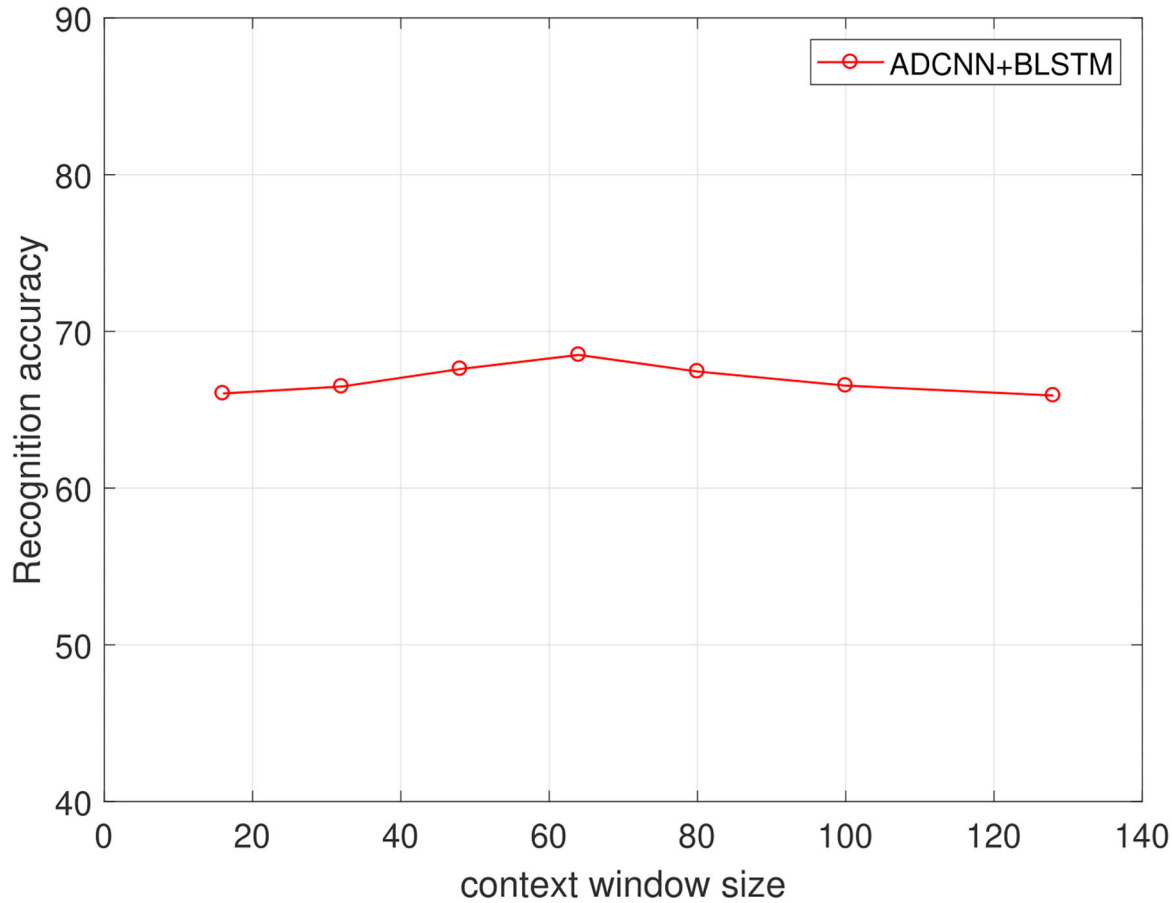

FIGURE 6 | The relationship of context window size and the recognition accuracy(%) on the IEMOCAP dataset.

TABLE 5 | The compare of our proposed approach with several popular methods on EMO-DB.

Author	Method	UAR(%)	Year
K. Han	DNN-ELM (Han et al., 2014)	71.56 ± 8.43	2014
Q. Mao	CNN (Mao et al., 2014)	85.20 ± 0.45	2014
S. Zhang	DCNN+DTPM (Zhang et al., 2017)	87.31 ± 6.95	2017
M. Chen	CRNN+Attention (Mingyi et al., 2018)	82.82 ± 4.99	2018
Baseline	2-CNN-LSTM	78.01 ± 6.91	2019
Proposed	DCNN+LSTM+Attention	87.86 ± 6.92	2019

TABLE 6 | The compare of our proposed approach with several popular methods on IEMOCAP.

Author	Method	UAR(%)	Year
K. Han	DNN-ELM (Han et al., 2014)	51.24 ± 8.24	2014
S. Mirsamadi	RNN+Attention (Mirsamadi et al., 2017)	58.80 ± 4.70	2017
Z. Zhao	Att-BLSTM-FCNs (Zhao et al., 2018)	60.10 ± 4.01	2018
M. Chen	CRNN+Attention (Mingyi et al., 2018)	64.74 ± 5.44	2018
D. Luo	HSF-CRNN (Luo et al., 2018)	63.98 ± 7.56	2018
Baseline	2-CNN-LSTM	58.23 ± 5.21	2019
Proposed	DCNN+LSTM+Attention	68.50 ± 6.20	2019

Mel-Spectrograms (static, delta, and delta-delta) and only adopted 1-D Mel-Spectrograms (static) when C = 1, and 2-D Mel-Spectrograms (static and delta) when C = 2 as DCNN input, respectively. The average accuracy as shown in **Table 1**. When C = 1, average accuracy obtained 80.37% of EMO-DB and 62.38% of IEMOCAP which is better than some traditional methods. This indicates that the high performance of log Mel-Spectrograms. The average accuracy reached 85.05% of C = 2 and 87.86% when C = 3, increasing 4.68 and 7.49%, respectively on EMO-DB. And The average accuracy obtained 66.25% of C = 2 and 68.50% when C = 3, increasing 3.87 and 6.12%, respectively on IEMOCAP. This demonstrated that the first order and second order derivatives of Mel-spectrogram contains helpful emotional information, and the combine of three-channel of Mel-spectrograms can improve the performance for SER.

(2) Secondly, we proved the effects of data enhancement by different multiples of samples with different speed of speech as shown in **Table 2**. Also, we used the proposed model to

test it. We found that the accuracy rate increased by 3.80% when we tripled the samples. And it increased by 6.94% when we expanded the data by 5 times. Because we use the different sampling frequency and the playing speed to enhance the samples, the important information will not lose in a speech. The good results prove that data enhancement can help the training of deep network model and improve the final classification accuracy.

(3) **Table 3** reveals the effects of the attention model. We found that after using the attention model, the average accuracy increased by 7.69% of EMO-DB and 3.36% of IEMOCAP. Abundant experiments proved the powerful performance of attention model, which can focus on important emotional features and help to extract higher utterance-level features and improve the accuracy for SER.

(4) **Table 4** shows the effects of pre-trained DCNN model. We tested the performance of without pre-trained DCNN model by using the same architecture of our proposed method. The initial parameters were randomly initialized with a standard normal distribution. The average accuracy of without pre-training was 81.31% of EMO-DB and 64.04% of IEMOCAP, we improved 6.55 and 4.46%, respectively by using initial parameters from ImageNet for pre-training. This demonstrates that pre-trained DCNN model not only speeds up the network convergence, but also improves the classification accuracy.

(5) Next, in order to find the best effects of the segment length, we designed different segment length to compare the relationship between context window size and the recognition accuracy. we tested context window size ranges

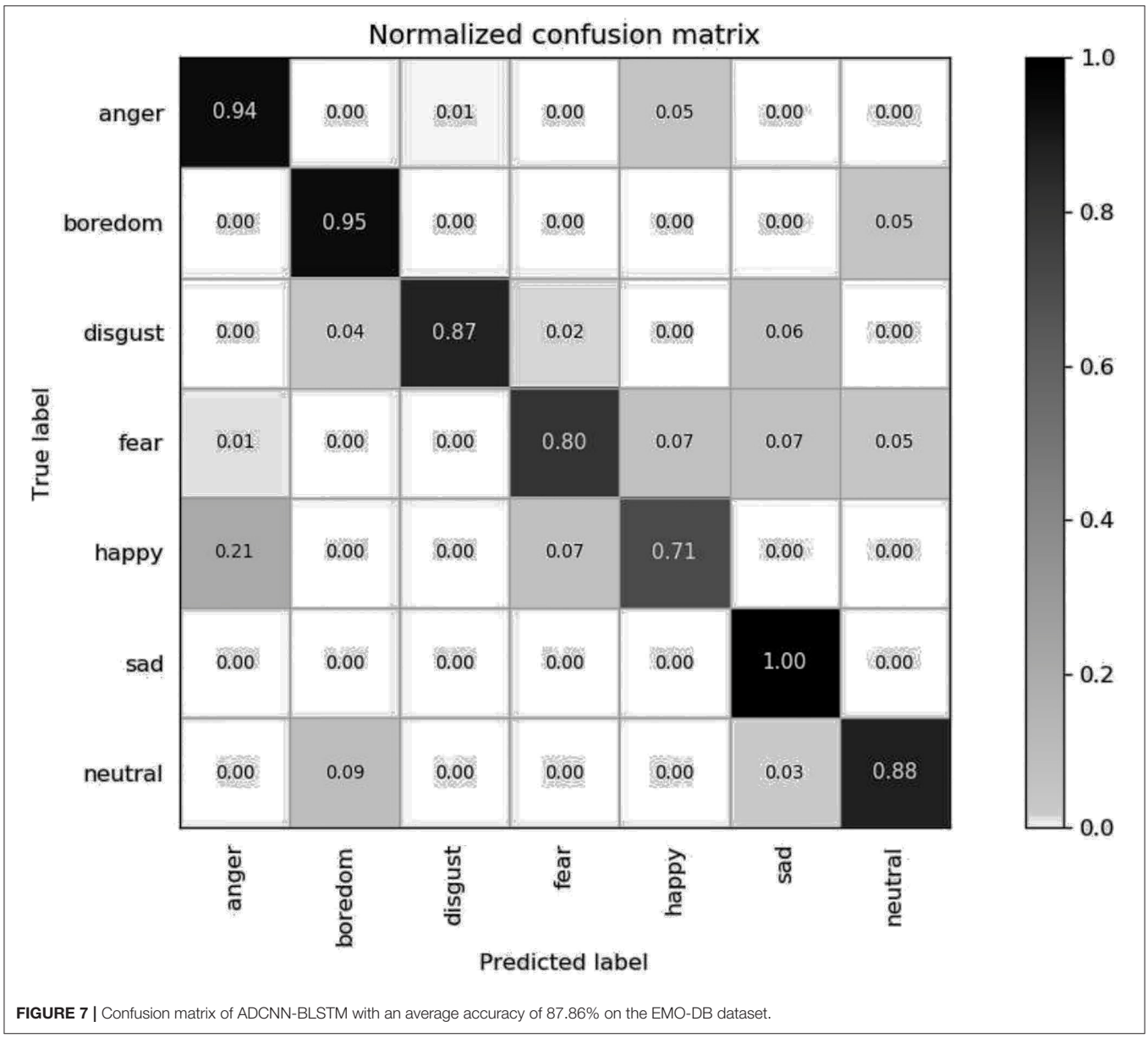

FIGURE 7 | Confusion matrix of ADCNN-BLSTM with an average accuracy of 87.86% on the EMO-DB dataset.

in (16, 32, 48, 64, 80, 100, 128), and the effects as shown in **Figures 5**, **6** of EMO-DB and IEMOCAP database respectively. We found that the context window size of 64 obtained best effects both on these two databases. This demonstrates that the image size of 64 × 64 reshape to 227 × 227 was best size of DCNN's input for training.

(6) Then, we compared our proposed approach with several popular methods as shown in **Tables 5**, **6**. We first built the baseline by using 2 convolution layers follow by LSTM. The first convolution layer contains 16 kernels of size 5 × 5 with the stride size of 1 × 1 and the second 32 kernels of size 5 × 5 with the same stride size of 1 × 1. Each convolution layer followed by a Max-pooling layer. The LSTM contains 512 cells with 0.5 dropout rate. Similarly, we also adopted 3-D log Mel-Spectrograms as input to extract the emotional features. The average accuracy of 2 CNN-LSTM is 78.09% of EMO-DB and 58.23% of IEMOCAP which are better than the method of DNN-ELM. As shown in **Tables 5**, **6**, the average accuracy is 87.86% of EMO-DB and 68.50% of IEMOCAP, and our proposed method is better than most of popular SER methods in recent years. This demonstrates the promising performance of our approach.

(7) Finally, we present the confusion matrix consistent with the results of DCNN-BLSTMwA to further analyze the recognition accuracy as shown in **Figure 7**, where the vertical axis represents the true label and the horizontal axis represents the predicted label. In the confusion matrix of EMO-DB, we find that *sad* achieves the highest recognition rate of 1 while *happy* is the lowest accuracy of 0.71, and the *anger* and *boredom* also obtain pretty good recognition rate

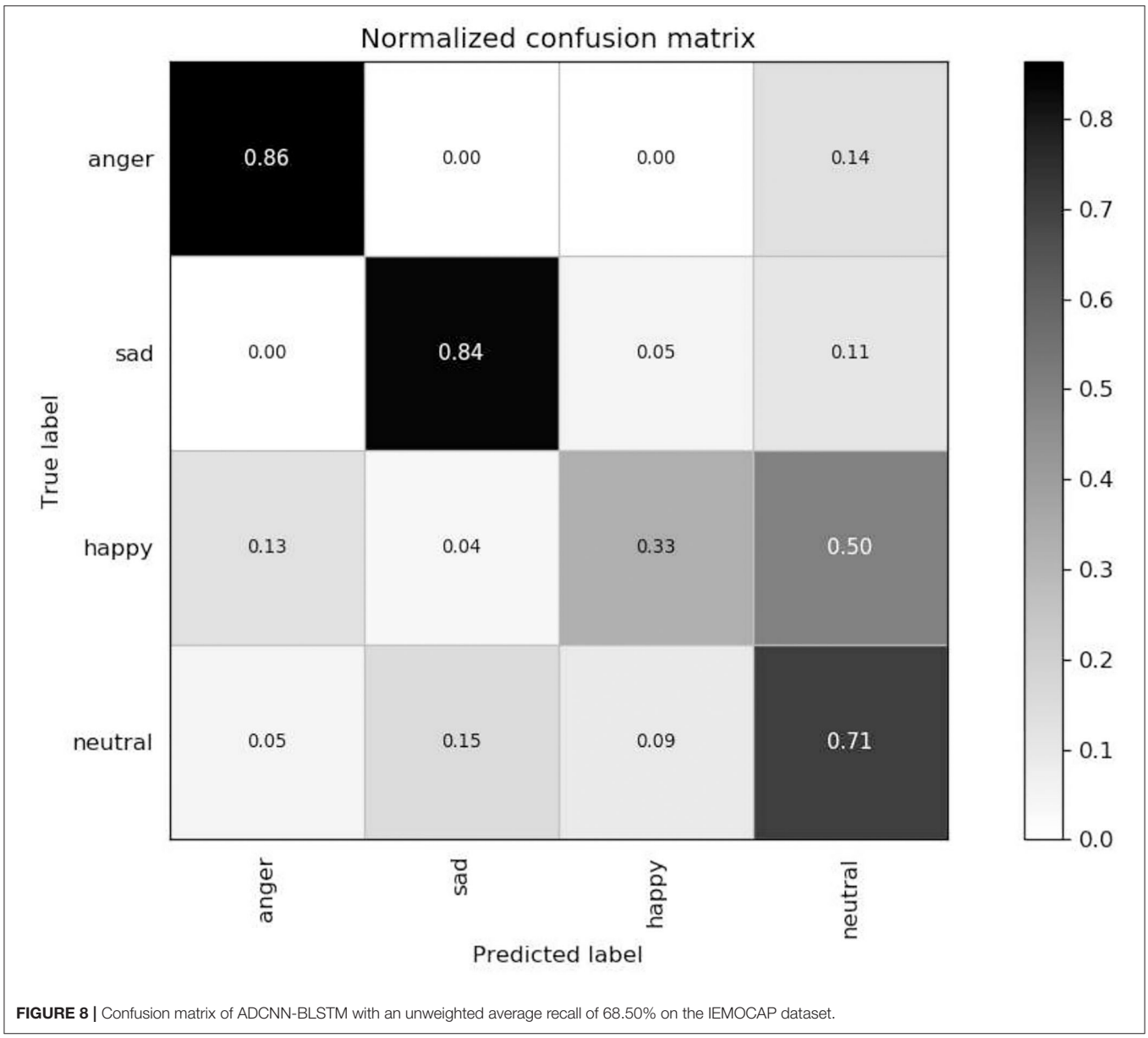

FIGURE 8 | Confusion matrix of ADCNN-BLSTM with an unweighted average recall of 68.50% on the IEMOCAP dataset.

at 0.94 and 0.95, respectively. Similarly, as shown in **Figure 8**, the *happy* is also the lowest accuracy of 0.33 on IEMOCAP database. And *anger* and *sad* achieve relatively good result of 0.86 and 0.84, respectively, *neutral* is 0.71. The reason may be that *sad* and *anger* are relatively intense emotions and highly diacritical among these emotions. Thus, these two emotions obtain better results. And to some extent, *happy* emotion features is a little similar with *anger*, so 21% *happy* samples are misclassified into *anger* on EMO-DB. However, there are 50% *happy* samples are misclassified into *neutral* on IEMOCAP. We attribute the misclassification to *neutral* is at the center of these emotions and the *happy* emotion closing to the center is hard to distinguish. In addition, we find that 9% *neutral* are misclassified into *boredom*, it may be the similar activation level between *neutral* and *boredom*. All in all, the average accuracy of 87.86% of EMO-DB and 68.50% of IEMOCAP are promising results and demonstrate the performance of our methods.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new method based on pre-trained DCNN model and BLSTM with attention model (DCNN-BLSTMwA) for speech emotion recognition. We first enhanced the speech samples and balanced datasets. Then 3-D log Mel-spectrograms (static, delta, delta and delta) were extracted from the speech signal as DCNN input. DCNN extracted the segment-level features which were stacked to obtain the utterance-level features. Then higher utterance-level features were further extracted through BLSTM with an attention model and finally, the DNN model was used for final SER. Experiments on EMO-DB database have shown the promising performance of our proposed method compared with some popular SER methods. The average accuracy in terms of UAR is 87.86% of EMO-DB and 68.50% of IEMOCAP, respectively, which are better than most popular SER methods of recent years. Additionally, we have also proved the robust performance and feasibility of the method.

In the future, we will try to construct a more stable deep neural network to fit more speech signals for SER. And we will combine the LLDs features and DCNN extracted features to realize speech emotion recognition. In addition, we are going to initial the parameters of other speech emotion database rather than ImageNet. It may help promote the training performance of deep neural model and increase the accuracy for SER.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://emodb.bilderbar.info/start.html; https://sail.usc.edu/iemocap/.

AUTHOR CONTRIBUTIONS

HZ and RG designed the core methodology of the study, carried out the implement, and drafted the manuscript. YW carried out the experiments and drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (No. 2017YFE0118200), the National Natural Science Foundation of China (No. 61471150, No. 61802094), the National Natural Science Foundation of Zhejiang Province (No. LQ19F020008, No. LY20F020012). Thanks for support and assistance from Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province. Thanks for support and assistance from Key Laboratory of Network Multimedia Technology of Zhejiang Province.

REFERENCES

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1533–1545. doi: 10.1109/TASLP.2014.2339736

Ayadi, M. E., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recogn.* 44, 572–587. doi: 10.1016/j.patcog.2010.09.020

Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)* (Busan), 1–5. doi: 10.1109/PlatCon.2017.7883728

Bhaykar, M., Yadav, J., and Rao, K. (2013). "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM," in *National Conference on Communications* (New Delhi). doi: 10.1109/NCC.2013.6487998

Burkhardt, F., Paeschke, A., Rolfes, M. A., Sendlmeier, W. F., and Weiss, B. (2005). "A database of German emotional speech," in *Ninth European Conference on Speech Communication and Technology* (Lisbon).

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Chan, W. and Lane, I. (2015). "Deep convolutional neural networks for acoustic modeling in low resource languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD). doi: 10.1109/ICASSP.2015.7178332

Chen, L., Mao, X., Xue, Y., and Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Process.* 22, 1154–1160. doi: 10.1016/j.dsp.2012.05.007

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 20, 30–42. doi: 10.1109/TASL.2011.2134090

Deng, J., Zhang, Z., Florian, E., and Björn, S. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 21, 1068–1072. doi: 10.1109/LSP.2014.2324759

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. (2013). "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Humaine Association Conference on Affective Computing and Intelligent Interaction* (Geneva). doi: 10.1109/ACII.2013.90

Han, K., Yu, D., and Tashev, I. (2014). "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association* (Singapore).

Kim, J., Truong, K., Englebienne, G., and Evers, V. (2017). "Learning spectro-temporal features with 3d cnns for speech emotion recognition," in *Seventh*

International Conference on Affective Computing and Intelligent Interaction (San Antonio, TX). doi: 10.1109/ACII.2017.8273628

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 1097–1105. doi: 10.1145/3065386

Lee, C. W., Song, K. Y., Jeong, J., and Choi, W. Y. (2018). Convolutional attention networks for multimodal emotion recognition from speech and text data. *arXiv [preprint]. arXiv:1805.06606*. doi: 10.18653/v1/W18-3304

Lee, J. and Tashev, I. (2015). "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association* (Dresden), 1537–1540.

Li, Y., Tao, J., Chao, L., Wei, B., and Liu, Y. (2017). Cheavd: a chinese natural emotional audio-visual database. *J. Ambient Intell. Human. Comput.* 8, 913–924. doi: 10.1007/s12652-016-0406-z

Luo, D., Zou, Y., and Huang, D. (2018). Investigation on joint representation learning for robust feature extraction in speech emotion recognition. *Interspeech* 2018, 152–156. doi: 10.21437/Interspeech.2018-1832

Mao, Q., Ming, D., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* 16, 2203–2213. doi: 10.1109/TMM.2014.2360798

Milton, A., Roy, S. S., and Selvi, S. T. (2014). Svm scheme for speech emotion recognition using mfcc feature. *Int. J. Comput. Appl.* 69, 34–39. doi: 10.5120/11872-7667

Mingyi, C., Xuanji, H., Jing, Y., and Han, Z. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* 25, 1440–1444. doi: 10.1109/LSP.2018.2860246

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA), 2227–2231. doi: 10.1109/ICASSP.2017.7952552

Mu, Y., Gómez, L. A. H., Montes, A. C., Martínez, C. A., Wang, X., and Gao, H. (2017). Speech emotion recognition using convolutional-recurrent neural networks with attention model. *DEStech Trans. Comput. Sci Eng.* 15, 341–350. doi: 10.12783/dtcse/cii2017/17273

Neumann, M., and Vu, N. T. (2017). "Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech," in *Proc. Interspeech 2017* (Stockholm), 1263–1267. doi: 10.21437/Interspeech.2017-917

Origlia, A., Galatà, V., and Ludusan, B. (2010). "Automatic classification of emotions via global and local prosodic features on a multilingual emotional database," in *Speech Prosody 2010-Fifth International Conference* (Chicago, IL).

Schuller, B., Rigoll, G., and Lang, M. (2003). "Hidden Markov model-based speech emotion recognition," in *International Conference on Multimedia & Expo* (Baltimore, MD). doi: 10.1109/ICME.2003.1220939

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M., Schuller, B., et al. (2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai). doi: 10.1109/ICASSP.2016.7472669

Weißkirchen, N., Böck, R., and Wendemuth, A. (2017). "Recognition of emotional speech with convolutional neural networks by means of spectral estimates," in *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos* (San Antonio, TX). doi: 10.1109/ACIIW.2017.8272585

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vis. Comput.* 31, 153–163. doi: 10.1016/j.imavis.2012.03.001

Yanai, K., and Kawano, Y. (2015). "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *IEEE International Conference on Multimedia & Expo Workshops* (Turin). doi: 10.1109/ICMEW.2015.7169816

Zhang, S., Zhang, S., Huang, T., and Wen, G. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* 20, 1576–1590. doi: 10.1109/TMM.2017.2766843

Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., and Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. *Proc. Interspeech 2018,* 272–276. doi: 10.21437/Interspeech.2018-1477

Zheng, L., Li, Q., Ban, H., and Liu, S. (2018). "Speech emotion recognition based on convolution neural network combined with random forest," in *2018 Chinese Control And Decision Conference (CCDC)* (Shenyang), 4143–4147. doi: 10.1109/CCDC.2018.8407844

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 Medicine

ORIGINAL RESEARCH
published: 09 March 2021
doi: 10.3389/fmed.2021.635771

Application of Machine Learning Algorithms to Predict Central Lymph Node Metastasis in T1-T2, Non-invasive, and Clinically Node Negative Papillary Thyroid Carcinoma

Jiang Zhu[1†‡], Jinxin Zheng[2†], Longfei Li[3], Rui Huang[4‡], Haoyu Ren[1,5], Denghui Wang[1], Zhijun Dai[2‡] and Xinliang Su[1*‡]*

[1] Department of Endocrine and Breast Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [2] Department of Breast Surgery, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China, [3] Department of Health Statistics, School of Public Health, Chongqing Medical University, Chongqing, China, [4] Department of Anesthesiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [5] Department of General, Visceral, and Transplant Surgery, Ludwig-Maximilians-University, Munich, Germany

Purpose: While there are no clear indications of whether central lymph node dissection is necessary in patients with T1-T2, non-invasive, clinically uninvolved central neck lymph nodes papillary thyroid carcinoma (PTC), this study seeks to develop and validate models for predicting the risk of central lymph node metastasis (CLNM) in these patients based on machine learning algorithms.

Methods: This is a retrospective study comprising 1,271 patients with T1-T2 stage, non-invasive, and clinically node negative (cN0) PTC who underwent surgery at the Department of Endocrine and Breast Surgery of The First Affiliated Hospital of Chongqing Medical University from February 1, 2016, to December 31, 2018. We applied six machine learning (ML) algorithms, including Logistic Regression (LR), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Neural Network (NNET), coupled with preoperative clinical characteristics and intraoperative information to develop prediction models for CLNM. Among all the samples, 70% were randomly selected to train the models while the remaining 30% were used for validation. Indices like the area under the receiver operating characteristic (AUROC), sensitivity, specificity, and accuracy were calculated to test the models' performance.

Results: The results showed that ∼51.3% (652 out of 1,271) of the patients had pN1 disease. In multivariate logistic regression analyses, gender, tumor size and location, multifocality, age, and Delphian lymph node status were all independent predictors of CLNM. In predicting CLNM, six ML algorithms posted AUROC of 0.70–0.75, with the extreme gradient boosting (XGBoost) model standing out, registering 0.75. Thus, we employed the best-performing ML algorithm model and uploaded the results to a

OPEN ACCESS

Edited by:
Juan Liu,
Huazhong University of Science and Technology, China

Reviewed by:
Hao Zhang,
Dalian Medical University, China
Changming An,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

****Correspondence:***
Xinliang Su
suxinliang@21cn.com
Zhijun Dai
dzj0911@126.com

† These authors have contributed equally to this work

‡ORCID:
Jiang Zhu
orcid.org/0000-0003-4194-2355
Rui Huang
orcid.org/0000-0002-3342-135X
Xinliang Su
orcid.org/0000-0001-5792-1407
Zhijun Dai
orcid.org/0000-0001-5209-8626

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *30 November 2020*
Accepted: *15 February 2021*
Published: *09 March 2021*

Citation:
Zhu J, Zheng J, Li L, Huang R, Ren H, Wang D, Dai Z and Su X (2021) Application of Machine Learning Algorithms to Predict Central Lymph Node Metastasis in T1-T2, Non-invasive, and Clinically Node Negative Papillary Thyroid Carcinoma. Front. Med. 8:635771. doi: 10.3389/fmed.2021.635771

self-made online risk calculator to estimate an individual's probability of CLNM (https://jin63.shinyapps.io/ML_CLNM/).

Conclusions: With the incorporation of preoperative and intraoperative risk factors, ML algorithms can achieve acceptable prediction of CLNM with Xgboost model performing the best. Our online risk calculator based on ML algorithm may help determine the optimal extent of initial surgical treatment for patients with T1-T2 stage, non-invasive, and clinically node negative PTC.

Keywords: papillary thyroid carcinoma, central lymph node metastasis, machine learning algorithms, lymph node dissections, prediction model

INTRODUCTION

Papillary thyroid carcinoma (PTC) is one of the most common type of endocrine malignancies with a favorable prognosis (1, 2). Nevertheless, central lymph node metastasis (CLNM), the first station of metastasis, occurs in 30–90% of patients following their first surgery and is correlated with an increased risk of local recurrence (3, 4).

The clinical community has reached a general consensus that central lymph node dissection (CLND) for therapeutic purposes is appropriate in PTC patients with suspected cervical lymph node metastasis (LNM) (5). By contrast, however, there is a growing controversy over the role of prophylactic central lymph node dissection (pCLND) due to the lack of randomized controlled data (6–8). Generally speaking, pCLND is not recommended for a subset of patients with small (T1 or T2), non-invasive, clinically node-negative (cN0) PTC according to the 2015 American Thyroid Association (ATA) guidelines (9), whereas the Japanese Society of Thyroid Surgery and the Chinese Thyroid Association both strongly recommend routine pCLND for cN0 PTC patients in order to stage disease and prevent recurrence. While an incomplete nodal resection in the first surgery may lead to disease recurrence and a second operation (10), it is also important to avoid unnecessary CLND in view of surgical complications such as hypoparathyroidism and recurrent laryngeal nerve injury. Ideal treatment decision-making should be based upon individual patients rather than "one size fits all" approach recommended by guidelines. This highlights the importance of accurate prediction of CLNM occurrence with a more personalized therapeutical schedule.

Machine learning (ML), as a novel type of artificial intelligence (AI), is starting to be widely applied to health-care data analysis (11, 12). By capitalizing on the robust prediction ability of ML algorithms, it may be possible to develop prediction tools which in some cases outperform traditional statistical modeling, and thus giving better prediction of CLNM status. Unfortunately, no current studies have trained ML algorithms to predict CLNM in this subset of PTC.

Hence, the purpose of this study is to develop ML-based models using preoperative and intraoperative clinicopathological characteristics to predict the likelihood of CLNM for individualized treatment and to obtain the best ML algorithms for online CLNM prediction in PTC.

METHODS

Study Population

We retrospectively retrieved the data of in-patients who underwent thyroid surgery at the Department of Endocrine and Breast Surgery of the First Affiliated Hospital of Chongqing Medical University from December 2016 to December 2018.

Data Collection

Criteria for inclusion were to be a PTC patient with a tumor size no larger than 40 mm (T1-T2), a non-invasive tumor, and no evidence for lymph nodes metastases (cN0) based on ultrasound (US) data. Tumor size was classified according to the 8th edition of American Joint Committee on Cancer (AJCC) Staging Standards. Criteria for exclusion were distant metastasis, previous thyroid surgery, or incomplete information. This study was approved by the local institutional ethics committee board. Demographic and clinicopathological characteristics data were collected as follows: gender, age, tumor size, tumor location, chronic lymphocytic thyroiditis (CLT), multifocality, bilaterality, and the presence of LNM.

Surgical Strategy

At our institution, it is customary to perform pCLND for PTC patients and the detailed surgical procedures were described in previous articles (13, 14). Soft tissues in the prelaryngeal and pretracheal regions were removed and marked as the Delphian (**Figure 1**) and pretracheal LNs, respectively. Those two subgroups were sent for intraoperative frozen section examination. Then, we proceeded to perform the thyroid lobectomy and ipsilateral paratracheal LN dissection and the paratracheal LN was also sent for frozen section examination. Lastly, all surgical specimens were sent for post-operative histopathologic evaluation. The Delphian lymph node (DLN) was not taken into account in the calculation of central compartment lymph nodes.

Statistical Analyses

The Fisher's exact test and Student's *t*-test were used for discrete and continuous parameters, respectively. For the independent risk factors of CLNM, a multivariable logistic regression analysis

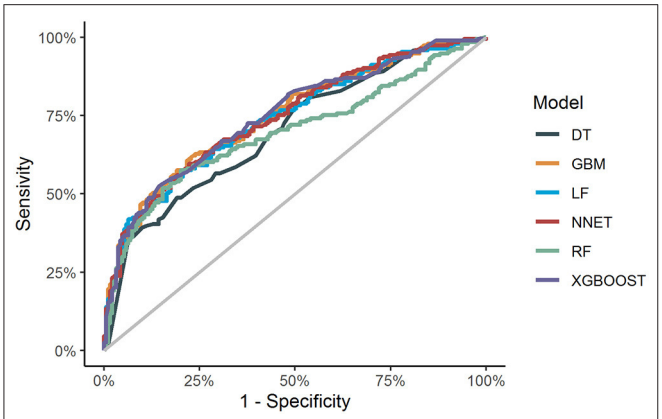

FIGURE 1 | ROC curve analysis of machine learning algorithms for prediction of CLNM patients with T1-T2 stage, non-invasive, and clinically node negative PTC in the validation set. LR, Logistic regression; GBM, Gradient boosting machine; RF, Random forest; DT, Decision tree; NNET, Neural network; Xgboost, Extreme gradient boosting; ROC, receiver operating characteristic; AUC, area under the curve.

with backward stepwise selection was used to calculate the odds ratios (ORs) with 95% confidence intervals (CIs).

ML algorithm is characterized by its extraordinary performance better than traditional regression approaches in predicting outcomes within large data bases (15–17). In this study, we randomly split our dataset into two groups, namely the training sets (70%) for ML model development and the validation sets (30%) for performance evaluation and we repeated this random splitting until the patient data were equally distributed in both sets (**Supplementary Table 1**). We developed six types of ML algorithms to model our data: Logistic regression (LR), Gradient boosting machine (GBM), Extreme gradient boosting (XGBoost), Random forest (RF), Decision tree (DT), and Neural network (NNET). In the training process, tuning was considered for ML-based models to avoid overfitting and the best hyper-parameter for ML models was 5-fold cross-validation. Then the ML algorithms were further trained by using the R software to predict the risk of CLNM and we evaluated the predictive ability of each ML classifier, with the same hyper-parameter, in validation sets where the area under the receiver operating characteristic (AUROC) value, and the corresponding sensitivity, specificity, as well as overall accuracy of ML algorithms were all calculated. In the comparison of ML algorithms' performance, the closer to 1 the AUC was, the better the classification model performed. Afterwards, based on the best-performing model, we created an online risk calculator that can make predictions with newly entered PTC patient data, and thus making the risk of CLNM in those patients easily accessible to clinicians. A total of 100 independent training simulation results were used to evaluate the variable importance of each CLNM-predicting ML model. All statistical analyses were performed by using R software, version 3.4.1 (R Foundation for Statistical Computing, Vienna, Austria). The R packages "caret," "e1071," "random-forest," "nnet," "gbm," "rpart," "GLM," "pROC" were used for ML algorithms and "shiny" package for web application. A two tailed $P < 0.05$ was deemed statistically significant.

RESULTS

Demographics Features

The clinicopathological characteristics of 1,271 PTC patients with T1-T2, non-invasive, clinically node-negative disease were summarized (**Table 1**). Of the 1,271 eligible patients, the average age was 42.15 ± 10.49 years (range 18–80 years). The ratio of male to female patients was 1:2.7. The mean tumor size was 9.92 mm (median = 8 mm). Eight hundred and ninety seven patients (70.6%) had papillary micro-carcinomas. Central lymph node metastases were positive in 652 (51.3%) cases.

Univariate and Multivariate Logistic Regression Analyses of CLNM

In univariable analysis, tumor size, gender, age, multifocality, bilateral lesions, and DLN status were all significantly associated with the occurrence of CLNM in overall population ($P < 0.001$), whereas there was no significant difference between CLNM-positive and CLNM-negative patients in terms of their tumor location or CLT status. In multivariable logistic regression analysis (**Table 2**), all parameters (age, gender, CLT, DLN, multifocality, bilaterality and tumor size, and location) were included. The results showed that male gender (OR 1.534, 95% CI 1.158–2.030), larger tumor size (OR 1.080, 95% CI 1.053–1.107), multifocality (OR 1.583, 95% CI 1.172–2.139), DLN metastasis (OR 6.454, 95% CI 4.246–9.651), and tumor located in inferior pole [vs. upper pole, (OR 1.507, 95% CI 1.080–2.103)] are independent positive predictors of CLNM while older age (OR 0.975, 95% CI 0.964–0.986) was a negative predictor. Variables of bilateral lesions and CLT were rejected by multivariable analysis.

Performance of Machine Learning Algorithms

Comparisons of the performance of prediction among the six ML algorithms models in validation sets are detailed in **Table 3** and **Figure 1**. It turned out that the XGBoost model demonstrated the highest performance of predicting CLNM, whose AUROC was 0.750, sensitivity 0.667, specificity 0.674, and accuracy 0.670 in validation sets. Accordingly, we chose the XGBoost model as the final prediction model.

Relative Importance of Variables in Machine Learning Algorithms

The relative importance of variables in each CLNM-predicting ML algorithm is shown in **Figure 2**. We can see there are general trends of evidence: although slight differences are shown in the importance of variables among those ML algorithms, factors including Delphian lymph node metastasis, tumor size, age, gender, multifocality rank top five without fail. On the contrary, variables like bilateral lesions, tumor location in middle or isthmus pole and CLT make little contribution to CLNM prediction. The importance of high-ranking variables in the XGBoost model is arranged as follows in a descending

TABLE 1 | Demographic and clinicopathologic variables of the whole cohort grouped by lymph node status.

Charteristics	Total (*N* = 1,271) No (%)	CLNM- (*N* = 619)	CLNM+(*N* = 652)	*P*-value
Gender				<0.001
Male	339 (26.67)	132 (38.94)	207 (61.06)	
Female	932 (73.33)	487 (52.25)	445 (47.75)	
Age (years)	41.38 ± 11.09	43.18 ± 11.39	39.68 ± 10.51	<0.001
≤55	1,140 (89.69)	534 (46.84)	606 (53.16)	<0.001
>55	131 (10.31)	85 (64.89)	46 (35.11)	
Tumor size (mm)	9.92 ± 5.69	8.53 ± 4.27	11.24 ± 6.34	<0.001
≤10 mm	897 (70.57)	491(54.74)	406 (45.26)	<0.001
10–20 mm	305 (24.00)	115 (37.70)	190 (62.30)	
>20 mm	69 (5.43)	13 (18.84)	56 (81.16)	
Bilateral				<0.001
No	1,071 (84.3)	544 (50.79)	527 (49.21)	
Yes	200 (15.7)	75 (37.50)	125 (62.50)	
Tumor location				0.127
Upper	304 (23.92)	152 (50.00)	152 (50.00)	
Middle	545 (42.88)	280 (51.38)	265 (48.62)	
Inferior	380 (29.90)	171 (45.00)	209 (55.00)	
Isthmus	42 (3.30)	16 (38.10)	26 (61.90)	
Multifocality				<0.001
Absence	1,002 (78.77)	514 (51.30)	488 (48.70)	
Presence	269 (21.23)	105 (39.03)	164 (60.97)	
CLT				0.573
No	988 (77.73)	477 (48.28)	511 (51.72)	
Yes	283 (22.27)	142 (50.18)	141 (49.82)	
DLN status				<0.001
Negative	1,051 (82.69)	589 (56.04)	462 (43.96)	
Positive	220 (17.31)	30 (13.64)	190 (86.36)	

Continuous data are shown as mean ± standard deviation.
–, negative; +, positive; CLNM, central lymph node metastasis; CLT, chronic lymphocytic thyroiditis; DLN, Delphian lymph node.

TABLE 2 | Univariate and multivariate logistic regression analysis of variables in predicting CLNM in whole cohort.

Variables	Univariate analysis		Multivariate analysis	
	OR (95%CI)	*P*	OR (95%CI)	*P*
Multifocality (+/–)	1.645 (1.250–2.165)	<0.001	1.583 (1.172–2.139)	0.003
Age	0.971 (0.961–0.981)	<0.001	0.975 (0.964–0.986)	<0.001
Gender (Male/Female)	1.716 (1.332–2.221)	<0.001	1.534 (1.158–2.030)	0.003
DLN status (+/–)	8.074 (5.392–12.092)	<0.001	6.454 (4.246–9.651)	<0.001
Tumor size (mm)	1.103 (1.077–1.130)	<0.001	1.080 (1.053–1.107)	<0.001
Tumor location		0.127		0.043
Upper	Reference		Reference	
Middle	0.946 (0.715–1.253)	0.701	1.059 (0.887–1.447)	0.719
Inferior	1.222 (0.903–1.654)	0.193	1.507 (1.080–2.103)	0.016
Isthmus	1.625 (0.838–3.151)	0.151	1.445 (0.692–3.018)	0.327
Bilateral (+/–)	1.720 (1.261–2.346)	0.001		
CLT (+/–)	0.927 (0.712–1.207)	0.573		

DLN, Delphian lymph node; CLT, chronic lymphocytic thyroiditis; CLNM, central lymph node metastasis; –, negative; +, positive.

TABLE 3 | Predictive performance comparison of the six types of machine learning algorithms in the validation sets.

Methods	AUROC	Sensitivity	Specificity	Accuracy
LR	0.739	0.693	0.648	0.670
GBM	0.748	0.661	0.663	0.662
RF	0.695	0.741	0.596	0.668
DT	0.701	0.603	0.622	0.613
NNET	0.745	0.693	0.663	0.678
XGBoost	0.750	0.667	0.674	0.670

LR, Logistic regression; GBM, Gradient boosting machine; RF, Random forest; DT, Decision tree; NNET, Neural network; XGBoost, Extreme gradient boosting.

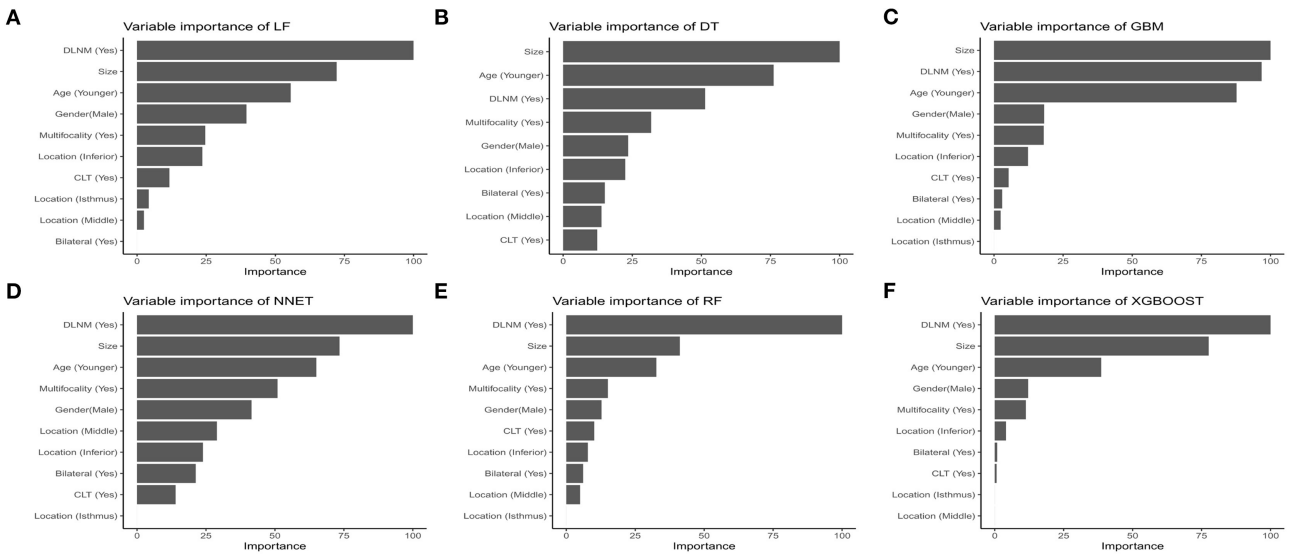

FIGURE 2 | Relative importance ranking of each input variable for predition of CLNM in the machine learning algorithms. **(A)** Logistic regression. **(B)** Decision tree. **(C)** Gradient boosting machine. **(D)** Neural network. **(E)** Random forest. **(F)** Extreme gradient boosting.

order: Delphian lymph node metastasis, tumor size, age, gender, multifocality and tumor location.

Web-Based Calculator

An online calculator based on the best-performing model was established for clinicians to predict patients' risk of developing CLNM by simply inputing readily available preoperative and intraoperative clinicopathological variables (https://jin63.shinyapps.io/ML_CLNM/) (**Figure 3**).

DISCUSSION

In this study, we developed and validated multiple popular machine learning algorithms to predict CLNM in patients with T1-T2, non-invasive, cN0 PTC. A comparison of ML algorithms identified that the XGBoost model gave the greatest performance. To make the application of this model available, we further established an online calculator for estimating the individual probability of CLNM in this subset patients with PTC. This ML-based model may potentially guide intraoperative decision-making.

It is noteworthy that the 2015 ATA guidelines (9) asserted that "thyroidectomy without pCLND is adequate for small (T1 or T2), non-invasive, clinically node-negative PTC." Yet, the risk of metastatic lymph nodes among this subgroup is unequal and a "one-size fits all" approach may raise concerns that in the long run it would bring potentially disastrous consequences for patients exempted from pCLND. Our data demonstrate that up to 51% of patients with T1-T2, non-invasive, cN0 PTC harbored central lymph node metastases. Such a high incidence of regional lymph node involvement is similar to other findings (18–20) and indicates that thyroid cancer is predisposed to LNM and that preoperative ultrasound currently fails to detect a massive number of patients with clinically significant lymph nodal disease (21, 22). Therefore, an accurate diagnosis of lymph node status carries much weight in helping clinicians determine the precise treatment for patients as well as informing the patients of prognoses and we advocate a selective approach to pCLND, particularly for cases with a high risk of CLNM.

Preoperative variables including larger tumors, younger age, male, multifocality, and tumor location in inferior portion are identified as the most important contributing predictors

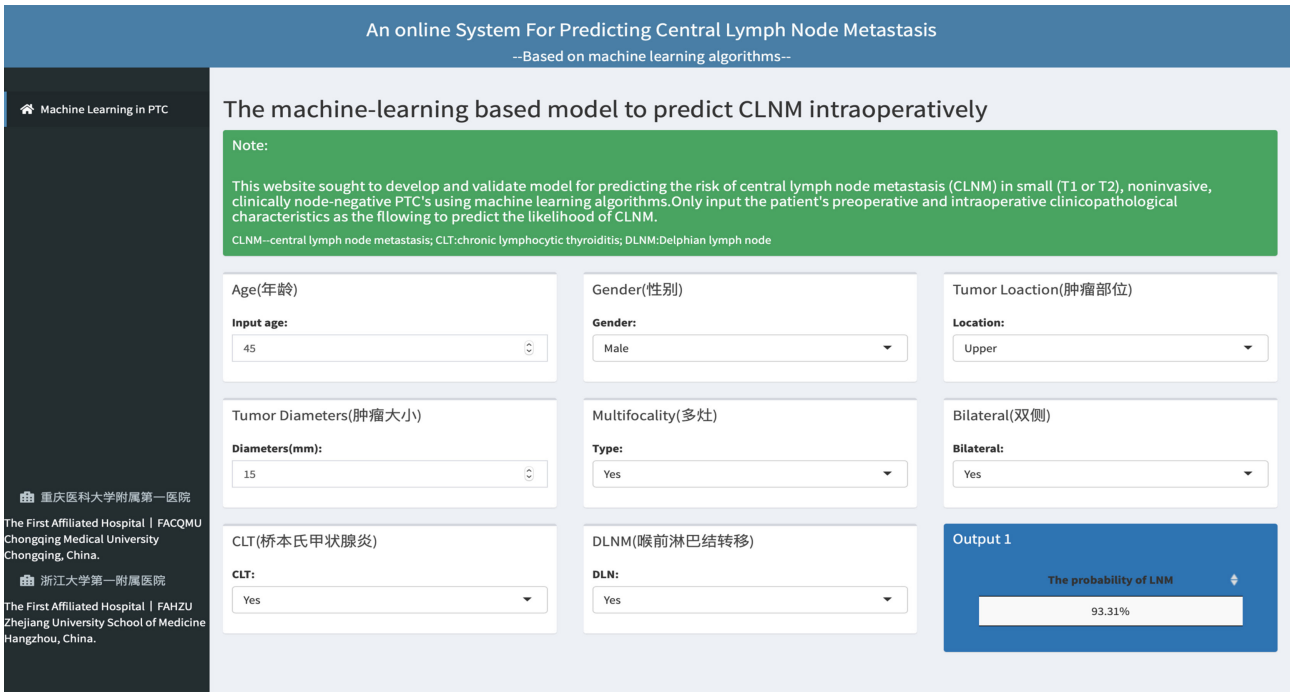

FIGURE 3 | The web-based calculator for predcting central lymph node metastasis in patients with T1-T2 stage, non-invasive, and clinically node negative PTC.

of CLNM-positive status by ML algorithms. The finding that younger age is highly predictive of CLNM in our research is similar to previous studies (23, 24). In addition, multifocal PTCs have been shown to be prone to CLNM and our results are consistent with previous reports, suggesting that multifocality is a positive predictor of CLNM (25, 26). It has been previously demonstrated by Thompson et al. (27) and Yang et al. (28) that larger tumors are significantly associated with an increased risk of nodal spread while we have found that rates of lymph node involvement surge in tumor sizes $>$ 20 mm, compared with those in tumor sizes of 10–20 mm and $<$ 10 mm (81.2 vs. 62.3 and 45.3%). Bilateral lesions are related with CLNM in the univariate analysis, but show insignificance in multivariate analyses after adjustment of confounders. All results have been confirmed in ML algorithms. Our study suggests that males are frequently found to be more susceptible to CLNM, which is supported by findings of previous studies (12, 29).

Nevertheless, the aforementioned factors in previous studies are mainly based on preoperative information and are still insufficient to achieve a reliable prediction. Besides, few studies have evaluated the predictive values of intraoperative factors. At our institution, lymph nodes in central compartment are classified as DLN, pretracheal and paratracheal nodes, respectively, and then routinely sent for frozen section examination separately. It was revealed in our previous study that the status of DLN based on frozen section examination was an independent predictor of CLNM and associated with poor prognostic features (14). And our findings of the present study further proves it, showing that 86.3% of DLN-positive patients have CLMN, compared with 43.9% of DLN-negative patients. The DLN status, in particular, is the strongest predictor in nearly all analytical approaches. Thus, we recommend routine intraoperative frozen section examination of DLN not only because the dissection of DLN can be performed safely without additional complications, but more importantly, it is a critical variable predicting further nodal metastases and aids in determining the extent of LN dissection. As intraoperative frozen section examination plays an essential role in immediate assessment of nodal status during an operation (30–32), it appears to be more promising in accurately predicting risks of LNM in subregion of central compartment when compared with preoperative evaluations alone.

Compared with studies attempting to predict the risk of central compartment lymph node metastases in PTC (12, 27, 28, 33, 34), our work has several strengths. First, few studies have ever focused on the subgroup of patients who suffer from clinically low-risk PTC. In fact, we found that a massive number of patients harbor clinically significant lymph node metastases which have not been detected by pre-oprative ultrasound. Furthermore, while ML approaches have shown unparalleled diagnostic performance in differentiating between benign and malignant thyroid nodules in recent reports (35, 36), there is, however, little research in the available literature on applying ML algorithms to lymph node metastases in PTC. To the best of our knowledge, this is the very first study to develop a prediction model using ML algorithms for real-time risk evaluation of CLNM with easy-to-use clinical data and fortunately, our model shows a great predictive power, which distinguishes itself from

linear models adopted by previous researches. Finally, in order to make this ML-based model easy to use, we established an online application based on it, which is now available for clinicians to facilitate individualized surgical treatment by calculating the risk for each patient: (https://jin63.shinyapps.io/ML_CLNM/). For instance, if a patient is identified to have a high probability of CLNM during surgery, then pCLND may be considered despite contradiction to the current ATA guidelines.

This study, however, also has limitations. First, the nature of a retrospective study might have resulted in selection bias. Second, the ML algorithm model we established, to some extent, was confined to one single institution, which might restrict its generalizability pending further validation in real-world scenarios. Third, predictve value was not high enough because the information in our current clinical database is to a certain degree limited.

CONCLUSIONS

We developed and validated ML algorithms for individualized prediction of CLNM in T1-T2 stage, non-invasive, and clinically node negative PTC patients by utilizing readily available preoperative variables and intraoperative frozen section examination. The ML-based prediction model can accurately identify whether patients are at high-risk of CLNM and its accompanying online risk calculator can serve as an easy-to-use tool for clinicians to make precise surgical decisions. In the future, our goal is to further integrate imaging, molecular and genetic data to improve our model performance in the realm of personalized medicine and more studies covering wider populations are also warranted for further validation.

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 First Affiliated Hospital of Chongqing Medical University. 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

JZhu, ZD, and XS conceived and designed the research. JZhu, JZhe, HR, RH, ZD, and XS prepared the manuscript. JZhu, JZhe, and DW collected and analyzed the data. JZhu, JZhe, and LL performed the statistical analysis and interpreted the results. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the Chongqing Science and Technology Committee (cstc2017shmsA1035).

ACKNOWLEDGMENTS

We appreciate the linguistic assistance provided by Sijia Xiong and TopEdit (www.topeditsci.com) during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.635771/full#supplementary-material

REFERENCES

1. Miller K, Nogueira L, Mariotto A, Rowland J, Yabroff K, Alfano C, et al. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin.* (2019) 69:363–85. doi: 10.3322/caac.21565
2. Deng Y, Li H, Wang M, Li N, Tian T, Wu Y, et al. Global burden of thyroid cancer from 1990 to 2017. *JAMA Netw Open.* (2020) 3:e208759. doi: 10.1001/jamanetworkopen.2020.8759
3. Ito Y, Kudo T, Kobayashi K, Miya A, Ichihara K, Miyauchi A. Prognostic factors for recurrence of papillary thyroid carcinoma in the lymph nodes, lung, and bone: analysis of 5,768 patients with average 10-year follow-up. *World J Surg.* (2012) 36:1274–8. doi: 10.1007/s00268-012-1423-5
4. Lee J, Song Y, Soh EY. Central lymph node metastasis is an important prognostic factor in patients with papillary thyroid microcarcinoma. *J Korean Med Sci.* (2014) 29:48. doi: 10.3346/jkms.2014.29.1.48
5. Wang TS, Sosa JA. Thyroid surgery for differentiated thyroid cancer — recent advances and future directions. *Nat Rev Endocrinol.* (2018) 14:670–83. doi: 10.1038/s41574-018-0080-7
6. Kim SK, Woo JW, Lee JH, Park I, Choe JH, Kim JH, et al. Prophylactic central neck dissection might not be necessary in papillary thyroid carcinoma: analysis of 11,569 cases from a single institution. *J Am Coll Surg.* (2016) 222:853–64. doi: 10.1016/j.jamcollsurg.2016.02.001
7. Hughes DT, Rosen JE, Evans DB, Grubbs E, Wang TS, Solórzano CC. Prophylactic central compartment neck dissection in papillary thyroid cancer and effect on locoregional recurrence. *Ann Surg Oncol.* (2018) 25:2526–34. doi: 10.1245/s10434-018-6528-0
8. Medas F, Canu G, Cappellacci F, Anedda G, Conzo G, Erdas E, et al. Prophylactic central lymph node dissection improves disease-free survival in patients with intermediate and high risk differentiated thyroid carcinoma: a retrospective analysis on 399 patients. *Cancers.* (2020) 12:1658. doi: 10.3390/cancers12061658
9. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid.* (2016) 26:1–133. doi: 10.1089/thy.2015.0020
10. Miller JE, Al-Attar NC, Brown OH, Shaughness GG, Rosculet NP, Avram AM, et al. Location and causation of residual lymph node metastasis after surgical treatment of regionally advanced differentiated thyroid cancer. *Thyroid.* (2018) 28:593–600. doi: 10.1089/thy.2017.0434

11. Singal AG, Mukherjee A, Elmunzer JB, Higgins PDR, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol.* (2013) 108:1723–30. doi: 10.1038/ajg.2013.332
12. Wu Y, Rao K, Liu J, Han C, Gong L, Chong Y, et al. Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. *Front Endocrinol.* (2020) 11:577537. doi: 10.3389/fendo.2020.577537
13. Zhu J, Huang R, Hu D, Dou Y, Ren H, Yang Z, et al. Individualized prediction of metastatic involvement of lymph nodes posterior to the right recurrent laryngeal nerve in papillary thyroid carcinoma. *OncoTargets Ther.* (2019) 12:9077–84. doi: 10.2147/OTT.S220926
14. Zhu J, Huang R, Yu P, Hu D, Ren H, Huang C, et al. Clinical implications of Delphian lymph node metastasis in papillary thyroid carcinoma. *Gland Surg.* (2021) 10:73–82. doi: 10.21037/gs-20-521
15. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. *Front Med.* (2019) 6:34. doi: 10.3389/fmed.2019.00034
16. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* (2019) 20:e262–73. doi: 10.1016/S1470-2045(19)30149-4
17. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. *Neurocomputing.* (2018) 300:70–9. doi: 10.1016/j.neucom.2017.11.077
18. Ma B, Wang Y, Yang S, Ji Q. Predictive factors for central lymph node metastasis in patients with cN0 papillary thyroid carcinoma: a systematic review and meta-analysis. *Int J Surg.* (2016) 28:153–61. doi: 10.1016/j.ijsu.2016.02.093
19. Feng JW, Pan H, Wang L, Ye J, Jiang Y, Qu Z. Determine the optimal extent of thyroidectomy and lymphadenectomy for patients with papillary thyroid microcarcinoma. *Front Endocrinol.* (2019) 10:363. doi: 10.3389/fendo.2019.00363
20. Fu GM, Wang ZH, Chen YB, Li CH, Zhang YJ, Li XJ, et al. Analysis of risk factors for lymph node metastases in elderly patients with papillary thyroid micro-carcinoma. *Cancer Manag Res.* (2020) 12:7143–9. doi: 10.2147/CMAR.S248374
21. Khokhar M, Day K, Sangal R, Ahmedli N, Pisharodi L, Beland M, et al. Preoperative high-resolution ultrasound for the assessment of malignant central compartment lymph nodes in papillary thyroid cancer. *Thyroid.* (2015) 25:1351–4. doi: 10.1089/thy.2015.0176
22. Liu C, Zhang L, Liu Y, Xia Y, Cao Y, Liu Z, et al. Ultrasonography for the prediction of high-volume lymph node metastases in papillary thyroid carcinoma: should surgeons believe ultrasound results? *World J Surg.* (2020) 44:4142–8. doi: 10.1007/s00268-020-05755-0
23. Shukla N, Osazuwa-Peters N, Megwalu UC. Association between age and nodal metastasis in papillary thyroid carcinoma. *Otolaryngol Head Neck Surg.* (2020). doi: 10.1177/0194599820966995
24. Luo X, Wang J, Xu M, Zou X, Lin Q, Zheng W, et al. Risk model and risk stratification to preoperatively predict central lymph node metastasis in papillary thyroid carcinoma. *Gland Surg.* (2020) 9:300–10. doi: 10.21037/gs.2020.03.02
25. Rosell R, Sun W, Lan X, Zhang H, Dong W, Wang Z, et al. Risk factors for central lymph node metastasis in CN0 papillary thyroid carcinoma: a systematic review and meta-analysis. *PLoS ONE.* (2015) 10:e0139021. doi: 10.1371/journal.pone.0139021
26. Lu S, Zhao R, Ni Y, Ding J, Qiu F, Peng Y, et al. Development and validation of a nomogram for preoperative prediction of cervical lymph node involvement in thyroid microcarcinoma. *Aging.* (2020) 12:4896–906. doi: 10.18632/aging.102915
27. Thompson AM, Turner RM, Hayen A, Aniss A, Jalaty S, Learoyd DL, et al. A preoperative nomogram for the prediction of ipsilateral central compartment lymph node metastases in papillary thyroid cancer. *Thyroid.* (2014) 24:675–82. doi: 10.1089/thy.2013.0224
28. Yang Z, Heng Y, Lin J, Lu C, Yu D, Tao L, et al. Nomogram for predicting central lymph node metastasis in papillary thyroid cancer: a retrospective cohort study of two clinical centers. *Cancer Res Treat.* (2020) 52:1010–8. doi: 10.4143/crt.2020.254
29. Liu C, Xiao C, Chen J, Li X, Feng Z, Gao Q, et al. Risk factor analysis for predicting cervical lymph node metastasis in papillary thyroid carcinoma: a study of 966 patients. *BMC Cancer.* (2019) 19:622. doi: 10.1186/s12885-019-5835-6
30. Antic T, Taxy J. Thyroid frozen section: supplementary or unnecessary? *Am J Surg Pathol.* (2013) 37:282–6. doi: 10.1097/PAS.0b013e318267aee6
31. Lim YS, Choi SW, Lee YS, Lee JC, Lee BJ, Wang SG, et al. Frozen biopsy of central compartment in papillary thyroid cancer: quantitative nodal analysis. *Head Neck.* (2013) 35:1319–22. doi: 10.1002/hed.23129
32. Raffaelli M, De Crea C, Sessa L, Fadda G, Bellantone C, Lombardi CP. Ipsilateral central neck dissection plus frozen section examination versus prophylactic bilateral central neck dissection in cN0 papillary thyroid carcinoma. *Ann Surg Oncol.* (2015) 22:2302–8. doi: 10.1245/s10434-015-4383-9
33. Wang Y, Guan Q, Xiang J. Nomogram for predicting central lymph node metastasis in papillary thyroid microcarcinoma: a retrospective cohort study of 8668 patients. *Int J Surg.* (2018) 55:98–102. doi: 10.1016/j.ijsu.2018.05.023
34. Zhao W, He L, Zhu J, Su A. A nomogram model based on the preoperative clinical characteristics of papillary thyroid carcinoma with Hashimoto's thyroiditis to predict central lymph node metastasis. *Clin Endocrinol.* (2020) 94:310–21. doi: 10.1111/cen.14302
35. Zhang B, Tian J, Pei S, Chen Y, He X, Dong Y, et al. Machine learning–assisted system for thyroid nodule diagnosis. *Thyroid.* (2019) 29:858–67. doi: 10.1089/thy.2018.0380
36. Zhao HN, Liu JY, Lin QZ, He YS, Luo HH, Peng YL, et al. Partially cystic thyroid cancer on conventional and elastographic ultrasound: a retrospective study and a machine learning—assisted system. *Ann Transl Med.* (2020) 8:495. doi: 10.21037/atm.2020.03.211

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 Medicine

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fmed.2021.653925

Brain Tumor Segmentation via Multi-Modalities Interactive Feature Learning

Bo Wang[1,2], Jingyi Yang[3], Hong Peng[4], Jingyang Ai[2], Lihua An[5], Bo Yang[6], Zheng You[1] and Lin Ma[4]**

[1] The State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing, China, [2] Beijing Jingzhen Medical Technology Ltd., Beijing, China, [3] School of Artificial Intelligence, Xidian University, Xi'an, China, [4] Department of Radiology, The 1st Medical Center, Chinese PLA General Hospital, Beijing, China, [5] Radiology Department, Affiliated Hospital of Jining Medical University, Jining, China, [6] China Institute of Marine Technology & Economy, Beijing, China*

Automatic segmentation of brain tumors from multi-modalities magnetic resonance image data has the potential to enable preoperative planning and intraoperative volume measurement. Recent advances in deep convolutional neural network technology have opened up an opportunity to achieve end-to-end segmenting the brain tumor areas. However, the medical image data used in brain tumor segmentation are relatively scarce and the appearance of brain tumors is varied, so that it is difficult to find a learnable pattern to directly describe tumor regions. In this paper, we propose a novel cross-modalities interactive feature learning framework to segment brain tumors from the multi-modalities data. The core idea is that the multi-modality MR data contain rich patterns of the normal brain regions, which can be easily captured and can be potentially used to detect the non-normal brain regions, i.e., brain tumor regions. The proposed multi-modalities interactive feature learning framework consists of two modules: cross-modality feature extracting module and attention guided feature fusing module, which aim at exploring the rich patterns cross multi-modalities and guiding the interacting and the fusing process for the rich features from different modalities. Comprehensive experiments are conducted on the BraTS 2018 benchmark, which show that the proposed cross-modality feature learning framework can effectively improve the brain tumor segmentation performance when compared with the baseline methods and state-of-the-art methods.

Keywords: brain tumor segmentation, deep neural network, multi-modality learning, feature fusion, attention mechanism

OPEN ACCESS

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Wei Wei Wei Wei,
Xi'an University of Technology, China
Puzhao Zhang,
Royal Institute of Technology, Sweden

****Correspondence:***
Zheng You
yz-dpi@mail.tsinghua.edu.cn
Lin Ma
malin.rad@301hospital.com.cn

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *15 January 2021*
Accepted: *04 March 2021*
Published: *13 May 2021*

Citation:
Wang B, Yang J, Peng H, Ai J, An L, Yang B, You Z and Ma L (2021) Brain Tumor Segmentation via Multi-Modalities Interactive Feature Learning. Front. Med. 8:653925. doi: 10.3389/fmed.2021.653925

1. INTRODUCTION

Brain cancer is an aggressive and highly lethal malignancy that has received more and more attention and presented multiple technical challenges for studies on brain tumors. Owing to the diversity of the appearance and morphology of brain tumors, accurately automatically segmenting tumor areas from multi-modality magnetic resonance image (MRI) sequences is a difficult but meaningful issue in field of artificial intelligence and assisted diagnosis (1). In this paper, we study a deep-learning based automatic brain tumor segmentation network to assist clinicians in improving

the diagnostic efficiency of brain tumors. For the automatically tumor segmentation task, the input medical images are multi-modality data and the corresponding segmentation masks contain multi areas of the brain tumor. Specifically, the input multi-modality medical image consist of four MRI modality, i.e., T1-weighted (T1) modality, contrast enhanced T1-weighted (T1c) modality, T2-weighted (T2) modality, and T2 Fluid Attenuation Inversion Recovery (FLAIR) modality. The goal of brain tumor segmentation is to determine the volume, shape, and localization of brain tumor areas, i.e., the whole tumor (WT) area, the tumor core (TC) area, and the enhancing tumor (ET) core area, which play crucial roles in brain tumor diagnosis and monitoring.

To achieve automatic brain tumor segmentation, some methods use the deep convolutional neural network (DCNNs) to extract the features of tumors and determine the labels of multi-class pixels in the end-to-end fashion. However, existing brain tumor segmentation methods (2–4) usually consider this task as a semantic segmentation problem for common nature images, which methods omit the great disparity between the medical image and the common nature image. Specifically, there are two-fold distinct properties between these two kinds of images: (1) As a departure from the common nature image, the medical image usually consist of multiple MRI modalities that capture different pathological properties. (2) The geometrical shape, spatial position, and texture structure of tumor in medical images are complex and changeable, and the tumor does not have a specific, regular pattern of appearance. Therefore, such existing approaches would not obtain the optimal solutions.

Due to the above discussions properties, for the brain tumor segmentation task, the deep learning based segmentation methods still has challenging issues needed to be addressed. First, the existing methods cannot fully mine the potential knowledge in multi-modalities. Specifically, the previous works use simple parameter-sharing feature extractors to obtain features of different modal data and directly concatenate the information from different modality data. Such feature extraction and processing methods lack a data mining strategy for effectively informational fusing and extracting knowledge from complex data structures. Second, due to the nonspecific structural pattern in the tumor area, the existing supervised learning-based segmentation methods, which are guided only by a manually annotated foreground and background segmentation ground truth, are difficult to learn the complete discriminant information of brain tumor.

To address these issues, in this paper, we proposed a novel interactive modality deep feature learning framework to learn the discriminant information of brain tumor from the multi-modality MRI data. Considering the fact that the texture and spatial position of normal organs in medical images have specific structural patterns, and deep neural networks can easily learn discriminant information from such regular patterns. Meanwhile, radiologists need to combine information from multiple modalities to determine the full range of areas of a brain tumor. For the multi-modality MRI data, the intra-modality information describes the discriminant feature between the normal organ and the lesion area (i.e., brain area and tumor area) in medical images, the inter-modality information provides additional cross-modal constraints for determining the visual boundaries and different regions of the brain tumor. Specifically, the proposed interactive modality deep feature learning framework consists of the cross-modality feature extraction and the normal region-guided feature fusion.

Figure 1 illustrates the proposed learning framework briefly. In the cross-modality feature extracting process, we adopt a two-step feature interacting strategy to extract the interactive features across different modality data. The first feature interacting step concatenates multi-modality image data in channel-wise to extract the low-level interactive features at input level, and the second feature interacting step integrates the high-level features of different modality pairs to extract the high-level interactive features. In the normal region-guided feature fusion, we propose a novel reverse attention-based feature fusion framework to collectively enhance the features of normal brain region from different modality data. This encourages the feature extracting network to learn intrinsic patterns that are helpful to determine the normal brain area from each modality data. The intuition behind this process is that the reverse attention mechanism enhance the non-tumor regions in the brain MRI data, and those regions contain rich structure and texture information of normal brain regions.

2. RELATED WORKS

2.1. Brain Tumor Segmentation

Brain tumor segmentation is a hot topic in the medical image analysis and machine learning community. It has received great attention in the past few years. Early efforts in this filed designed hand-crafted features and adopted the classic machine learning models to predict the brain tumor areas. Due to the rapid development of the deep learning technique (5–9), the recent brain tumor segmentation approaches mainly apply the deep features and classifiers from the DCNN models. Based on the type of the convolutional operation used in the DCNN models, we briefly divide the existing methods into two groups, i.e., the 2D CNN-based methods and 3D CNN-based method. The 2D CNN-based methods (10–12) apply the 2D convolutional operations and split the 3D volume samples into 2D slices or 2D patches. While the 3D CNN-based methods (13–16) apply the 3D convolutional operations, which can take the whole 3D volume samples or the extracted sub-3D patches as the network input.

2.2. Multi-Modality Feature Learning

Multi-modality feature learning is gaining more and more attention in the recent years as the multi-modality data can provide richer information for sensing the physical world. Existing works have applied multi-modality feature learning in many computer vision-based tasks such as 3D shape recognition (17–20) and retrieval (21–24), survival prediction (25), RGB-D object recognition (26), and person re-identification (27). Among these methods, Bu et al. (21) built a multi-modality fusion head to fuse the deep features learnt by

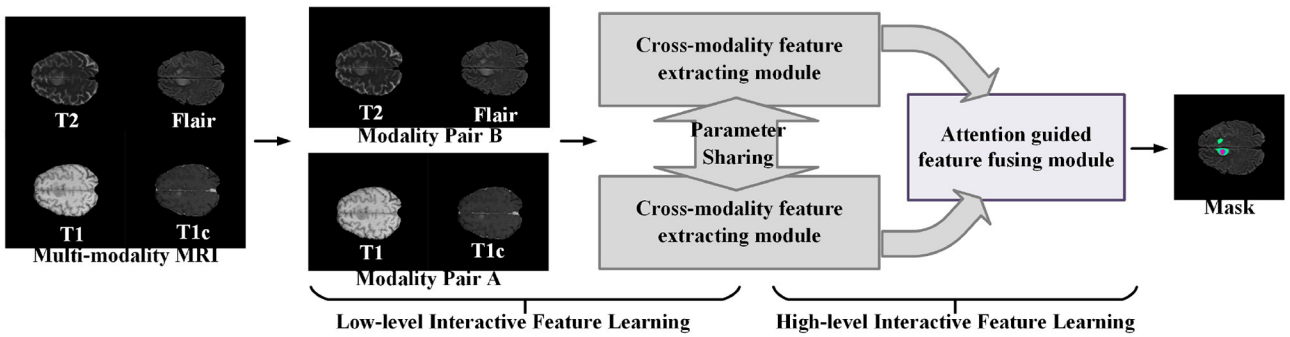

FIGURE 1 | A brief illustration of the proposed multi-modality interactive feature learning framework for brain tumor segmentation.

a CNN network branch and a deep belief network (DBN) branch. To integrate multiple modalities and eliminate view variations, Yao et al. (25) designed a deep correlational learning module for learning informative features on the pathological data and the molecular data. Wang et al. (28) proposed a large-margin multi-modal deep learning framework to discover the most discriminative features for each modality and harness the complementary relationship between different modalities.

3. DATASET DESCRIPTION

We implement all experiments on BraTS 2018 benchmark (29–31) to evaluate the performance of proposed brain tumor segmentation. The BraTS 2018 benchmark dataset contains four modalities, i.e., T1, T1-c, T2, and FLAIR, for each patient. The BraTS 2018 benchmark has two subsets: a training set, which contains 285 subjects, and a validation set containing 66 subjects with hidden ground truth. Each subject holds a manual expert segmentation of three tumor sub-compartments: edema (ED), ET, and necrotic tissue combined with non-enhancing tumor (NCR/NET). In the official BraTS evaluation, these sub-compartments are combined into three hierarchical labels: WT, TC, and ET. WT is a combination of all tumor sub-compartments (i.e., ET, NCR/NET), TC combines ET and NCR/NET, and ET is defined by the ET sub-compartment. Aiming at yielding uncertainty estimates for these hierarchical tumor regions, we combined the tumor sub-compartment labels into the hierarchical labels before the training of the automated segmentation models. The BraTS 2018 dataset comes preprocessed; the subjects and MR images are co-registered to the same anatomical template, resampled to unit voxel size ($1 \times 1 \times 1$), and skull stripped. When implementing the experiments on each of the benchmarks, we randomly select the 80% data in training set to train the brain tumor segmentation models while use the rest of the data in training set to test the segmentation performance. We additionally normalized each MR image subject-wise to zero mean and unit variance.

4. METHODS

The aim is to segment the brain tumor regions including the WT region, the TC region, and the enhancing TC region from multi-modality MRI data. For this purpose, we propose to build a multi-modality-based single prediction multi-region segmentation method that utilizes the cross-modalities interactive features from MRI data. In this work, we propose to train a cross-modalities interactive feature extracting and fusing network using reverse attention guidance and use the trained network for segmenting brain tumor regions in MRI data.

In this section, we first describe the network architecture and the workflow of the proposed multi-modalities brain tumor segmentation framework and also including the details of the cross-modality feature extracting process and the attention-guided feature fusion that are two important interactive feature learning modules. Then, we introduce the implement details of the training process and experiments.

4.1. Multi-Modalities Brain Tumor Segmentation Network

Given an input MRI data $\mathbf{X} = \{x_{T1}, x_{T1c}, x_{T2}, x_{FLAIR}\}$, where the variables x_{T1},x_{T1c},x_{T2}, and x_{FLAIR} represent the T1-weighted modality, the contrast-enhanced T1-weighted modality, the T2-weighted modality, and the fluid attenuation inversion recovery modality, respectively, we follow the work (32) to split the multi-modalities input $\mathbf{X}$ to form two modality pairs $X_{g1} = \{x_{T1}, x_{T1c}\}$ and $X_{g2} = \{x_{T2}, x_{FLAIR}\}$, which encourages the information within each modality pair tends to be consistent while the information from different modality pairs tends to be distinct and complementary. The cross-modality feature extracting module takes the modality pair as input, and outputs the interactive features of the multi-modalities data. Then, the attention-guided feature fusion module takes the interactive features as input and output the fused cross-modality interactive feature. Finally, the segmentation results of the brain tumor region are generated from the fused cross-modality interactive feature. The network architecture of our proposed multi-modalities brain tumor segmentation framework is shown in **Figure 2**. Each component will be elaborated as follows.

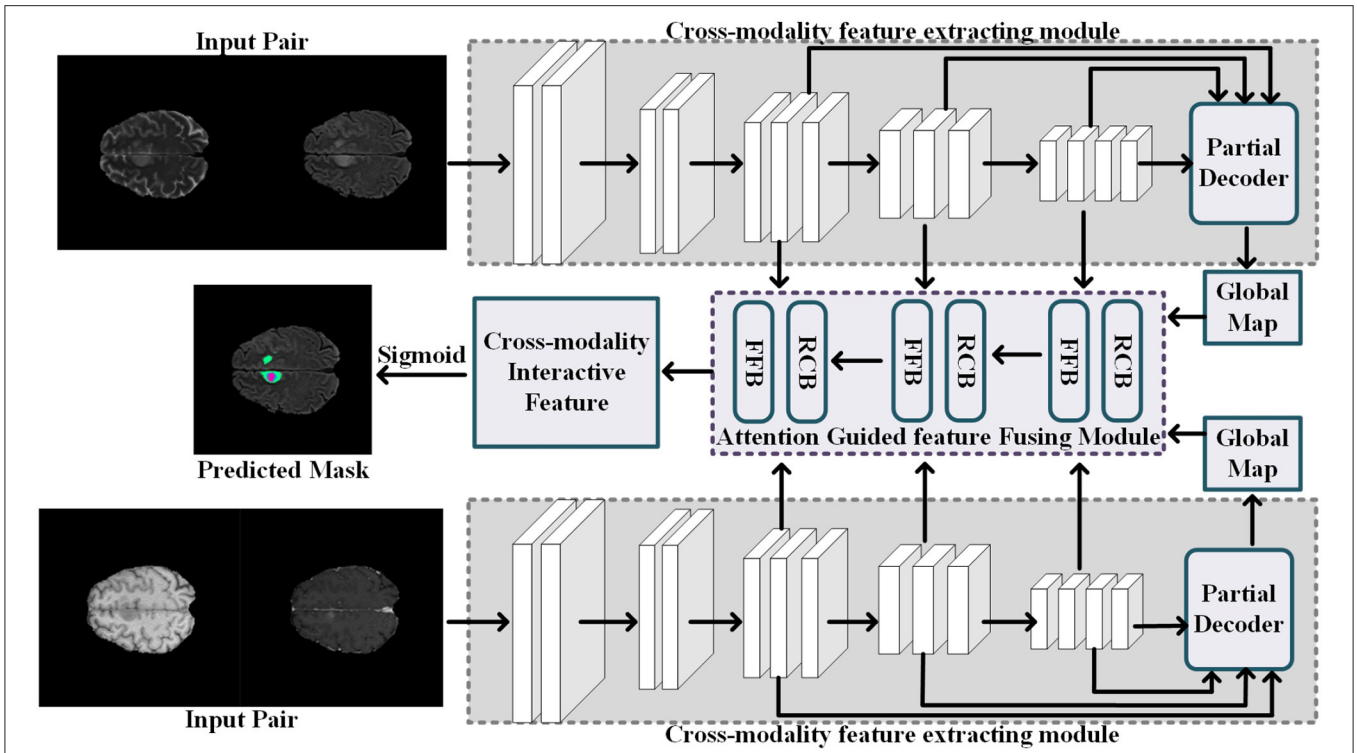

FIGURE 2 | Illustration of the network architecture of the segmentation process.

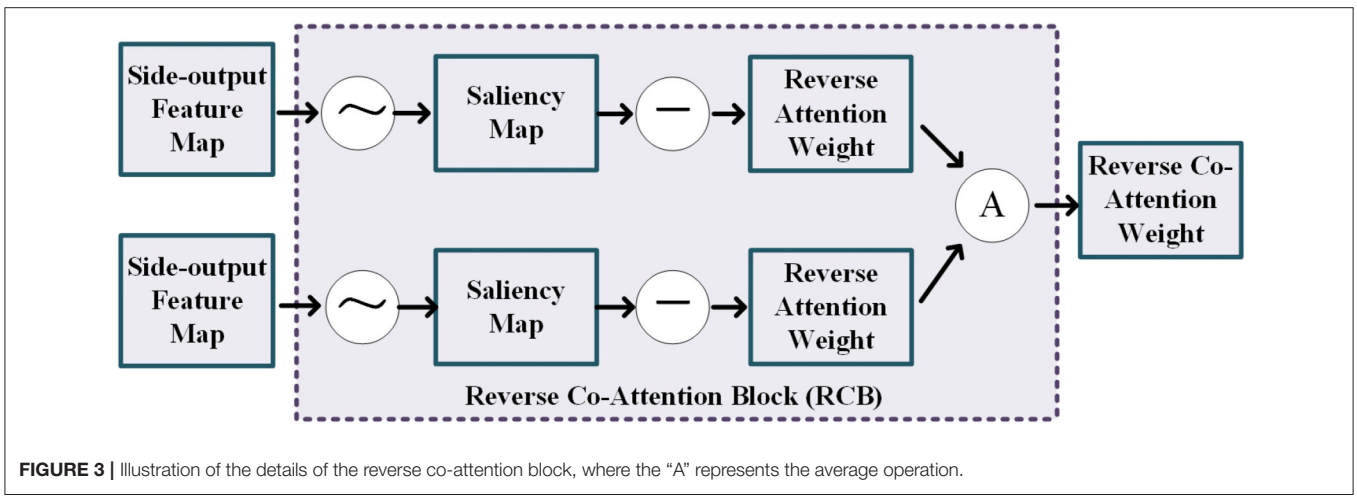

FIGURE 3 | Illustration of the details of the reverse co-attention block, where the "A" represents the average operation.

4.1.1. Cross-Modality Feature Extracting Module

Current popular multi-modalities feature extracting network usually rely on a single simple interactive strategy, i.e., the channel concatenation (33) or the parameters sharing (34). The channel concatenation strategy only considers the common features among different modalities, but ignore the richness of the features brought by the modes; conversely, the parameters sharing strategy only pays attention to the richness of features brought by multi-modalities, but ignores the common features among different modalities. To effectively interact features between different modalities, we employ the combinational strategy of both the channel concatenation and the parameters sharing to extract the common features among similar modalities and use the information between different modalities to improve the richness of the features. Specifically, we use a CNN-based network to extract the common features in a modality pair where the modalities sharing consistent feature for common pathological areas and normal areas, as shown in **Figure 2**. The cross-modality feature extracting module has two input channels corresponding to the two MR images from one modality pair,

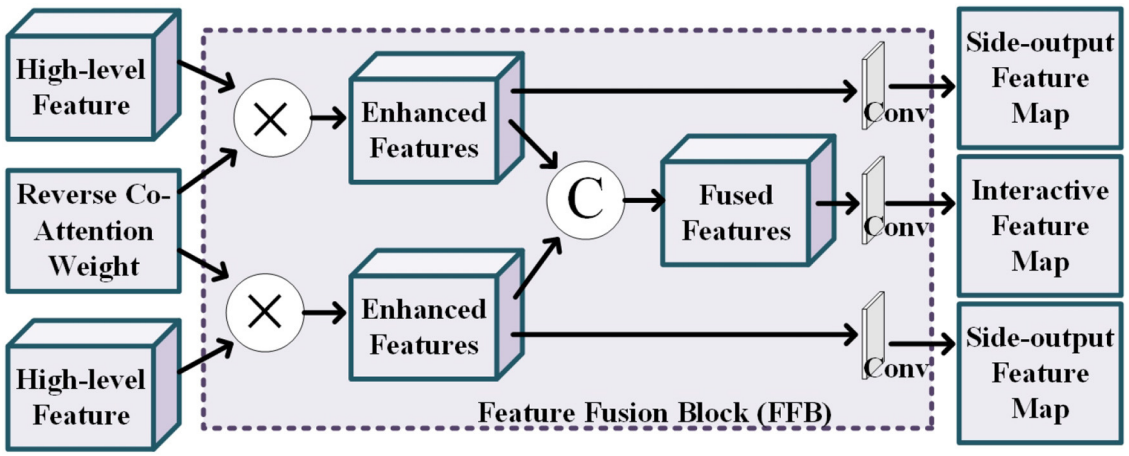

FIGURE 4 | Illustration of the details of the feature fusion block, where the operation "C" represents the channel-wise concatenation.

i.e., $X_{g1} = \{x_{T1}, x_{T1c}\}$ or $X_{g2} = \{x_{T2}, x_{FLAIR}\}$. Meanwhile, the feature extractor is sharing parameters for extracting the interactive features of the different modality pairs.

Considering the low-level features contribute less to segmentation performance but demand more computational resources, we aggregate the high-level features to predict the common brain tumor areas in each modality pair. Specifically, for an input modality pair $x_{g1} = \{x_{T1}, x_{T1c}\}$ (or $x_{g2} = \{x_{T2}, x_{FLAIR}\}$), each modality data with size $h \times w \times l$, five levels of features $f_i, i = 1, \ldots, 5$ with resolution $[h/2^{k-1}, w/2^{k-1}, l/2^{k-1}]$ can be extracted from the cross-modality feature extracting network. Then, we follow the work (35) to divide interactive features f_i into low-level features group $\{f_i, i = 1, 2\}$ and high-level features group $\{f_i, i = 3, 4, 5\}$. The low-level features contain lots of modality information, which are not applicable to interactive features fusion between multi-modalities. Thus, we employ the partial decoder D_p (35) to only aggregate the high-level feature $\{f_i, i = 3, 4, 5\}$ with a cascade fashion. The interactive feature of one modality pair is computed by the $f_{Dp} = D_p(f_3, f_4, f_5)$, and we also can obtain the global map M_g of the input modality pair.

4.1.2. Attention Guided Feature Fusing Module

The global map M_g is formed by the high-level features $\{f_i, i = 3, 4, 5\}$, which captures the high-level information such as normal brain areas and tumor areas. However, the rich diversity of brain tumor regions makes it impossible for feature extraction models to extract a learnable structural pattern from this region. Compared with brain tumor regions, the normal brain regions in the training images are regularly distributed, and these structural patterns are easier to perceive and extract. Motivated by this observation, we propose a cross-modality features fusing strategy to progressively discriminative brain regions through an erasing foreground object manner [pranet 27,4]. Instead of predicting the non-normal regions (brain tumor areas) directly, we propose to determine the normal brain regions in the multi-modalities MR data by learning the reverse attention (35) from the high-level features. The proposed attention-guided feature fusing module consists of two blocks: the feature fusion block and the reverse co-attention block.

As shown in **Figure 3**, the reverse co-attention block takes two side-output feature maps from two modality pairs as input and outputs a reverse co-attention weight. The side-output feature maps $M_i, i = 3, 4, 5$ are generated by the previous FFD (feature fusing block). In each reverse co-attention block, a sigmoid operation and a reverse operation are used to generate the reverse attention weight R_i. The reverse attention weight R_i is a negative salient object detection in the computer vision community (36–39) and can be formulated as Equation (1):

$$R_i = \ominus(\sigma(M_i)) \tag{1}$$

where the $\ominus$ denotes a reverse operation subtracting the input from all 1's matrix **E** and σ is the Sigmoid function. To explore the high-level interactive features of the two modality pairs, we average the reverse attention weights from the two cross-modalities feature extracting module to generate a reverse co-attention weights $\overline{R}_i$.

The details of feature fusing block is shown in **Figure 4**. This block tasks the high-level features of the two modality pairs and a reverse co-attention weight as input to generate the side-output feature maps and the interactive feature map. The reverse co-attention weight enhances the features of the common interest regions in the two modality pairs, and weakens the features of the common no interest regions, which will enable deep integration of features between multiple modality pairs. Specifically, the output interactive features $\{\overline{f}_i, i = 3, 4, 5\}$ of each modality pair can be obtained by element-wise multiplying ($\bigotimes$) the high-level feature $\{f_i, i = 3, 4, 5\}$ by the reverse co-attention weight $\overline{R}_i$, as Equation (2):

$$\overline{f}_i = f_i \bigotimes \overline{R}_{i+1} \tag{2}$$

We concatenate the reverse co-attention feature of the two modality pairs in channel-wise to deeply fuse the features of the

two modality pairs. The final segmentation result is obtained by progressively superpose the fused features.

4.2. Learning Process and Implementation Details

4.2.1. Loss Function

Our loss function consist of segmentation loss $\mathcal{L}_{sg}$ and saliency detection loss $\mathcal{L}_{sd}$. The $\mathcal{L}_{sg}$ is Dice Similarity Coefficient (DSC) (32), which evaluates the similarity between two higher-dimensional sets, i.e., the segmentation masks and the ground-truth masks, and can be formulated as Equation (4):

$$\mathcal{L}_{sg}(\mathbf{Y}, \mathbf{U}) = 1 - \frac{2 \times |\mathbf{Y} \bigcap \mathbf{S}|}{|\mathbf{Y}| + |\mathbf{S}|} \tag{3}$$

where $\mathbf{Y}$ and $\mathbf{S}$ represent the ground-truth annotation and the segmentation mask for the desired brain tumor areas, respectively.

The saliency detection loss $\mathcal{L}_{sd}$ implements deep supervision for the three side-output feature maps $\{M_3, M_4, M_5\}$ and the global map M_g, which prevents the model from being heavily affected by the unbalance among different types of tumor areas. We adopt weighted binary cross entropy (BCE) loss to achieve this proposal. The weighted BCE loss pays more attention to hard pixels rather than assigning all pixels equal weights (35). The definitions of these losses are the same as in [21,26] and their effectiveness has been validated in the field of salient object detection. Each map is up-sampled M_i^{up} to the same size as the ground-truth map $\mathbf{G}$, which is obtained by dividing the tumor regions annotation into three separate binary maps (i.e., WT map, ET map, and TC map). The deep supervision loss L_{deep} can be formulated as Equation (4):

$$L_{deep} = \mathcal{L}_{sd}(\mathbf{G}, M_g^{up}) + \sum_{i=3}^{5} \mathcal{L}_{sd}(\mathbf{G}, M_i^{up}) \tag{4}$$

The total loss function L_{total} can be formulated as Equation (5):

$$L_{total} = \alpha \mathcal{L}_{sg}(\mathbf{Y}, \mathbf{U}) + (1 - \alpha)(\mathcal{L}_{sd}(\mathbf{G}, M_g^{up}) + \sum_{i=3}^{5} \mathcal{L}_{sd}(\mathbf{G}, M_i^{up})) \tag{5}$$

where the weight α is empirically set to 0.7.

4.2.2. Implementation Details

We follow the work (32) to adopt the pre-trained parameters of transition generative networks to initialize the feature extracting network in our methods. Specifically, each of the input modality data was normalized to have zero mean and unit variance, and the inputs of both the cross-modality feature transition are randomly sampled from the training data set, and the input patch size is $128 \times 128 \times 128$. We also employ U-net as backbone, where the base number of filters is 16 and increased to twice after each down-sampling layer. We use Adam optimizer with an initial learning rate is 10-e4 and λ is 10 to optimize the objective function. The network branches were implemented in Pytorch on four NVIDIA GTX 1080TI GPU. It totally takes 5 h to complete the training process and the test speed is 2.5 s per subject.

5. EVALUATION METRICS

The performance of the segmentation algorithm is evaluated based on two metrics, i.e., the Dice score, and the 95th percentile of the Hausdorff Distance (Hausdorff95).

The Dice score is a commonly used metric for measuring the segmentation accuracy at the pixel level. It is a statistical gauge of the similarity between two sets of samples. Given S, a set of pixels belonging to a ground truth of the segmentation mask of brain tumor regions, and P, a set of pixels belonging to a predicted segmentation mask of the brain tumor regions.The Dice score is defined as in Equation (6), where $|\cdot|$ denotes set cardinality. The Dice score ranges from 0 (no overlap between S and P) to 1 (perfect overlap between S and P), and the lower is better.

$$Dice = \frac{2 \times |\mathbf{S} \bigcap \mathbf{P}|}{|\mathbf{S}| + |\mathbf{P}|} \tag{6}$$

The 95th percentile of the Hausdorff Distance (Hausdorff95) is a boundary-based segmentation accuracy evaluation metric. It calculates the distance between the two point sets. Considering the predicted segmentation mask P and the ground-truth mask S, the Hausdorff distance between the two set is defined as Equation (7):

$$d_H(S, P) = max[max_{p \in P} min_{s \in S}[D(S, P)], max_{s \in S} min_{p \in P}[d(S, P)]] \tag{7}$$

where the $d_H(x, y)$ denotes the distance between pixels $x \in P$ and $y \in S$. We follow the work (40) to use Euclidean distance to calculate the pixel-wise distance. The Hausdorff distance represents the longest distance from P (respectively S) to its closest point in S (respectively P). It is the most extreme value from all distances between the pairs of the nearest pixel on the boundaries of S and P. Finally, the score of Hausdorff distance is multiplied by 95% to eliminate the interference from outlier points.

In this work, the predicted segmentation masks are compared with the ground-truth masks via Dice score and the 95th percentile of Hausdorff distance (Hausdorff95). A higher Dice coefficient and a lower Hausdorff distance indicate the efficacy of the brain tumor segmentation method.

6. RESULTS

This section presents quantitative and qualitative evaluations of the performance of the proposed segmentation method to segment the three brain tumor regions in the multi-modality MRI data.

6.1. Ablation Study of the Proposed Approach

For the analysis of the contribution and the effect of our proposed branches for brain tumor segmentation task, we conduct the

TABLE 1 | Ablation study of the proposed approach and the other baseline models on the BraTS 2018 validation set.

Methods	Dice score				Hausdorff95			
	WT	ET	TC	Average	WT	ET	TC	Average
"f_{g1}"	0.698	0.793	0.808	0.766	4.412	9.614	8.184	7.403
"f_{g2}"	0.517	0.876	0.749	0.714	10.461	5.668	9.472	8.534
"f_{g1+2}"	0.674	0.818	0.782	0.758	5.072	6.101	8.562	6.578
"Ours w/o CA"	0.778	0.885	0.819	0.827	3.841	5.912	7.291	5.681
"Ours w AT"	0.789	0.897	0.836	0.841	4.690	4.912	6.912	5.505
Ours	0.801	0.909	0.854	0.855	3.879	4.571	6.411	4.954

Higher Dice scores indicate the better results, while lower Hausdorff95 scores indicate the better results.

TABLE 2 | Comparison results of the proposed approach and the other state-of-the-art models on the BraTS 2018 validation set.

Methods	Dice score				Hausdorff95			
	WT	ET	TC	Average	WT	ET	TC	Average
Myronenko (33)	0.823	0.910	0.867	0.866	3.926	4.516	6.855	5.099
Isensee et al. (42)	0.809	0.913	0.863	0.861	2.410	4.270	6.520	4.400
Puch et al. (2)	0.758	0.895	0.774	0.809	4.502	10.656	7.103	7.420
Chandra et al. (3)	0.767	0.901	0.813	0.827	7.569	6.680	7.630	7.293
Ma et al. (4)	0.743	0.872	0.773	0.796	4.690	6.120	10.400	7.070
Chen et al. (43)	0.733	0.888	0.808	0.810	4.643	5.505	8.140	6.096
Zhang et al. (32)	0.791	0.903	0.836	0.843	3.992	4.998	6.369	5.120
Ours	0.801	0.909	0.854	0.855	3.879	4.571	6.411	4.954

Higher Dice scores indicate the better results, while lower Hausdorff95 scores indicate the better results.

experiments on the following baseline models. The first two baseline models train the single-modality-pair feature extracting modules "f_{g1}" and "f_{g2}" with the input modality data $X_{g1} = \{x_{T1}, x_{T1c}\}$ or $X_{g2} = \{x_{T2}, x_{FLAIR}\}$, respectively. The third baseline model "f_{g1+2}" fuses the prediction of "f_{g1}" and "f_{g2}" by directly computing the average of the obtained segmentation maps without using any feature fusing strategies proposed in this paper. The first three baselines are designed to analyze the contribution of the multi-modalities of the MRI data for segmenting the brain tumor regions. We also introduce two baseline models "Ours w AT" and "Ours w/o CA" to analyze the contribution of the proposed reverse attention-guided feature fusion and segmentation module. Specifically, "Ours w AT" represents the feature fusion module use, a saliency attention strategy (41) to fuse the cross-modalities features, and "Ours w/o CA" represents the feature fusion module use, the independent reverse attention that do not interact between the modality pairs to guide the feature fusing. We use the parameters of the pre-trained generative feature transition network (32) to initialize all the aforementioned baseline models, and these baselines are fine-tuned on the same training data as our method. The experimental results are reported in top rows of **Table 1**.

By comparing single-modality pair modules ("f_{g1}" and "f_{g2}") and the multi-modality pair baseline "f_{g1+2}", we observe that the baseline achieves more stable performance than the single-modality pair modules, but it does not achieve the better comprehensive performance than baseline "f_{g1}." This can demonstrate that the arbitrary feature fusion has limited improvement on segmentation performance due to the lack of effective fusion strategy. By comparing the attention-guided feature fusion baselines (i.e., "Ours w/o CA" and "Ours w AT") with the "f_{g1+2}", we can observe that the attention-guided feature fusion can improve the segmentation performance. It demonstrates that the performance improvement of our method mainly comes from the well-designed multi-modalities feature fusion and learning strategy. By comparing "Ours" with baselines "Ours w/o CA" and "Ours w AT," we can observe that the common attention of modality pairs plays an important role in fusing informative features and predicting accurate tumor areas (see "Ours w/o CA" vs. "Ours w AT"), and the reverse attention mechanism can further improve the segmentation performance (see "Ours" vs. "Ours w AT"). The ablation analysis demonstrates the contribution of our proposed cross-modality feature extracting module and attention guided feature fusing module for improving the performance of brain tumor segmentation.

6.2. Comparison With State-of-the-Art Methods

To evaluate the effectiveness of the proposed brain tumor segmentation model, on the BraTs2018 dataset, we follow the work of (32) to compare the segmentation performance of the proposed method with seven state-of-the-art methods including three ensemble-models methods, i.e., Myronenko (33), Isensee et al.(42), Puch et al.(2), and four single-prediction methods: Chandra et al. (3), Ma et al. (4), Chen et al.(43), and Zhang et al. (32). The quantitative results are reported in **Table 2**. The performances of the segmentation models were evaluated with the Disc score and Hausdorff95. From **Table 2**, we can observe that our methods achieve the best performance when comparing with the state-of-the-art single-prediction methods both in terms of Dice score and Hausdorff95. When comparing with the ensemble-models methods, our method has the second best performance. Usually, the ensemble-models methods can usually obtain better performance than the single-prediction methods, since the ensemble models methods integrate multiple brain tumor segmentation models that are trained by using different views or different training subsets, while the single prediction methods only use one segmentation model to implement multi-brain tumor areas segmentation tasks. However, the ensemble-models methods require training multiple models with more training data, which means higher complexity both in computational cost and time consumption. Considering the balance between time cost and algorithm performance, the performance of our method is satisfactory. Thus, the comparison results in **Table 2** demonstrate the effectiveness of the proposed approach.

In **Figure 5**, we also show some examples of the brain tumor segmentation results for quantitative analysis. From

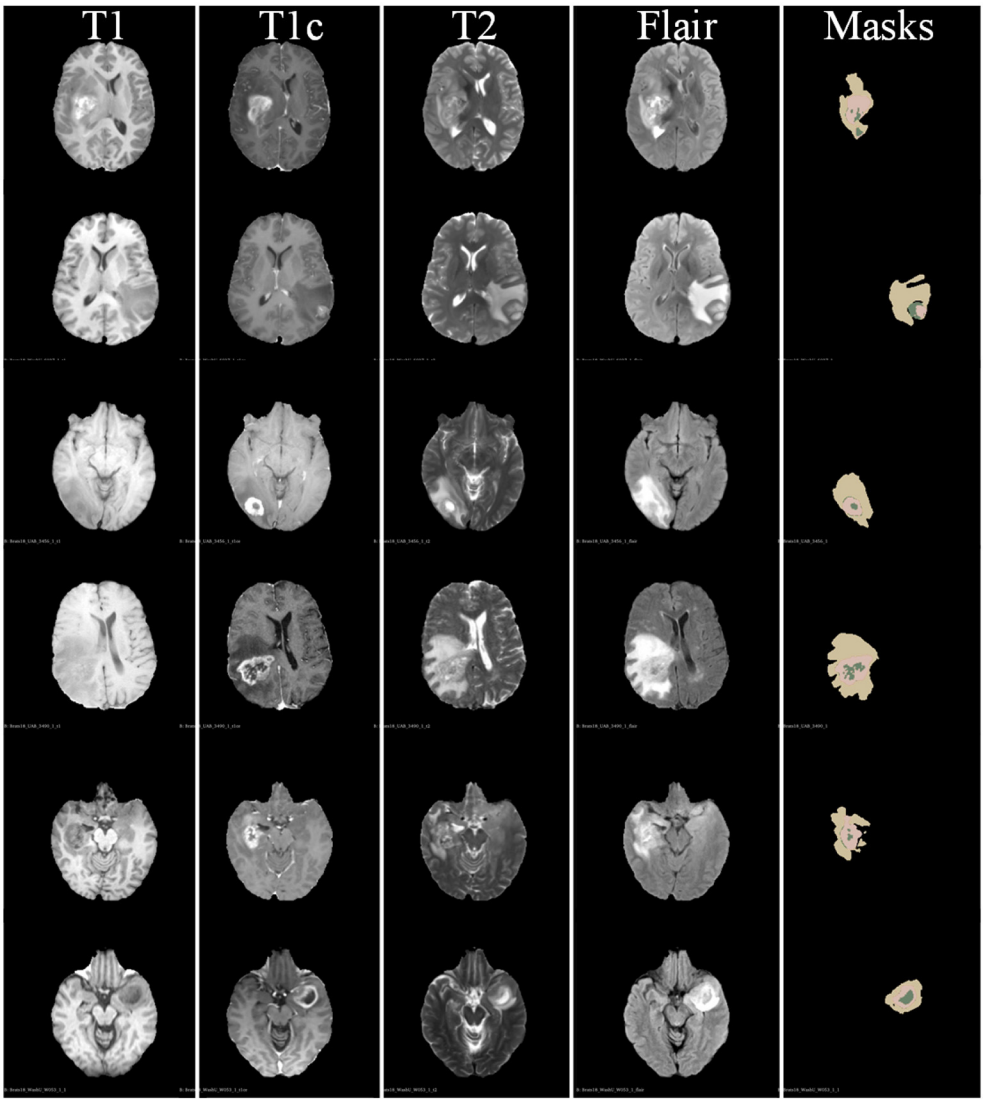

FIGURE 5 | Some examples of segmentation results of our proposed brain tumor segmentation on BraTs 2018 dataset.

Figure 5, we can observe that our method is more able to segment the details of the tumor areas, including TC areas, enhancing TC areas, and WT areas. The quantitative analysis results further illustrate the effectiveness of our proposed segmentation method.

7. CONCLUSION

In this work, we have proposed a novel attention-guided cross-modality feature learning framework for segmenting brain tumor areas from the multi-modality MRI data. Considering the fact that the texture and spatial position of normal organs in medical images have specific structural patterns, and deep neural networks can easily learn discriminant information from such regular patterns, we propose to mine the common normal patterns across the multi-modality data to captures the discriminative features between brain tumor areas and normal brain areas. The proposed learning framework consists of a cross-modality feature extracting module and an attention guided feature fusing module. By building a two-step feature interacting strategy, our proposed feature extracting module explores the multi-modalities interactive features that capture the rich information of the multi-modalities MRI data. The attention-guided feature fusing module encourages the feature extracting module to learn the structure patterns of the normal brain areas and aggregates the cross-modalities features in reasonable manner. Comprehensive experiments are conducted on BraTS 2018 benchmark, which demonstrate the effectiveness of our approach when compared to baseline models and state-of-the-art methods.

DATA AVAILABILITY STATEMENT

The dataset BraTs2018 for this study can be found in the MICCAI Brain Tumor Segmentation Challenge: http://braintumorsegmentation.org/.

AUTHOR CONTRIBUTIONS

BW and ZY contributed to the conception and design of the study. BW implemented the experiments and wrote the first draft of the manuscript. HP and LM contributed to clinical experience for the design of MRI segmentation model. All authors contributed to the result analysis, manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the China Postdoctoral Science Foundation (2019T120945) and Natural Science Basic Research Plan in Shaanxi Province of China (2019JQ-630).

REFERENCES

1. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system. *Appl Soft Comput*. (2020) 10:6897. doi: 10.1016/j.asoc.2020.106897
2. Puch S, Sanchez I, Hernandez A, Piella G, Prcckovska V. Global planar convolutions for improved context aggregation in brain tumor segmentation. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 393–405. doi: 10.1007/978-3-030-11726-9_35
3. Chandra S, Vakalopoulou M, Fidon L, Battistella E, Estienne T, Sun R, et al. Context aware 3D CNNs for brain tumor segmentation. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 299–310. doi: 10.1007/978-3-030-11726-9_27
4. Ma J, Yang X. Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3D lightweight CNNs. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 25–36. doi: 10.1007/978-3-030-11726-9_3
5. Yan Q, Gong D, Zhang Y. Two-stream convolutional networks for blind image quality assessment. *IEEE Trans Image Process*. (2019) 28:2200–11. doi: 10.1109/TIP.2018.2883741
6. Yan Q, Zhang L, Liu Y, Zhu Y, Sun J, Shi Q, et al. Deep HDR imaging via a non-local network. *IEEE Trans Image Process*. (2020) 29:4308–22. doi: 10.1109/TIP.2020.2971346
7. Yan Q, Wang B, Li P, Li X, Zhang A, Shi Q, et al. Ghost removal via channel attention in exposure fusion. *Comput Vis Image Understand*. (2020) 201:10. doi: 10.1016/j.cviu.2020.103079
8. Yan Q, Wang B, Gong D, Luo C, Zhao W, Shen J, et al. COVID-19 chest CT image segmentation-a deep convolutional neural network solution. (2020) *arXiv preprint arXiv:20041(0987)*.
9. Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020). p. 3667–76. doi: 10.1109/CVPR42600.2020.00372
10. Shaikh M, Anand G, Acharya G, Amrutkar A, Alex V, Krishnamurthi G. Brain tumor segmentation using dense fully convolutional neural network. In: *International MICCAI Brainlesion Workshop*. Quebec, Qc: Springer (2017). p. 309–19. doi: 10.1007/978-3-319-75238-9_27
11. Islam M, Ren H. Fully convolutional network with hypercolumn features for brain tumor segmentation. In: *Proceedings of MICCAI Workshop on Multimodal Brain Tumor Segmentation Challenge (BRATS)*. Quebec, QC: (2017).
12. Lopez MM, Ventura J. Dilated convolutions for brain tumor segmentation in MRI scans. In: *International MICCAI Brainlesion Workshop*. Quebec, QC: Springer (2017). p. 253–62. doi: 10.1007/978-3-319-75238-9_22
13. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. (2017) 36:61–78. doi: 10.1016/j.media.2016.10.004
14. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: *International Conference on Information Processing in Medical Imaging*. Boone, NC : Springer (2017). p. 348–60. doi: 10.1007/978-3-319-59050-9_28
15. Castillo LS, Daza LA, Rivera LC, Arbeláez P. Volumetric multimodality neural network for brain tumor segmentation. In: *13th International Conference on Medical Information Processing and Analysis*. Vol. 1. San Andres Island: International Society for Optics and Photonics (2017). p. 105720E.
16. Fidon L, Li W, Garcia-Peraza-Herrera LC, Ekanayake J, Kitchen N, Ourselin S, et al. Scalable multimodal convolutional networks for brain tumour segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Quebec, QC: Springer (2017). p. 285–93. doi: 10.1007/978-3-319-66179-7_33
17. Han J, Yang Y, Zhang D, Huang D, Torre FDL. Weakly-supervised learning of category-specific 3D object shapes. *IEEE Trans Pattern Anal Mach Intell*. (2019) 99:1–1. doi: 10.1109/TPAMI.2019.2949562
18. Wei W, Xu Q, Wang L, Hei X, Shen P, Shi W, et al. GI/Geom/1 queue based on communication model for mesh networks. *Int J Commun Syst*. (2014) 27:3013–29. doi: 10.1002/dac.2522
19. Wei W, Fan X, Song H, Fan X, Yang J. Imperfect information dynamic stackelberg game based resource allocation using hidden Markov for cloud computing. *IEEE Trans Serv Comput*. (2016) 11:78–89. doi: 10.1109/TSC.2016.2528246
20. Wei W, Song H, Li W, Shen P, Vasilakos A. Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. *Inform Sci*. (2017) 408:100–14. doi: 10.1016/j.ins.2017.04.042
21. Bu S, Wang L, Han P, Liu Z, Li K. 3D shape recognition and retrieval based on multi-modality deep learning. *Neurocomputing*. (2017) 259:183–93. doi: 10.1016/j.neucom.2016.06.088
22. Wei W, Su J, Song H, Wang H, Fan X. CDMA-based anti-collision algorithm for EPC global C1 Gen2 systems. *Telecommun Syst*. (2018) 67:63–71. doi: 10.1007/s11235-017-0321-4
23. Wei W, Xia X, Wozniak M, Fan X, Damaševičius R, Li Y. Multi-sink distributed power control algorithm for cyber-physical-systems in coal mine tunnels. *Comput Netw*. (2019) 161:210–9. doi: 10.1016/j.comnet.2019.04.017
24. Wei W, Zhou B, Połap D, Woźniak M. A regional adaptive variational PDE model for computed tomography image reconstruction. *Pattern Recogn*. (2019) 92:64–81. doi: 10.1016/j.patcog.2019.03.009
25. Yao J, Zhu X, Zhu F, Huang J. Deep correlational learning for survival prediction from multi-modality data. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Quebec, QC: Springer (2017). p. 406–14. doi: 10.1007/978-3-319-66185-8_46
26. Xu X, Li Y, Wu G, Luo J. Multi-modal deep feature learning for RGB-D object detection. *Pattern Recogn*. (2017) 72:300–13. doi: 10.1016/j.patcog.2017.07.026
27. Liu X, Ma X, Wang J, Wang H. M3L: Multi-modality mining for metric learning in person re-Identification. *Pattern Recogn*. (2018) 76:650–61. doi: 10.1016/j.patcog.2017.09.041
28. Wang A, Lu J, Cai J, Cham TJ, Wang G. Large-margin multi-modal deep learning for RGB-D object recognition. *IEEE Trans Multim*. (2015) 17:1887–98. doi: 10.1109/TMM.2015.2476655
29. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. (2014) 34:1993–2024. doi: 10.1109/TMI.2014.2377694
30. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with

expert segmentation labels and radiomic features. *Sci Data*. (2017) 4:17. doi: 10.1038/sdata.2017.117
31. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. (2018) *arXiv preprint arXiv:18110(2629)*.
32. Zhang D, Huang G, Zhang Q, Han J, Han J, Yu Y. Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recogn*. (2020) 110:107562. doi: 10.1016/j.patcog.2020.107562
33. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 311–20. doi: 10.1007/978-3-030-11726-9_28
34. Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, et al. PHISEG: Capturing uncertainty in medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Shenzhen: Springer (2019). p. 119–27. doi: 10.1007/978-3-030-32245-8_14
35. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, et al. PraNet: Parallel reverse attention network for polyp segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer (2020). p. 263–73. doi: 10.1007/978-3-030-59725-2_26
36. Yan Q, Gong D, Shi Q, Hengel Avd, Shen C, Reid I, et al. Attention-guided network for ghost-free high dynamic range imaging. (2019) *arXiv preprint arXiv:19041(0293)*. doi: 10.1109/CVPR.2019.00185
37. Yan Q, Gong D, Zhang P, Shi Q, Sun J, Reid I, et al. Multi scale dense networks for deep high dynamic range imaging. In: *IEEE Winter Conference on Applications of Computer Vision*. Waikoloa Village, HI (2019). p. 41–50. doi: 10.1109/WACV.2019.00012
38. Yan Q, Sun J, Li H, Zhu Y, Zhang Y. High dynamic range imaging by sparse representation. *Neurocomputing*. (2017) 269:160–9. doi: 10.1016/j.neucom.2017.03.083
39. Zhang C, Yan Q, Zhu Y, Li X, Sun J, Zhang Y. Attention-based network for low-light image enhancement. In: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. London (2020). p. 1–6. doi: 10.1109/ICME46284.2020.9102774
40. Jungo A, Balsiger F, Reyes M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Front Neurosci*. (2020) 14:282. doi: 10.3389/fnins.2020.00282
41. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich (2018). p. 3–19. doi: 10.1007/978-3-030-01234-2_1
42. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No newnet. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 234–44. doi: 10.1007/978-3-030-11726-9_21
43. Chen W, Liu B, Peng S, Sun J, Qiao X. S3D-UNet: separable 3D U-Net for brain tumor segmentation. In: *International MICCAI Brainlesion Workshop*. Granada: Springer (2018). p. 358–68. doi: 10.1007/978-3-030-11726-9_32

Conflict of Interest: BW and JA are employed by company Beijing Jingzhen Medical Technology Ltd.

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

frontiers in Medicine

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fmed.2021.653913

Accurate Tumor Segmentation via Octave Convolution Neural Network

Bo Wang[1,2,3], *Jingyi Yang*[4]*, *Jingyang Ai*[3], *Nana Luo*[5], *Lihua An*[5], *Haixia Feng*[5], *Bo Yang*[6] *and Zheng You*[1,2]*

[1] The State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing, China, [2] Innovation Center for Future Chips, Tsinghua University, Beijing, China, [3] Beijing Jingzhen Medical Technology Ltd., Beijing, China, [4] School of Artificial Intelligence, Xidian University, Xi'an, China, [5] Affiliated Hospital of Jining Medical University, Jining, China, [6] China Institute of Marine Technology & Economy, Beijing, China

Three-dimensional (3D) liver tumor segmentation from Computed Tomography (CT) images is a prerequisite for computer-aided diagnosis, treatment planning, and monitoring of liver cancer. Despite many years of research, 3D liver tumor segmentation remains a challenging task. In this paper, we propose an effective and efficient method for tumor segmentation in liver CT images using encoder-decoder based octave convolution networks. Compared with other convolution networks utilizing standard convolution for feature extraction, the proposed method utilizes octave convolutions for learning multiple-spatial-frequency features, thus can better capture tumors with varying sizes and shapes. The proposed network takes advantage of a fully convolutional architecture which performs efficient end-to-end learning and inference. More importantly, we introduce a deep supervision mechanism during the learning process to combat potential optimization difficulties, and thus the model can acquire a much faster convergence rate and more powerful discrimination capability. Finally, we integrate octave convolutions into the encoder-decoder architecture of UNet, which can generate high resolution tumor segmentation in one single forward feeding without post-processing steps. Both architectures are trained on a subset of the LiTS (Liver Tumor Segmentation) Challenge. The proposed approach is shown to significantly outperform other networks in terms of various accuracy measures and processing speed.

Keywords: liver, liver tumor, deep learning, octave convolution, segmentation

OPEN ACCESS

Edited by:
Juan Liu,
Huazhong University of Science and Technology, China

Reviewed by:
Hua Zhang,
Hangzhou Dianzi University, China
Yi Liu,
Changzhou University, China

***Correspondence:**
Jingyi Yang
yangjingyi16@stu.xidian.edu.cn
Zheng You
yz-dpi@mail.tsinghua.edu.cn

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *15 January 2021*
Accepted: *24 March 2021*
Published: *19 May 2021*

Citation:
Wang B, Yang J, Ai J, Luo N, An L, Feng H, Yang B and You Z (2021) Accurate Tumor Segmentation via Octave Convolution Neural Network. Front. Med. 8:653913. doi: 10.3389/fmed.2021.653913

1. INTRODUCTION

According to the World Health Organization, liver cancer was the second most common cause of cancer deaths in 2015. Hepatocellular carcinoma (HCC) is the most common primary liver cancer and the sixth most common cancer. Each year, the incidence and death rates of liver cancer are steadily increasing. In addition, the liver is also a common site for secondary tumors. It is an important factor leading to human death. With the rapid development of tumor radiation technology, radiotherapy has entered the stage of precision radiotherapy represented by image guidance and adaptive radiotherapy. Precision radiotherapy needs to accurately delineate the target area (tumor) of radiotherapy to guide treatment and subsequent radiation plans. But at this stage, accurate target area delineation in clinical medicine needs to be done manually by experienced physicians, and its accuracy and efficiency completely depend on the physician's clinical experience. This work is not only time-consuming, but also poorly reproducible.

Using computer image processing technology, combined with medical imaging diagnostic technology, early diagnosis, three-dimensional modeling, and quantitative analysis of liver diseases can enable doctors to have sufficient data before surgery, make preoperative planning, improve the success rate of surgery, and make reasonable preparations for an effective treatment plan. The accurate and reliable segmentation of liver contours from abdominal CT images is the first step in the early diagnosis of liver disease, the estimation of liver size and condition, and three-dimensional modeling. It is also a very critical step. The segmentation results have a direct effect on subsequent work. In actual clinical applications, the liver contour is usually manually segmented from CT images by physicians with relevant practical experience and professional knowledge. However, this process is very time-consuming and energy-consuming, and is subject to the subjective factors, experience, and knowledge of different physicians. The effect of the difference will often result in different segmentation results. Therefore, in order to reduce the workload of doctors, improve work efficiency, and obtain more objective and accurate segmentation results, computer-aided diagnosis technology must be introduced to help professional doctors segment liver CT images.

To solve this problem, researchers have invested in the research and come up with a number of approaches. Over the past few decades, they have focused on developing algorithms such as level sets, watershed, statistical shape models, region growth, active contour models, threshold processing, pattern cutting, and traditional machine learning methods that require manual extraction of tumor features.

Traditional liver segmentation methods are based on image processing methods, and mainly rely on some shallow features of the image, such as grayscale, statistical structure, and texture to segment liver contours. This feature can be obtained directly from the image or obtained by artificially designed extraction operators. These shallow features are less robust, not representative, and susceptible to noise interference. Practice has proved that it is often those abstract and deep features that are more representative. Deep learning technology can mine the deep abstract features of data from a large amount of data and apply them to liver segmentation tasks to improve the accuracy and robustness of segmentation.

Region growing, thresholding, or clustering methods have been widely used in medical image segmentation because they are fast, easy to implement, and have relatively low computational costs. However, the main drawback of these methods is that they use only strength information. As a result, this method is prone to boundary leakage at blurred tumor boundaries. Therefore, prior knowledge or other algorithms are integrated to reduce under-segmentation or over-segmentation (1–3). Anter et al. (1) present an automatic tumor segmentation method using adaptive region growth. A marker-controlled watershed algorithm was used to detect the initial seed points of regional growth. Yan et al. (4) present a semi-automatic segmentation method based on watershed transformation. They first manually placed seed points in the tumor area as markers, and then performed watershed transformation to delineate and extract tumor contours in the image. Therefore, the density information of the tumor can be obtained as a threshold to separate the hepatic lesion from its adjacent tissues. Then, the threshold is refined from the segmented lesion to obtain accurate results. DAS and Sabut (3) used adaptive thresholding, morphological processing, and nucleated fuzzy C-means (FCM) algorithms to segment liver tumors from CT images. Moghbel et al. (5) present an automatic tumor segmentation scheme based on supervised random Walker method. FCM with the function of cuckoo optimization is used for PIXEL marking of final random Walker segmentation.

Active contour methods, such as fast moving and level set algorithms, are popular segmentation techniques. However, good initialization and velocity function are needed to obtain accurate segmentation results, especially for tumors with uneven intensity and weak boundaries. Li et al. (6) present a new level set model that combines edge- and region-based information with prior information. An FCM algorithm is used to estimate the probability of tumor tissue. Li et al. (6) present a semi-automatic method for segmentation of liver tumors from magnetic resonance (MR) images, which uses a fast-moving algorithm to generate initial labeled regions and then classifies other unlabeled voxels through a neural network. A graph cutting method has also been widely used in medical image segmentation (7, 8), which can achieve global optimization solutions. Stawiaski et al. (7) present an interactive segmentation method based on watershed and graph cutting. When held in conjunction with the 2008 Liver Tumor Segmentation Challenge (LTSC08) competition [in conjunction with the 2008 Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference], the method achieved the highest accuracy compared to other semi-automatic or automated methods. Linguraru et al. (8) present an automatic pattern segmentation method that uses pattern cutting with Hessian-based shape constraints to bias speckle-like tumors. However, the main drawback of such techniques based on level sets or graphic cuttings is their high computational cost, especially for 3D volume data.

Recently, deep learning (9–21) has penetrated into a variety of applications and surpassed the state-of-the-art performance in many fields such as image detection, classification, and segmentation (22–26), which also excites us to use this technique in the liver tumor segmentation task. Many researchers have already used deep learning methods to explore the task of liver tumor segmentation. In practical applications, CNN shows excellent feature extraction capabilities. Among them, fully convolutional neural networks (FCN) as an improved network of CNN have been widely used in the field of image segmentation. Different from image classification, semantic segmentation needs to determine the category of each pixel to achieve accurate segmentation. FCN replaces the last fully connected layer of CNN with a deconvolution layer to achieve pixel-to-pixel classification. The application of FCN and its derivatives in image segmentation continues to expand. Its encoder is the same as the 13 convolutional layers in VGG-16. The decoder maps the features extracted by the encoder to the encoder with the same resolution as the input. When the feature is extracted from small to small, the decoder gradually enlarges the extracted feature to the size of the input image from small to large. However,

the traditional FCN network has poor edge segmentation and low accuracy, which cannot meet the requirements of medical image segmentation. Li et al. (6) propose a H-DensU-Net, which consists of 2D and 3D U-Net, for the segmentation of liver tumors. 2D U-Net is used to extract tumor features in individual sections, while 3D U-Net is used to understand tumor spatial information between sections. Sun et al. (27) present a method of liver tumor segmentation based on multi-channel full convolutional network (MC-FCN). They designed an MC-FCN to train contrast-enhanced CT images at different imaging stages, because each stage of the data provides unique information about the pathological features of the tumor. However, these neural networks are fully connected between adjacent layers, which leads to problems such as over-parameterization and over-fitting for tumor segmentation tasks. In addition, the number of trainable parameters in a fully connected neural network is related to the size of the input image, which results in higher computational costs when processing high-resolution images.

One of the challenges of deep learning for medical image processing is that the samples provided are often relatively small, and U-Net still performs well under this limitation. As an image semantic segmentation network, U-Net was mainly used to process medical images when it was proposed. The U-Net network is a CNN-based image segmentation network, mainly used for medical image segmentation. When it was first proposed, it was used for cell wall segmentation. Later, it has excellent performance in lung nodule detection and blood vessel extraction on the fundus retina. Including the CT image segmentation of liver tumor lesions. In specific implementation, this type of method can use deep features to locate liver tissue regions and use shallow features to achieve accurate segmentation results. Many medical image segmentation problems are improved based on U-Net. According to the adopted form of U-Net network architecture, it can be divided into single network liver tumor segmentation method, multi-network liver tumor segmentation method, and u. A liver tumor segmentation method combining Net network and traditional methods. Regardless of the calculation and memory performance, the 3D network can combine the image layer information to ensure a change continuity between the interlayer image masks, and the segmentation effect is better than 2D.

Considering clinical suitability and segmentation accuracy as well as processing time, our goal is to develop an efficient, robust, and accurate method for tumor segmentation. Therefore, in this paper, a deep learning method based on learning and decoding layered features with multiple spatial frequencies is proposed to achieve 3D liver tumor segmentation from CT images. The main contributions of this work are three-fold:

- Due to observe the CT liver tumor image can be decomposed to describe the structure of the smooth change (such as the shape of the tumor) mutations in the low spatial frequency components and describe the details of (the edge of the tumor, for example) the high spatial frequency components, so we use the octave convolution (28) encoder block for building characteristics, and use them to study neural network layered multiple frequency characteristics of multiple levels.
- We propose to decompose the convolution feature graph into two groups at different spatial frequencies and process them with different extended convolution at their corresponding frequencies (one octave apart). Storage and computation can be saved because the resolution of low frequency graphs can be reduced. This also helps each layer to have a larger receive field to capture more contextual information. Importantly, the proposed blocks are fast in practice and can reach speeds close to the theoretical limit.
- More importantly, we introduce deep supervision to the hidden layer, which can accelerate the optimization convergence speed and improve the prediction accuracy.
- In addition, the proposed network is superior to the benchmark U-Net in terms of segmentation performance and computing overhead, while achieving better or comparable performance to the latest approach on open data sets.

2. METHODS

U-Net is modified on the basis of the existing CNN structure for classification, that is, the original fully connected layer of CNN is changed into a convolutional layer. FCN is composed of convolution and deconvolution. Through the process of convolution and deconvolution, based on end-to-end learning, the classification of each pixel of the image is completed, thereby realizing the segmentation of the entire input image. U-Net realizes the semantic segmentation of images through an end-to-end network structure. The end-to-end network can reduce manual preprocessing and subsequent processing and make the model from the original input to the final output as much as possible. The network learns the features by itself, and the extracted features are also integrated into the algorithm. The network model can be automatically adjusted according to the data, thereby increasing the overall fit of the model, and the cost of end-to-end network learning is lower than that of non-end-to-end network structure.

2.1. Encoder Part

The liver tumors often have varying sizes and shapes. The low- and high- frequency components of tumors focus on capturing the style of tumor and edge information, respectively. Motivated by this observation, we hypothesize that adopting a multi-frequency feature learning approach may be beneficial for segmenting the tumor from liver CT images. Therefore, the octave convolution (28) is adopted as an extractor for multifrequency features in this work. The computational graph for multifrequency feature transformations of the octave convolution is illustrated in **Figure 1**. Let X^H and X^L denote the inputs of high- and low- frequency feature maps, respectively. The high- and low-frequency outputs of the octave convolution are given by $\hat{Y}^H = f^{H\to H}(X^H) + f^{L\to H}(X^L)$ and $\hat{Y}^L = f^{L\to L}(X^L) + f^{H\to L}(X^H)$, where $f^{H\to H}$ and $f^{L\to L}$ denote two standard convolution operations for intra-frequency information update, whereas $f^{H\to L}$ and $f^{L\to H}$ denote the process of inter-frequency information exchange. Specifically, $f^{H\to L}$ is equivalent to first down-sampling the input by average-pooling with a scale of two and then applying a standard convolution for feature

transformation, and $f^{L\to H}$ is equivalent to up-sampling the output of a standard convolution by nearest interpolation with a scale of two.

To calculate these items, working (28) splits the convolution kernel Winto two components $W = [W^H, W^L]$is responsible for convolved with $X^H$and X^L. Each component can be further divided into in-frequency and in-frequency parts: $W^H = [W^{H\to H}, W^{L\to H}]$and $W^L = [W^{L\to L}, W^{H\to L}]$, whose parameter tensor shape is shown in **Figure 2**. Especially for the high-frequency feature graph, we use A to calculate its regular convolution for in-frequency update at the position $(p, q)$and for inter-frequency communication. We can fold the up-sampling of the feature tensor X^Linto convolution without explicit calculation and storage of the up-sampling function as follows:

$$\begin{aligned} Y^H_{p,q} &= Y^{H\to H}_{p,q} + Y^{L\to H}_{p,q} \\ &= \sum_{i,j\in\mathcal{N}_k} W^{H\to H}_{i+\frac{k-1}{2},j+\frac{k-1}{2}} X^H_{p+i,q+j} \\ &\quad + \sum_{i,j\in\mathcal{N}_k} W^{L\to H}_{i+\frac{k-1}{2},j+\frac{k-1}{2}} X^L_{(\lfloor\frac{p}{2}\rfloor+i),(\lfloor\frac{q}{2}\rfloor+j)} \end{aligned} \tag{1}$$

where $\lfloor\cdot\rfloor$represents a lower bound operation. Similarly, for low-frequency characteristic graphs, we use regular convolution to calculate in-frequency update. Note that since the graph is an octave lower, the convolution is also low frequency W.R.T. High frequency coordinate space. For inter-frequency communication, we can fold the subsample of the feature tensor X^Hinto the convolution again, as shown below:

$$\begin{aligned} Y^L_{p,q} &= Y^{L\to L}_{p,q} + Y^{H\to L}_{p,q} \\ &= \sum_{i,j\in\mathcal{N}_k} W^{\top}_{i+\frac{k-1}{2},j+\frac{k-1}{2}} X^L_{p+i,q+j} \\ &\quad + \sum_{i,j\in\mathcal{N}_k} W^{H\to L}_{i+\frac{k-1}{2},j+\frac{k-1}{2}} X^{\top}_{(2*p+0.5+i),(2*q+0.5+j)} \end{aligned} \tag{2}$$

where multiplying a factor 2 to the locations (p, q) performs down-sampling, and further shifting the location by a half step is to ensure the down-sampled maps are well-aligned with the input.

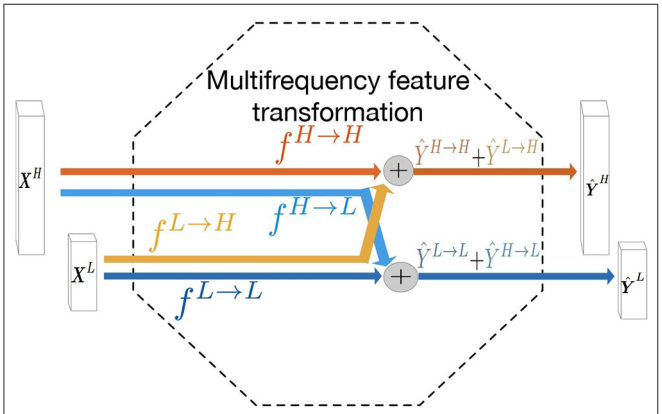

FIGURE 1 | Computation graph of the multifrequency feature transformation of octave convolution. The operation mainly contains two processes of the inter-frequency information exchange ($f^{L\to H}$ and $f^{H\to L}$) and intra-frequency information update ($f^{L\to L}$ and $f^{H\to H}$).

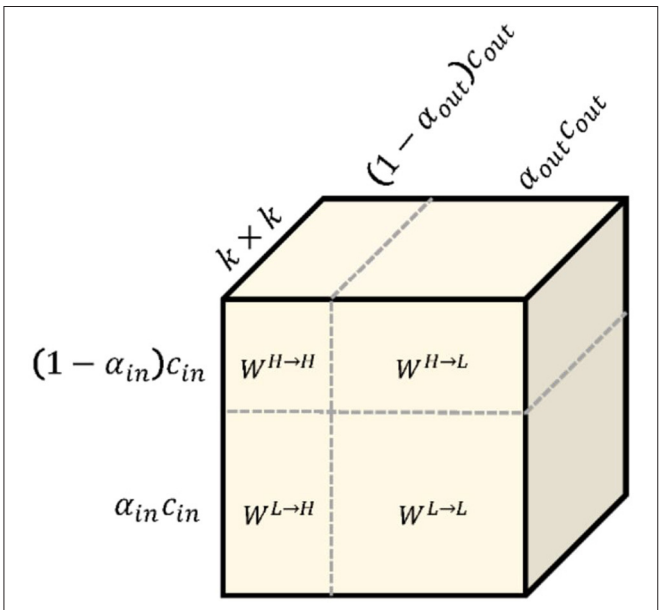

FIGURE 2 | The octave convolution kernel. $k \times k$ octave convolution kernel W is equivalent to vanilla convolution kernel because they have exactly the same number of parameters.

2.2. Decoder Part

Deconvolution is a convolution operation, which is the inverse process of pooling. In U-Net, the pooling operation reduces the size of the input picture, but in the image segmentation process, each pixel needs to be classified, and finally a segmented image with the same dimension as the input picture is obtained. Therefore, the generated heat map (heat map) is restored to the original image dimensions. Through reverse training, deconvolution can achieve the effect of output reconstruction and input, so that the output image can be restored to the same dimension as the input image.

On the one hand, in the process of feature coding as shown in the **Figure 3**, although the spatial size of the feature graph gradually decreases, the feature graph gradually loses spatial details. This compression effect forces the kernel to learn more discriminations with higher levels of abstraction. On the other hand, multi-frequency feature extraction alone is not sufficient to perform dense pixel classification for liver tumor segmentation. A process is needed to decode the feature map to recover spatial detail and generate a high-resolution probabilistic map of the tumor. A simple way to do this is to use bilinear interpolation, which unfortunately lacks the ability to learn the decoding transformation that transpose convolution has. Therefore, we choose the transpose convolution to up-sample the feature.

2.3. The Proposed Network

In this section, a novel encoder-decoder based neural network architecture called **OCunet** is proposed. After end-to-end training, the proposed OCunet is able to extract and decode layered multifrequency features for the segmentation of liver tumors in full-size CT images. The computational pipeline of OCunet consists of two main processes, namely feature encoding and decoding. By using octave convolution, we design multi-frequency feature encoder block and decoder block for hierarchical multi-frequency feature learning and decoding. By sequentially stacking multiple encoder blocks (as shown in **Figure 4**), layered multifrequency features can learn to capture details of the low frequency components that describe smooth changes in the structure (such as the main blood vessels) and the high frequency components that describe details of sudden changes (including the fine components), as shown in **Figure 3**.

2.4. Loss Function

The learning of the 3D network is formulated as a problem of minimizing the per-pixel binary classification error relative to the ground mask, but the optimization process is challenging. A major problem is the disappearance of the gradient, which makes the loss back propagation ineffective in the early layers. This problem is likely to be more serious in 3D and will inevitably slow down the convergence rate and the discriminating ability of the model. To address this challenge, we used additional monitoring injected into some hidden layers to counteract the negative effects of gradient disappearance. Specifically, we used an additional deconvolution layer to amplify some of the lower- and mid-level feature quantities, and then used the Softmax layer to obtain dense predictions for calculating classification errors. Using the gradient obtained from the prediction of these branches and the last output layer, the effect of gradient disappearance can be effectively mitigated.

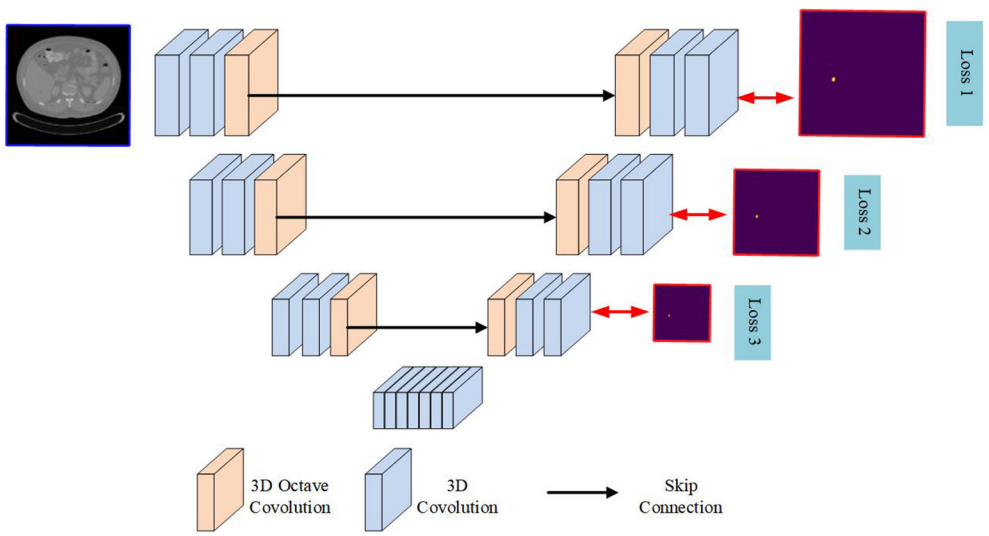

FIGURE 3 | Detailed network architecture of the proposed network.

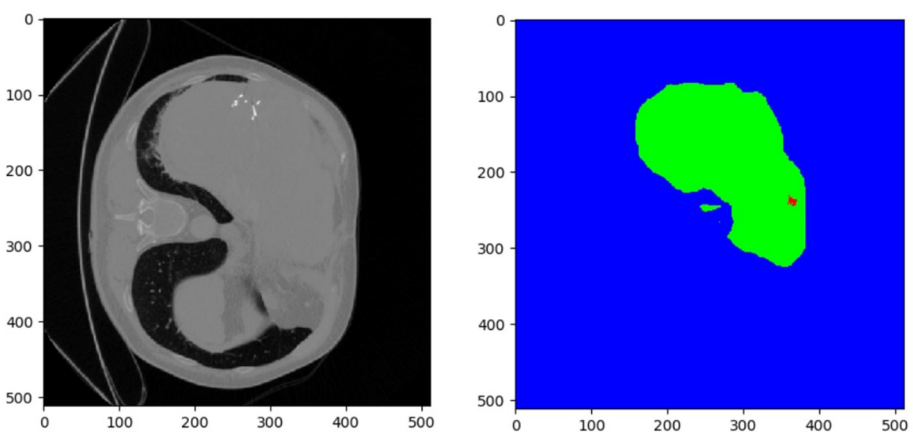

FIGURE 4 | Training data provided by LiTS-challenge.

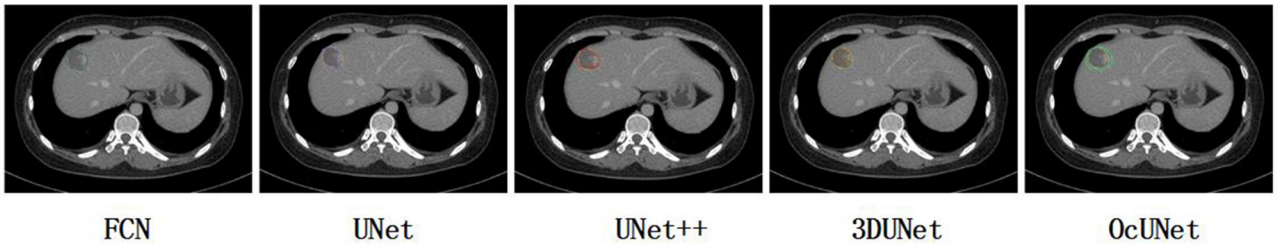

FIGURE 5 | Example of tumor segmentation results from a testing image.

Since the number of voxels belonging to the foreground is much smaller than the number belonging to the background (i.e., the liver), this problem of data imbalance usually leads to a prediction bias when using traditional loss functions. In order to solve this problem, the loss function, Dice coefficient (DICE), which represents the similarity measure between the ground truth and the predicted score graph, is proposed.

3. EXPERIMENTS

3.1. Datasets

The LiTs dataset[1] includes 130 contrast-enhanced 3D abdominal CT scan images from 6 different clinical sites, of which 130 cases are used for training and the remaining 70 are used for testing. The CT scan is accompanied by reference annotations of the liver and tumors made by a trained radiologist. The data set contains 908 lesions. The data set has significant differences in image quality, spatial resolution, and vision. The in-plane resolution is $0.6 \times 0.6mm$-$1.0 \times 1.0mm$, slice thickness (layer spacing) is 0.45–6.0 mm, the axial slice size of all scans is fixed at 512×512 pixels, but the number of slices per scan It ranges from 42 to 1,026 sheets.

Further test data were provided by the Radiology Centre of the Medical University of Innsbruck. The data set contains CT scans of patients with liver cancer, with reference notes drawn up by medical scientists. Because deep learning methods can achieve better performance if the data has a consistent size or distribution, all data is normalized to strength values between [0,1] before starting optimization.

3.2. Implementation Details

Our OCunet was implemented with PyTorch library. We trained the network from scratch with weights initialized from Gaussian distribution. The learning rate was initialized as 0.1 and divided by 10 every 1,000 epochs. Each training epoch took around 2 min using a GPU of NVIDIA GTX 2080Ti.

3.3. Metrics

(1) Precision: Precision, or the positive predictive value, refers to the fraction of relevant instances among the total retrieved instances.

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

(2) Recall: Recall, also known as sensitivity, refers to the fraction of relevant instances retrieved over the total amount of relevant instances.

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

(3) Accuracy: Accuracy refers to the fraction of relevant instances among the total instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5}$$

(4) Specificity: Accuracy refers to the fraction of retrieved instances among the total amount of relevant instances.

$$Specificity = \frac{TN}{FP + TN}. \tag{6}$$

(5) DICE Score: also called the overlap index, is the most commonly used index to verify the segmentation of medical images, and it usually represents the repetition rate between the segmentation result and the mark. The value range of DICE is 0 1, 0 means real. The experimental segmentation result and the labeling result deviate seriously, and 1 means that the experimental segmentation result and the labeling result completely coincide. It is defined as follows:

$$Dic(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{7}$$

where A is the estimated maps, B denotes the ground truth, $|A \cap B|$ represents the number of pixels common to both images. The higher value of the dice coefficient denotes the better segmentation accuracy.

[1]https://competitions.codalab.org/competitions/17094

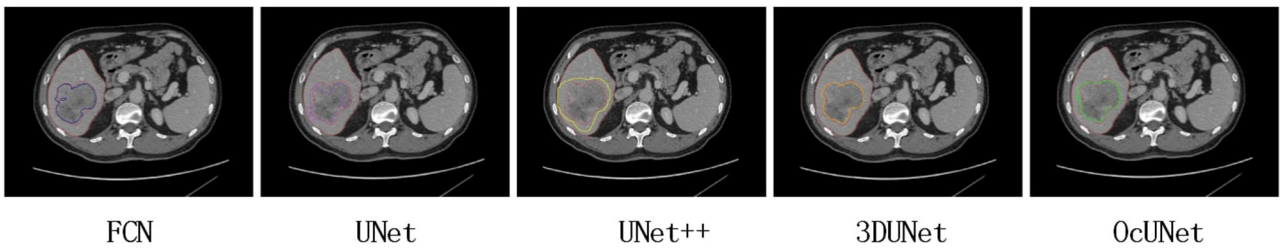

FIGURE 6 | Compared results of tumor segmentation with different methods.

3.4. Evaluation on Test Data

Figures 5, **6** show tumor segmentation results from training and test images, respectively. The red is a liver tumor. We compare the basic facts with the results generated by FCN, U-Net, UNet++, and 3DU-Net. In order to visualize the simplicity of the results caused by network differences, here we train the network only on the axial plane. In **Figure 5**, by FCN, U-Net, UNet++, and 3DU-Net provide results showed in the first, second, third, and fourth columns, we can see that in the FCN and U-Net segmentation results, residual connection can distinguish to some extent of tumor, but will miss part should belong to the tumor tissue. In UNet++ more accurate segmentation results can be predicted through intense connection, but compared with the result of a split, a split less than 3DU-Net still exists, thanks to combat training strategy, and can recognize more voxels belonging to the tumor. In the **Figure 6**, the results obtained from the test image show a similar appearance to the training image. However, it can be seen that liver tumors produced by 3DUNet are segmented more accurately. Although the segmentation results provided by 3DU-Net still have some unsegmented tumor tissue, it has been significantly improved compared to the other two methods, demonstrating the effectiveness of the algorithm. The quantitative results are reported in **Table 1**.

TABLE 1 | Comparing different methods with the proposed dataset on the liver tumor segmentation task.

Metrics	FCN	U-Net	UNet++	3DU-Net	3D Attention	OCunet
Precision	0.872	0.896	0.901	0.914	0.926	0.939
Recall	0.923	0.930	0.931	0.925	0.951	0.962
Accuracy	0.912	0.930	0.942	0.951	0.956	0.959
Specificity	0.909	0.917	0.918	0.957	0.966	0.967
DICE	0.923	0.942	0.945	0.958	0.961	0.963

TABLE 2 | Ablation study results.

Metrics	Precision	Recall	Accuracy	Specificity	DICE
w/o Octave Conv.	0.921	0.939	0.938	0.942	0.947
w Octave Conv.	0.928	0.946	0.944	0.950	0.951
Add 1 Loss	0.930	0.951	0.948	0.958	0.957
Add 2 Loss	0.936	0.959	0.952	0.962	0.960
OCunet	0.939	0.962	0.959	0.967	0.963

3.5. Ablation Study

In this section, we conduct experiments to investigate the effectiveness of different modules of our model. Starting from our baseline, we gradually inject our modifications on the whole structure. The results are summarized in **Table 2**, from which we can see that octave convolution is an effective block for liver tumor segmentation. In addition, we can find that the deep supervision can promote the performance the proposed method.

4. CONCLUSION

In this work, we propose a new network for segmentation of liver tumors. We solve the problem of reducing the extensive spatial redundancy in the original CNN model, and propose a novel Octave convolution operation to store and process the low frequency and high frequency features respectively to improve the model efficiency. In addition to octave convolution, the well-designed OCunet can also extract layered features with multiple spatial frequencies and reconstruct accurate tumor segmentation. Thanks to the design of layered multi-frequency features, OCunet is superior to the baseline model in terms of segmentation performance and computational overhead. A large number of experiments show that the proposed method based on octave convolution converges quickly and can produce high quality segmentation results.

At present, the development direction of deep learning in liver tumor segmentation is mainly concentrated in the following points: (1) The training of deep learning algorithms needs to rely on a large number of data sets, and due to its particularity and sensitivity, medical images need to be manually obtained and labeled by experts. The process is very time-consuming. Therefore, it is not only necessary for medical providers to provide more data support, but also to adopt enhanced methods for the data set to increase the size of the data set. The use of three-dimensional neural network and network deepening is a future research direction of this field; (2) The use of multi-modal liver images for segmentation and the combination of multiple different deep neural networks to extract deeper image information and improve the accuracy of liver tumor segmentation are also a major research direction in this field; (3) Currently most medical image segmentation uses supervised deep learning algorithms. However, for some rare diseases that lack a large amount of data support, supervised deep learning algorithms cannot exert their performance. To

overcome the lack of data for the available problems, some researchers will transfer the supervised field to the semi-supervised or unsupervised field. For example, the GAN network is proposed. Combining the GAN network with other higher-performance networks, further research can be carried out in the future.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

BW and JY: writing. LA, BY, and ZY: supervised. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Science Foundation of China under Grant 61806167, 61936007 and U1801265, the China Postdoctoral Science Foundation (2019T120945), Natural Science Basic Research Plan in Shaanxi Province of China (2019JQ-630), and research funds for the interdisciplinary subject, NWPU.

REFERENCES

1. Anter AM, Azar AT, Hassanien AE, El-Bendary N, ElSoud MA. Automatic computer aided segmentation for liver and hepatic lesions using hybrid segmentations techniques. In: *2013 Federated Conference on Computer Science and Information Systems*. Los Angeles, CA: IEEE (2013). p. 193–8.
2. Zhou JY, Wong DW, Ding F, Venkatesh SK, Tian Q, Qi YY, et al. Liver tumour segmentation using contrast-enhanced multi-detector CT data: performance benchmarking of three semiautomated methods. *Eur Radiol*. (2010) 20:1738–48. doi: 10.1007/s00330-010-1712-z
3. Das A, Sabut SK. Kernelized fuzzy C-means clustering with adaptive thresholding for segmenting liver tumors. *Proc Comput Sci*. (2016) 92:389–95. doi: 10.1016/j.procs.2016.07.395
4. Yan J, Schwartz LH, Zhao B. Semiautomatic segmentation of liver metastases on volumetric CT images. *Med Phys*. (2015) 42:6283–93. doi: 10.1118/1.4932365
5. Moghbel M, Mashohor S, Mahmud R, Saripan MIB. Automatic liver tumor segmentation on computed tomography for patient treatment planning and monitoring. *EXCLI J arXiv [Preprint]*. (2016) 15:406.
6. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans Med Imaging*. (2018) 37:2663–74. doi: 10.1109/TMI.2018.2845918
7. Stawiaski J, Decenciere E, Bidault F. Interactive liver tumor segmentation using graph-cuts and watershed. In: *11th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2008)*. Los Angeles, CA (2008).
8. Linguraru MG, Richbourg WJ, Liu J, Watt JM, Pamulapati V, Wang S, et al. Tumor burden analysis on computed tomography by automated liver and tumor segmentation. *IEEE Trans Med Imaging*. (2012) 31:1965–76. doi: 10.1109/TMI.2012.2211887
9. Yan Q, Sun J, Su S, Zhu Y, Li H, Zhang Y. Blind image quality assessment via deep recursive convolutional network with skip connection. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Los Angeles, CA: Springer (2018). p. 51–61. doi: 10.1007/978-3-030-03335-4_5
10. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. *Appl Soft Comput*. (2020) 11:106897. doi: 10.1016/j.asoc.2020.106897
11. Zhang C, Yan Q, Zhu Y, Li X, Sun J, Zhang Y. Attention-based network for low-light image enhancement. In: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. Los Angeles, CA: IEEE (2020). p. 1–6. doi: 10.1109/ICME46284.2020.9102774
12. Yan Q, Gong D, Zhang Y. Two-stream convolutional networks for blind image quality assessment. *IEEE Trans Image Process*. (2019) 28:2200–11. doi: 10.1109/TIP.2018.2883741
13. Han J, Yang Y, Zhang D, Huang D, Torre FDL. Weakly-supervised learning of category-specific 3D object shapes. *IEEE Trans Pattern Anal Mach Intell*. (2019) 5:15–23.
14. Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020). p. 3667–76. doi: 10.1109/CVPR42600.2020.00372
15. Yan Q, Gong D, Shi Q, Hengel AVD, Shen C, Reid I, et al. Attention-guided network for ghost-free high dynamic range imaging. *arXiv preprint arXiv:190410293*. (2019). doi: 10.1109/CVPR.2019.00185
16. Yan Q, Gong D, Zhang P, Shi Q, Sun J, Reid I, et al. Multi-scale dense networks for deep high dynamic range imaging. In: *IEEE Winter Conference on Applications of Computer Vision*. Los Angeles, CA (2019). p. 41–50. doi: 10.1109/WACV.2019.00012
17. Yan Q, Wang B, Zhang W, Luo C, Xu W, Xu Z, et al. An attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation. *IEEE J Biomed Health Inform*. (2020) 3:113–23. doi: 10.1109/JBHI.2020.3042069
18. Yan Q, Wang B, Zhang L, Zhang J, You Z, Shi Q, et al. Towards accurate HDR imaging with learning generator constraints. *Neurocomputing*. (2020) 7:23–8. doi: 10.1016/j.neucom.2020.11.056
19. Yan Q, Wang B, Gong D, Luo C, Zhao W, Shen J, et al. COVID-19 chest CT image segmentation-a deep convolutional neural network solution. *arXiv preprint arXiv:200410987*. (2020).
20. Yan Q, Wang B, Li P, Li X, Zhang A, Shi Q, et al. Ghost removal via channel attention in exposure fusion. *Comput Vis Image Understand*. (2020) 201:103079. doi: 10.1016/j.cviu.2020.103079
21. Yan Q, Zhang L, Liu Y, Zhu Y, Sun J, Shi Q, et al. Deep HDR imaging via a non-local network. *IEEE Trans Image Process*. (2020) 29:4308–22. doi: 10.1109/TIP.2020.2971346
22. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. (2017) 60:84–90. doi: 10.1145/3065386
23. Xu Y, Du J, Dai LR, Lee CH. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process Lett*. (2013) 21:65–8. doi: 10.1109/LSP.2013.2291240
24. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans Med Imaging*. (2016) 36:994–1004. doi: 10.1109/TMI.2016.2642839
25. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset

characteristics and transfer learning. *IEEE Trans Med Imaging*. (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162
26. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. (2017) 35:18–31. doi: 10.1016/j.media.2016.05.004
27. Sun C, Guo S, Zhang H, Li J, Chen M, Ma S, et al. Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs. *Artif Intell Med*. (2017) 83:58–66. doi: 10.1016/j.artmed.2017.03.008
28. Chen Y, Fan H, Xu B, Yan Z, Kalantidis Y, Rohrbach M, et al. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In: *Proceedings of the IEEE International Conference on Computer Vision*. Los Angeles, CA (2019). p. 3435–44. doi: 10.1109/ICCV.2019.00353

Conflict of Interest: JA was employed by company Beijing jingzhen Medical Technology Ltd.

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

frontiers in Medicine

OPINION
published: 28 May 2021
doi: 10.3389/fmed.2021.655298

Opinions on Computer Audition for Bowel Sounds Analysis in Intestinal Obstruction: Opportunities and Challenges From a Clinical Point of View

*Zhu Yang, Luming Huang, Jingsun Jiang, Bing Hu, Chengwei Tang and Jing Li**

Department of Gastroenterology, West China Hospital, Sichuan University, Chengdu, China

Keywords: intestinal obstruction, bowel sounds, computer audition, machine learning, deep learning

INTRODUCTION

Intestinal obstruction (IO) is a common acute abdominal disease with abdominal pain, distension, vomiting, and constipation. IO, especially mechanical IO (MIO), may need emergency surgery, since it possibly has high morbidity and mortality in cases of perforation, intestinal fistula, peritonitis, etc. Until now, the diagnosis of IO has still relied on abdominal imaging examination (computerized tomography, X-ray), but its application is limited by its radiation, high cost, and requirement for expensive, large equipment as well as professional technicians.

Bowel sound (BS) auscultation is not only safe and effective but also non-invasive for the diagnosis of IO. However, BS has strong subjectivity and randomness, and susceptibility to noises. Doctors have poor accuracy for BS auscultation, which was only 84.5% in normal people and even lower in patients with IO (only 70–80%) (1).

Notably, the etiology, location, and severity of all IO patients cannot be determined depending on imaging examination, symptoms and signs, and traditional auscultation of BS. As a result, intestinal ischemic necrosis or intestinal fistula may not be found in some patients until surgery, which possibly leads to delayed diagnosis and appropriate treatment with increasing the incidence of complications. Computer audition (CA), including machine learning (ML) and deep learning (DL), deals with the complex problem of understanding and analyzing sounds, such as heart sound, lung sound, and BS (2). Characteristics of BS can be automatically extracted and analyzed by powerful ML and DL. That might provide a new way to solve the above problems.

This opinion article aims to highlight the opportunities and challenges of CA for BS analysis in IO.

CHARACTERISTICS OF BS

Different from heart sounds and breath sounds, there is no standard definition or classification of BS at present. This may be due to difference in duration, location, sensor of BS acquisition, and inconsistent acoustic characteristics used for classification.

In 1975, Dalle et al. used computers to analyze BS for the first time and divided BS into three types by using their duration as the classification index (3). Recently, according to duration, frequency, waveform, auditory perception, and mechanisms for the production, Du et al. classified BS as a single burst, multiple bursts, continuous random sound, harmonic sound, and a combination sound (4). When it comes to characteristics of BS in IO, there are few studies. Unfortunately, the available study showed that auscultation of BS was non-specific for diagnosing IO since there was no significant difference in sound-to-sound interval, dominant

Edited by:
Kun Qian,
The University of Tokyo, Japan

Reviewed by:
Jian Guo,
RIKEN Center for Computational Science, Japan
Meishu Song,
University of Augsburg, Germany

****Correspondence:***
Jing Li
melody224@163.com

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *18 January 2021*
Accepted: *22 April 2021*
Published: *28 May 2021*

Citation:
Yang Z, Huang L, Jiang J, Hu B, Tang C and Li J (2021) Opinions on Computer Audition for Bowel Sounds Analysis in Intestinal Obstruction: Opportunities and Challenges From a Clinical Point of View. Front. Med. 8:655298. doi: 10.3389/fmed.2021.655298

frequency, and peak frequency between patients with IO and those without IO (5). In the near future, CA may help us reveal more information about BS in IO.

CA FOR BS ANALYSIS IN IO

Recently, CA has greatly facilitated BS analysis, including the following stages, acoustic sensor, the advanced digital signal processing, ML, and DL (2). DL has brought about breakthroughs in processing images, videos, and audios. The pivotal aspect of DL is that the features of signals can be learned from data using a general-purpose learning procedure. Although its application in audio recognition is still in the initial stage, it shows great advantages in terms of non-invasiveness and big data processing ability. Domestic and foreign scholars have established a DL model based on auscultation data of heart sounds and breath sounds for rapid identification of COVID-19 and its real-time and remote diagnosis (6). Unfortunately, there are few studies on DL for BS analysis in IO. Wang used back-propagation neural networks (BPNNs) for BS analysis based on spectral frequency. It was found that BPNNs may have the ability to recognize BS, with possible applications in digestive function evaluation, recovery monitoring after operation, and auxiliary diagnosis of bowel problems (7). The application of BPNNs in IO remains to be further confirmed. In addition, there is still no database of BS auscultation data for normal people and patients with IO.

CLINICAL DEMAND FOR CA OF BS ANALYSIS IN THE DIAGNOSIS AND TREATMENT OF IO

Difficulty in Etiological Diagnosis of IO

IO can be roughly divided into three categories based on etiology, including MIO, dynamic IO, and mesenteric vascular obstruction (MVO) (8). Patients with different causes of IO are supposed to have different treatment and prognosis. Therefore, it is crucial to precisely discriminate etiologies. Usually, abdominal imaging examinations, endoscopy, and traditional BS auscultation may reveal causes of most IO cases. However, it is difficult to identify MVO and MIO in some insidious condition by using the above traditional diagnosis methods, which could easily lead to missed diagnosis or misdiagnosis. CA of BS analysis, especially ML and DL with advantages in non-invasiveness and big data processing ability, has the potential to diagnose IO of unknown etiology. Zaborski et al. demonstrated that the number of impulses of BS contributed to identify MIO caused by some tumors and diffuse peritonitis (9). Nevertheless, it is still unable to distinguish between benign and malignant tumors. Therefore, CA of BS analysis may provide a new way for etiologic diagnosis of IO.

Difficulty in Identification of IO Location

According to the location, IO can be classified into high IO (duodenum and jejunum), low IO (small intestine), and colorectal obstruction (10). The site of IO determines the choice of internal treatment and surgical operation. However, in some cases, there are discrepancies between imaging findings and clinical conditions. So, endoscopy is needed for further diagnosis. However, application of endoscopy is not suitable for the patients with severe cardiopulmonary disease, unstable vital sign, acute cerebral accident with complications of cardiac infarction, respiratory depression, hypotension, infection, perforation, etc. Ching et al. found that multi-channel acquisition of BS could be used to identify the possible location of IO with significant difference in sound characteristics (sound duration and peak frequency) between large bowel and small bowel obstruction (5). Unfortunately, the location of IO by BS analysis is relatively rough at present with no information about specific intestinal segment provided. Therefore, there is still a long way to go for BS analysis to guide clinical work.

Difficulty in Identification of IO Severity

In severe cases of IO, perforation, intestinal fistula, intestinal ischemia and necrosis, peritonitis, and even death may occur. They probably need emergency surgery to alleviate the condition, while mild incomplete IO can be relieved by conservative treatments. Early identification of the severity of IO may help to develop appropriate treatment and improve the prognosis of patients. In most cases, the patients can get timely surgical treatment since most complications may be recognized by imaging examination, symptoms, and signs of the patients. However, some other patients had abdominal pain relieved after medical treatment with no imaging findings of perforation, intestinal fistula, etc. Surprisingly, ischemic necrosis of the intestinal segment, and even intestinal fistula were found during surgery of those patients. Yoshino et al. used a signal processor to analyze the BS among 21 patients with MIO to evaluate the severity of IO based on the frequency and peak values of BS (11). However, there has been a lack of severity scoring systems for IO in clinical practice up to now. In relevant studies, the results could not be compared with clinical grading standards, which resulted in insufficient strength of evidence. More large sample prospective clinical studies of BS analysis may be expected to solve this problem.

Difficulty in Monitoring of Intestinal Motility for IO Patients

Clinicians usually judge the recovery of intestinal motility in patients with MVO, dynamic IO by traditional BS auscultation, so as to guide the timing of enteral nutrition initiation. However, in clinical work, we observed that vomiting and abdominal distension occurred again after eating among some patients whose BS returned to normal. The above symptoms were relieved again after fasting with slow peristalsis and poor motility in intestinal radiography. Therefore, traditional BS auscultation cannot accurately determine the recovery of intestinal motility. In addition, repeated imaging examinations cause increased radiation exposure. Non-invasive CA of BS analysis is expected to break through this bottleneck. Spiegel used an acoustic gastrointestinal surveillance (AGIS) biosensor to identify and predict the patients at high risk of postoperative IO and to help to determine the timing of enteral nutrition initiation after surgery (12). In the near future, BS analysis is expected to judge the

recovery of intestinal motility more accurately and to determine optimal timing for enteral nutrition.

DISCUSSION

At present, application research on CA for BS analysis in IO is relatively rare and not deep enough, and there are many aspects worthy of further improvement. ML and DL have been successfully applied in the field of acoustic-based disease diagnosis (13, 14). Since BS analysis is also one kind of acoustic-based disease diagnosis technology, we expect that the introduction of ML and DL techniques into the BS field will contribute to the research in the field.

First of all, the characteristics of BS in normal people and IO patients are still unclear. As a result, there is still no standard definition or classification of BS. We need to establish BS information database for ML and DL, analyze the characteristics of BS and unify its clinical classification and definition with the same duration, location, sensor of BS acquisition and acoustic characteristics used for classification. Secondly, in a certain situation, it is still difficult to identify the cause of IO by using only conventional diagnostic methods. We need to analyze and summarize the characteristics of BS in different causes of IO, and confirm that through clinical trials, so as to achieve etiological diagnosis of IO by using ML and DL for BS analysis. Thirdly, location accuracy for IO still needs improvement by using ML and DL to extract and analyze the characteristics of BS during IO in different intestinal segments with more specific information about IO location and sensors that can realize simultaneous auscultation of different parts of the intestine. In addition, there has been a lack of severity scoring systems for IO. Consequently, that possibly leads to the delay of treatment due to inaccurate judgment for the progress and prognosis of some patients with IO. We need to develop large sample prospective clinical trials of BS analysis by using ML and DL to promote establishment of severity scoring systems for IO. That would help clinicians to judge the severity of IO in time and effectively, and to improve the prognosis of the patients. Last but not least, we still have difficulty in judging the recovery of intestinal function in some IO patients by traditional BS auscultation, imaging examination, symptoms, and signs. The characteristics of BS should be analyzed by ML and DL in patients with different course of IO to judge the recovery of intestinal mobility more accurately and to determine optimal timing for enteral nutrition.

In conclusion, we are looking forward to making better use of ML and DL in the diagnosis and treatment of IO, so as to optimize decision-making for treatment strategy, provide precise treatment of IO and realize real-time diagnosis and monitoring of IO as soon as possible.

AUTHOR CONTRIBUTIONS

ZY wrote the review. JL and CT designed and revised the manuscript. ZY, LH, JJ, and BH searched and collected the literature. All authors contributed to the article and approved the submitted version.

FUNDING

This paper was financially supported in part by the following funds: Grant No. #2018GZ0088, Key research and development program of science and technology Department of Sichuan Province, China.

REFERENCES

1. Gu Y, Lim HJ, Moser MA. How useful are bowel sounds in assessing the abdomen? *Dig Surg.* (2010) 27:422–6. doi: 10.1159/000319372
2. Qian K, Li X, Li H, Li S, Li W, Ning Z, et al. Computer audition for healthcare: opportunities and challenges. *Front Digital Health.* (2020) 2:5. doi: 10.3389/fdgth.2020.00005
3. Dalle D, Devroede G, Thibault R, Perrault J. Computer analysis of bowel sounds. *Comput Biol Med.* (1975) 4:247–56. doi: 10.1016/0010-4825(75)90036-0
4. Du X, Allwood G, Webberley KM, Osseiran A, Marshall BJ. Bowel sounds identification and migrating motor complex detection with low-cost piezoelectric acoustic sensing device. *Sensors.* (2018) 18:4240. doi: 10.3390/s18124240
5. Ching SS, Tan YK. Spectral analysis of bowel sounds in intestinal obstruction using an electronic stethoscope. *World J Gastroenterol.* (2012) 18:4585–92. doi: 10.3748/wjg.v18.i33.4585
6. Han J, Qian K, Song M, Yang Z, Ren Z, Liu S, et al. An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety. *Paper Presented Interspeech.* (2020) 2020:4946–50. doi: 10.21437/Interspeech.2020-2223
7. Wang F, Wu D, Jin P, Zhang Y, Yang Y, Ma Y, et al. A flexible skin-mounted wireless acoustic device for bowel sounds monitoring and evaluation. *Sci China Information Sci.* (2019) 62:202402. doi: 10.1007/s11432-019-9906-1
8. Madl C, Druml W. Gastrointestinal disorders of the critically ill. Systemic consequences of ileus. *Best Pract Res Clin Gastroenterol.* (2003) 17:445–56. doi: 10.1016/S1521-6918(03)00022-2
9. Zaborski D, Halczak M, Grzesiak W, Modrzejewski A. Recording and analysis of bowel sounds. *Euroasian J Hepatogastroenterol.* (2015) 5:67–73. doi: 10.5005/jp-journals-10018-1137
10. Gao F. Diagnosis and treatment of lower intestinal obstruction. *Chinese J Pract Surg.* (2000) 20:461–2. doi: 10.3321/j.issn:1005-2208.2000.08.008
11. Yoshino H, Abe Y, Yoshino T, Ohsato K. Clinical application of spectral analysis of bowel sounds in intestinal obstruction. *Dis Colon Rectum.* (1990) 33:753–7. doi: 10.1007/BF02052320
12. Spiegel BM, Kaneshiro M, Russell MM, Lin A, Patel A, Tashjian VC, et al. Validation of an acoustic gastrointestinal surveillance biosensor for postoperative ileus. *J Gastrointest Surg.* (2014) 18:1795–803. doi: 10.1007/s11605-014-2597-y
13. Kun Qian, Christoph Janott, Maximilian Schmitt, Zixing Zhang, Clemens Heiser, Werner Hemmert, et al. Can machine learning assist locating the excitation of snore sound? A review. *IEEE J Biomed Health Inform.* (2020) 25:1233–46. doi: 10.1109/JBHI.2020.3012666
14. Kun Qian, Maximilian Schmitt, Huaiyuan Zheng, Tomoya Koike, Jing Han, Juan Liu, et al. Computer audition for fighting the SARS-

CoV-2 Corona Crisis — Introducing the multi-task speech corpus for COVID-19". *IEEE Internet Things J.* (2021). doi: 10.1109/JIOT.2021.3067605. [Epub ahead of print].

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 Medicine

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fmed.2021.654696

A Novel Hierarchical Deep Learning Framework for Diagnosing Multiple Visual Impairment Diseases in the Clinical Environment

Jiaxu Hong[1,2,3,4†], Xiaoqing Liu[5*†], Youwen Guo[6‡], Hao Gu[2], Lei Gu[7,8], Jianjiang Xu[1], Yi Lu[1], Xinghuai Sun[1], Zhengqiang Ye[1], Jian Liu[2], Brock A. Peters[9] and Jason Chen[9‡]*

[1] Department of Ophthalmology and Visual Science, Eye, and Ear, Nose, and Thorat Hospital, Shanghai Medical, College Fudan University, Shanghai, China, [2] Department of Ophthalmology, Affiliated Hospital of Guizhou Medical University, Guiyang, China, [3] Key Laboratory of Myopia, Ministry of Health (Fudan University), Shanghai, China, [4] Shanghai Engineering Research Center of Synthetic Immunology, Fudan University, Shanghai, China, [5] AI Laboratory, Deepwise Healthcare, Beijing, China, [6] Wuhan Servicebio Technology, Wuhan, China, [7] Epigenetics Laboratory, Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany, [8] Cardiopulmonary Institute (CPI), Bad Nauheim, Germany, [9] Complete Genomics Inc., San Jose, CA, United States

Early detection and treatment of visual impairment diseases are critical and integral to combating avoidable blindness. To enable this, artificial intelligence–based disease identification approaches are vital for visual impairment diseases, especially for people living in areas with a few ophthalmologists. In this study, we demonstrated the identification of a large variety of visual impairment diseases using a coarse-to-fine approach. We designed a hierarchical deep learning network, which is composed of a family of multi-task & multi-label learning classifiers representing different levels of eye diseases derived from a predefined hierarchical eye disease taxonomy. A multi-level disease–guided loss function was proposed to learn the fine-grained variability of eye disease features. The proposed framework was trained for both ocular surface and retinal images, independently. The training dataset comprised 7,100 clinical images from 1,600 patients with 100 diseases. To show the feasibility of the proposed framework, we demonstrated eye disease identification on the first two levels of the eye disease taxonomy, namely 7 ocular diseases with 4 ocular surface diseases and 3 retinal fundus diseases in level 1 and 17 subclasses with 9 ocular surface diseases and 8 retinal fundus diseases in level 2. The proposed framework is flexible and extensible, which can be inherently trained on more levels with sufficient training data for each subtype diseases (e.g., the 17 classes of level 2 include 100 subtype diseases defined as level 3 diseases). The performance of the proposed framework was evaluated against 40 board-certified ophthalmologists on clinical cases with various visual impairment diseases and showed that the proposed framework had high sensitivity and specificity with the area under the receiver operating characteristic curve ranging from 0.743 to 0.989 in identifying all identified major causes of blindness. Further assessment of 4,670 cases in a tertiary eye center also demonstrated that the proposed framework achieved a high identification accuracy rate for different visual impairment diseases compared with that of human

OPEN ACCESS

Edited by:
Liang Zhang,
Xidian University, China

Reviewed by:
Madhura Ingalhalikar,
Symbiosis International University, India
Guangming Zhu,
Xidian University, China

***Correspondence:**
Xiaoqing Liu
xiaoqing.liu@ieee.org
Jiaxu Hong
jiaxu_hong@163.com

[†]These authors have contributed equally to this work and share first authorship

[‡]ORCID:
Youwen Guo
orcid.org/0000-0001-9932-1047
Jason Chen
orcid.org/0000-0002-7627-3019

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *17 January 2021*
Accepted: *07 May 2021*
Published: *07 June 2021*

Citation:
Hong J, Liu X, Guo Y, Gu H, Gu L, Xu J, Lu Y, Sun X, Ye Z, Liu J, Peters BA and Chen J (2021) A Novel Hierarchical Deep Learning Framework for Diagnosing Multiple Visual Impairment Diseases in the Clinical Environment. Front. Med. 8:654696. doi: 10.3389/fmed.2021.654696

graders in a clinical setting. The proposed hierarchical deep learning framework would improve clinical practice in ophthalmology and broaden the scope of service available, especially for people living in areas with a few ophthalmologists.

Keywords: artificial intelligence, hierarchical deep learning framework, visual impairment disease, coarse-to-fine, multi-task multi-label

INTRODUCTION

Eye diseases leading to visual impairment are a significant source of social burden. It is estimated that, as of 2017, 1 billion people were living with vision impairment worldwide, including those with moderate or severe distance vision impairment or blindness caused by unaddressed refractive error (123.7 million), cataract (65.2 million), glaucoma (6.9 million), corneal opacities (4.2 million), diabetic retinopathy (3.0 million), and trachoma (2.0 million), as well as near vision impairment caused by unaddressed presbyopia (826.0 million) (1). In China, the most frequent cause of visual impairment is cataract, which is followed by corneal disease and glaucoma (2, 3). In contrast, age-related macular degeneration and diabetic retinopathy are more prevalent in the United States (4). Early detection and treatment of visual impairment diseases are critical and integral to combating this avoidable blindness worldwide.

A slit-lamp investigation of the ocular surface and retina using manual interpretation is a widely accepted screening tool to detect visual impairment diseases. However, this is highly dependent on the ophthalmologist's clinical experience, which is time-consuming and may have an interobserver variation on the same patient. Automated identification of various visual impairment diseases via slit-lamp photography has benefits such as increased efficiency, reproducibility, and access to eye care. To enable this, artificial intelligence (AI)-based approaches for the identification of visual impairment diseases are greatly needed, especially for people living in areas with a limited number of ophthalmologists.

Recent advances in AI, particularly convolutional neural networks (CNN)-based deep learning algorithms, have made it possible to learn the most predictive disease features directly from medical images given a large dataset of labeled examples (5, 6). Esteva et al. (7) proposed a dermatologist-level classification of skin cancer by fine-tuning a pretrained Inception-v3 network (8). Menegola et al. (9) also conducted experiments comparing training from scratch with fine-tuning of pretrained networks on skin lesion images. Their study showed that fine-tuning of pretrained networks worked better than training from scratch. Setio et al. (10) applied a multi-view CNN to classify points of interest in chest computed tomography as nodules or non-nodules. Similarly, Nie et al. (11) used a three-dimensional CNN on magnetic resonance images to assess the survival of patients suffering from brain tumors.

Because of the fine-grained variability in the appearance of eye lesions, most of the existing eye disease identification methods focused on a single disease type (such as retinopathy and macular diseases) via retinal fundus or optical coherence tomography (OCT) images. Gulshan et al. (12) demonstrated the detection of diabetic retinopathy by fine-tuning a pretrained Inception-v3 network on retinal fundus images. Similarly, Gargeya and Leng (13) performed automated identification of diabetic retinopathy using a ResNet-based architecture. Li et al. (14) adopted an Inception-v3 network to detect glaucomatous optic neuropathy using color fundus images, whereas Burlina et al. (15) applied both a pretrained model and a newly trained from a scratch model for automated grading of age-related macular degeneration from color fundus images. Schlegl et al. (16) and Treder et al. (17) proposed automated detection of macular diseases using OCT images. Long et al. (18) developed a technique for the diagnosis of congenital cataracts. However, their method was focused on images covering the pupil area only; therefore, their algorithm could not detect diseases affecting the peripheral cornea and limbus. To date, there have been few studies diagnosing ocular surface diseases or identifying various disease types simultaneously. Ting et al. (19) proposed a deep learning system for diabetic retinopathy and related eye diseases using retinal images. Fauw et al. (20) proposed an Ensemble-based deep learning framework that could make referral suggestions on retinal diseases by analyzing OCT images. Li et al. (21) presented a workflow for the segmentation of anatomical structures and annotation of pathological features in slit-lamp images, which improved the performance of a deep learning algorithm for diagnosing ophthalmic disorders. As most of these algorithms have been derived from datasets of one or a few ocular diseases, they struggle to detect visual impairment diseases accurately in large-scale, heterogeneous datasets.

To maximize the clinical utility of AI, we developed a hierarchical deep learning framework, which enables early screening and differentiation of a large variety of visual impairment diseases simultaneously in a coarse-to-fine manner. Here, a hierarchical architecture means that multiple classification layers are arranged in a hierarchical way for different levels. To test the feasibility of the proposed framework, we identified eye diseases on two different levels of the eye disease taxonomy. Thereby, in our case, the proposed framework would first perform disease classification for a lower level (i.e., level 1) and then perform a higher-level disease classification (i.e., level 2). Also, algorithm performance was tested against 40 ophthalmologists in a clinic-based dataset. Finally, we performed an observational diagnostic assessment comparison of visual impairment disease screening between the algorithm and the ophthalmologists in a tertiary eye center.

MATERIALS AND METHODS

Datasets

Our dataset came from two major eye centers in China: (i) the Eye and ENT Hospital of Fudan University, Shanghai, and (ii)

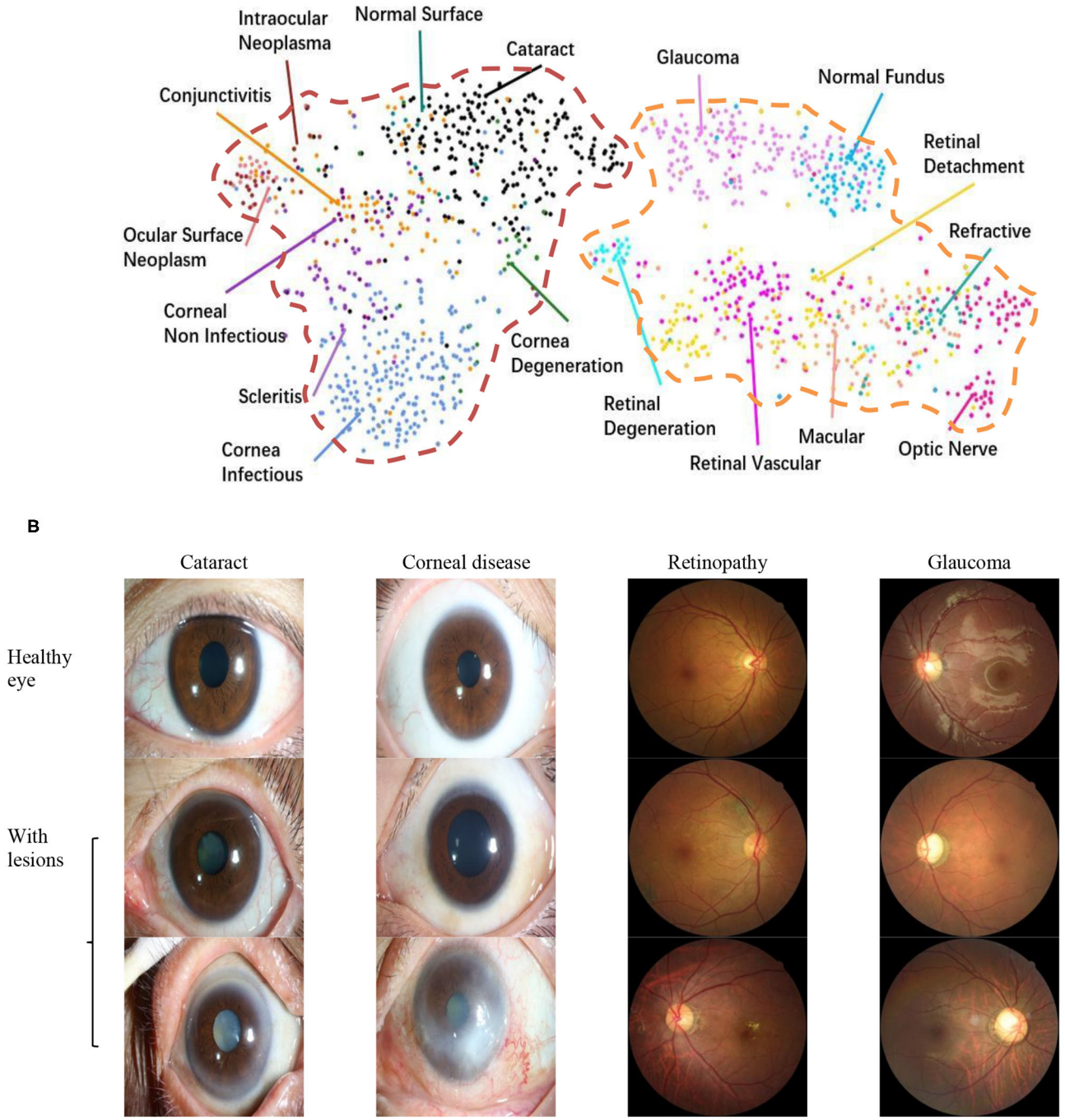

FIGURE 1 | Dataset. **(A)** t-Distributed stochastic neighbor embedding visualization of the collected dataset consisting of 17 major ocular disease classes (100 subtypes), leading to visual impairment, clustered according to deep features generated from the last layer of trained networks. Colored point clouds represent images with different visual impairment diseases. This visualization represents the ability of our method to objectively separate normal patients from early cases of visual impairment diseases for referral. **(B)** Example ocular surface and retinal images for the eye with some common diseases or healthy eye. In this study, the first two levels of the taxonomy consisting of 17 major ocular disease classes (100 subtypes) were used in performance evaluation.

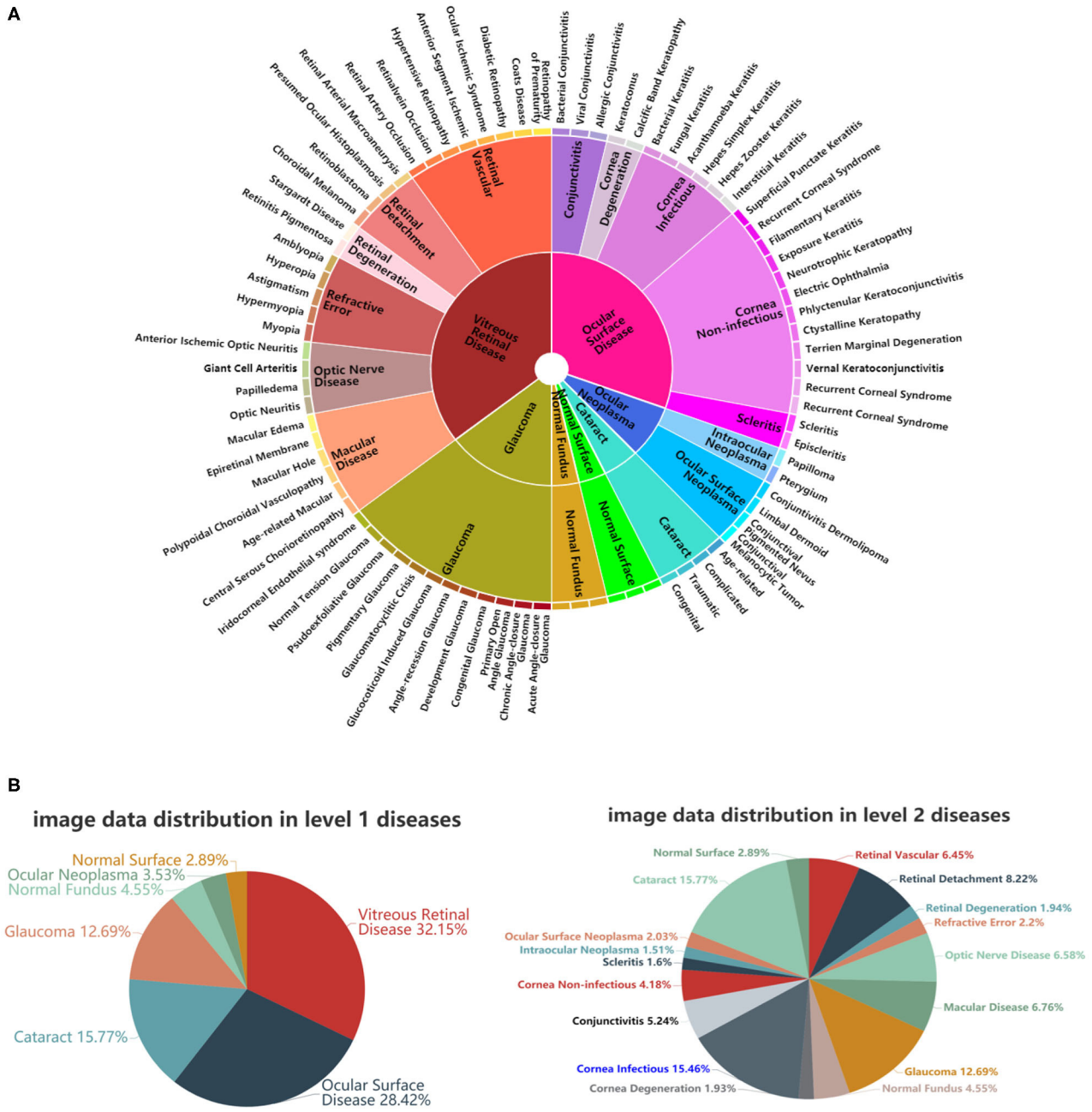

FIGURE 2 | A schematic illustration of the predefined eye disease taxonomy and example test set images. **(A)** Pie-structured eye disease taxonomy. **(B)** Data distribution for the first two levels of diseases.

the Affiliated Hospital of Guizhou Medical University, Guizhou. We used the IM 900 or 600 digital slit-lamp photography system (Haag-Streit, Switzerland) and CR-2 digital non-mydriatic retinal cameras (Canon, Japan). All images were annotated by senior ophthalmologists, where 50% of the proportion included retinal photographs and no images with the dilated pupil were included. Our objective was to provide a fast and cost-effective tool for screening patients with visual impairments. A suspected participant would be referred to a doctor for further assessment, including the dilated examination.

Retrospective Dataset

Thirty-two ophthalmologists were invited to grade the images of the retrospective database. During the training process of ophthalmologists, a dataset of 100 images (including 25 corneal disease cases, 25 cataract cases, 25 glaucoma cases, and 25 retinal

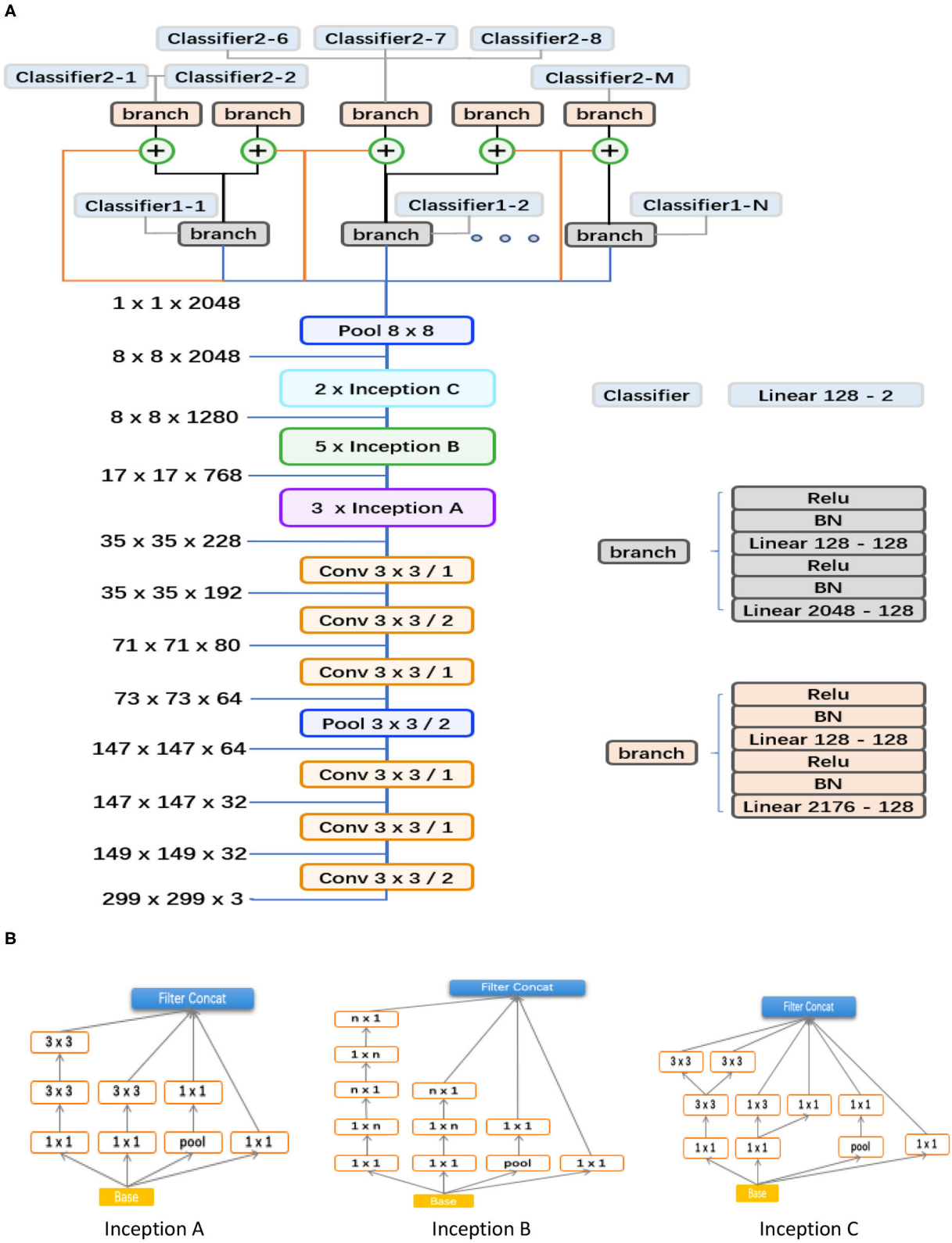

FIGURE 3 | Abstraction of the proposed hierarchical deep learning framework. **(A)** The proposed network architecture based on the feature network of Inception v3 (Conv 3 × 3/2 indicates that a 3 × 3 convolution kernel was used and stride = 2). The corresponding sizes of the input and output for each module are also shown. In

(Continued)

FIGURE 3 | our framework, a family of multi-task & multi-label classification layers were used hierarchically to represent various levels of eye diseases. The individual multi-task classifier layer is defined on the basis of a predefined eye disease taxonomy. Here, the data flow in blue indicates that the backbone is directly connected to the branch of level 1; the orange means that the backbone is directly connected to the branch of level 2; the flow in black means connecting from the branch of level 1 to the branch of level 2; and the $\bigoplus$ is a feature concatenation operation, where features from the black and orange are superimposed; finally, this 8*8 pooling layer is a global average pooling, which turns the 8*8 feature map into a 1*1 feature map. **(B)** Different spatial factorized Inception modules are presented here. Inception A contains the factorization of the original 5 × 5 convolutions, factorizes general n × n convolutions ($n = 5$ in our study), and has expanded the filter bank outputs.

disease cases) was used for the test. The participants' results were compared with those of two senior corneal specialists (H.G. and J.H.). The participants would not complete the training until they achieved a κ-value of 0.75 or more. A κ-value of 0 indicates that observed agreement is the same as that expected by chance; 1 indicates perfect agreement; 0.75 or more indicates substantial agreement and/or almost perfect agreement. As a result, 20 ophthalmologists were qualified as graders to classify images. Each photograph was reviewed with the same standard and annotated via face-to-face communication between two ophthalmologists. As all 7,100 images from 1,600 patients collected already had original diagnoses recorded in medical charts, graders were asked to review, validate, and classify the images.

Prospective Dataset

A total of 4,670 outpatients agreed to receive the test and got their ocular surface slit-lamp photographs taken before their physician visits. Informed consent was obtained from all the participants. A software practitioner participating in this study fed these images as input to the trained deep learning software model. The algorithm generates a probability/confidence score over the classification nodes in a sequential manner, i.e., level by level. If the probability/confidence score of any disease subtype was greater than a predefined threshold, the disease subtype was diagnosed as positive. To quantitatively compare the sensitivity and specificity of our algorithm to that of the other 40 ophthalmologists on the diagnostic task of these cases, receiver operating characteristic (ROC) curves were plotted where each ophthalmologist was asked about the diagnosis on the basis of the images. Thirteen additional cases were also independently collected from clinics for our direct performance test sets.

To explore the visual characteristics of different clinical classes, we examined the internal image features learned by the proposed framework using t-distributed stochastic neighbor embedding (22). As demonstrated in **Figure 1A**, each point represents an eye image projected from the n-dimensional output of the last hidden layer of Inception-v3 backbone into two dimensions. We see clusters of points of the same clinical classes. This visualization represents the ability of our method to objectively separate normal patients from early cases of visual impairment diseases for a referral. **Figure 1B** shows a few examples of images that demonstrate the visual features using which the proposed hierarchical deep learning framework can identify and make a diagnosis.

Taxonomy

Inspired by Esteva et al. (7), who defined skin diseases in a tree structure, we adopted a similar approach to define our domain taxonomy structure for eye diseases, taking advantage of fine-grained information embedded within the images. Our taxonomy represented 100 individual diseases hierarchically arranged in a Pie structure. It was derived based on the collected retrospective database with 7,100 images from 1,600 patients by ophthalmologists using a bottom-up procedure: Individual diseases—initialized were defined as leaf nodes, and then were merged on the basis of clinical and visual similarity until the entire structure was connected.

As shown in **Figure 2A**, the taxonomy is useful in generating hierarchical training classes that are both well-suited for machine learning classifiers and medically relevant. In this study, the first two levels of the taxonomy were used in performance validation. **Figure 2B** illustrates the corresponding data distributions. It is worth mentioning that due to insufficient numbers of images for each of the level 3 diseases, we did not perform the level 3 classification. However, the extension to more levels can be implemented via our flexible and extensive framework with sufficient training data.

Proposed Hierarchical Deep Learning Framework

As shown in **Figure 3**, the proposed hierarchical deep learning framework is composed of a family of multi-task & multi-label learning classifiers representing different levels of eye disease classification derived from the hierarchical eye disease taxonomy. Here, we used an Inception-v3 CNN as the backbone of the proposed framework, and the final classification layer of the Inception-v3 network was replaced with our novel hierarchical multi-task & multi-label classification layers. Each task branch consists of several stacked fully connected units, hierarchically representing various levels of eye disease classification. As a result, the classification results of lower levels of classifiers can be used as priors for higher levels of classifiers, thereby improving the final classification performance.

We trained the model by minimizing our novel multi-level eye disease–guided loss function consisting of multiple levels of losses. The objective function for two levels can be represented as follows:

$$Loss_T = \alpha^* loss_{l1} + (1 - \alpha)^* loss_{l2} \quad (1)$$

where the term $Loss_T$ is the total loss of the final model, and $loss_{l1}$ and $loss_{l2}$ represent the corresponding losses for levels 1 and 2 of eye disease identification, respectively. α is a weight parameter that is used to control the balance between the two losses. For the two levels, $\alpha \in (0, 0.5)$, setting more weight for the higher level because the ultimate goal was to classify

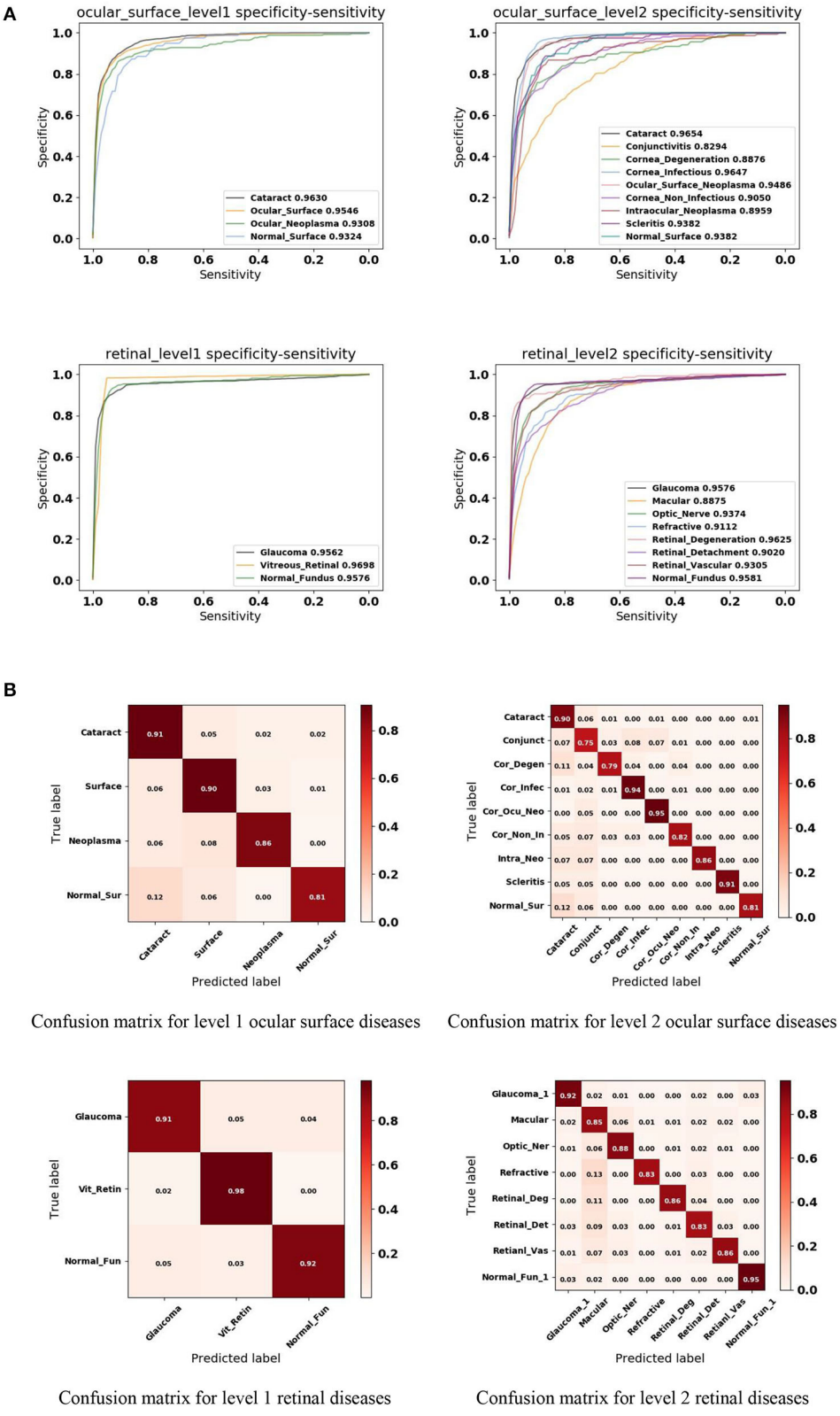

FIGURE 4 | Performance of the proposed hierarchical deep learning framework. **(A)** The mean receiver operating characteristic (ROC) curve for various eye diseases of the first two levels of the eye disease taxonomy. AUC is the area under the ROC curve. **(B)** Confusion matrices for the first two levels of the eye disease taxonomy. Conjunct, Conjunctivitis; Cor_Degen, Corneal_Degeneration; Cor_Infec, Corneal_Infectious; Ocu_Cor_Neo, Ocular_Corneal_Neoplasma; Cor_Non_In, Corneal_Non_Infectious; Intra_Neo, Intraocular_Neoplasma; Normal_Sur, Normal_Surface; Optic_Ner, Optic_Ner; Retinal_Deg, Retinal_Degeneration; Retinal_Det, Retinal_Detachment; Retinal_Vas, Retinal_Vascular; Normal_Fun, Normal_Fundus.

higher levels of diseases. Through experiments, we found that $\alpha = 0.3$ performed well (i.e., the loss weight ratio 3:7 between level 1 and 2 classifiers). In this study, we used the sigmoid function for each class instead of the commonly used SoftMax function, for multiple diseases may simultaneously exist. Because of the unbalanced property of data, we applied the focal loss (23) for the loss function of each level, which reduced the impact of data imbalance and made the training focus on hard negatives as well. The focal loss function can be represented as follows:

$$FL\left(p_t\right) = -\left(1-p_t\right)^{\gamma} \ log\left(p_t\right) \tag{2}$$

where

$$p_t = \begin{cases} p & if\ y=1 \\ 1-p & otherwise \end{cases} \tag{3}$$

$\left(1-p_t\right)^{\gamma}$ is a modulating factor of the cross-entropy loss, with a tunable focusing parameter $\gamma \geq 0$, $p \in [0, 1]$. During the training process, various data augmentation methods (including horizontal and vertical flipping, color jitter, rotation, etc.) were also applied to all classes independently on-the-fly. It is worth mentioning that the online data augmentation was aimed at increasing the diversity of data for generalization rather than balancing and/or increasing the amount of training data.

Instead of training from scratch, we applied a fine-tuning strategy on a pretrained model using a multi-step retraining strategy. In this study, all images were resized to the size of 299 × 299 since that is the default input size for the Inception-v3 model. We used the Inception-v3 model pretrained on the ImageNet dataset (24) as the initial model and fine-tuned all layers with our dataset. First, the multi-task branches were trained by freezing the backbone's weights for 5 epochs. The Adam optimizer and a learning rate of 0.0001 and epsilon of 0.1 were used. Then, we performed a multi-step retraining strategy. In this strategy, we gradually unfroze the layer weights in steps, with the first few layers being unfrozen last. The learning rates were progressively reduced from 0.0001 to 0.000001, whereas other parameters were kept unchanged. Every step lasted 20 epochs. We used Facebook's PyTorch deep learning framework (25) to train, validate, and test the algorithm networks.

RESULTS

Performance Evaluation

Algorithm performance was measured by the area under the ROC curve (AUC) and the accuracy rate. The accuracy rate calculated the percentage of correctly predicted individuals among the whole test set, whereas the ROC curve was generated by plotting the curve of sensitivity against specificity, which can be defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative rates, respectively. TP and TN represent correctly predicted positives and negatives with respect to the ground truth labels. FP and FN represent incorrectly predicted positives and negatives with respect to the ground truth labels.

In this study, we applied a 5-fold cross-validation strategy to evaluate the effectiveness of the proposed framework. This strategy randomly divides the entire dataset into five subsets, each containing around 20% of the data. Model training and validation were performed five times. **Figure 4A** shows that our framework achieved high sensitivity, specificity, and AUC for most of the identified diseases. **Figure 4B** illustrates the corresponding confusion matrices for disease classification. As shown in level 1 confusion matrices, the CNN model performed extremely well on all three retinal fundus diseases, with an accuracy of 0.91 for glaucoma, 0.98 for vitreoretinal disease, and 0.92 for normal fundus. Meanwhile, the CNN model performed moderately well on all four ocular surface diseases, with an accuracy of 0.91 for cataract, 0.90 for surface disease, 0.90 for neoplasma, and 0.81 for normal surface images. This may be because fundus images contain more discriminative features than do ocular surface images. The model confused normal surface cases with cataract (12.0%) and confused cataract with

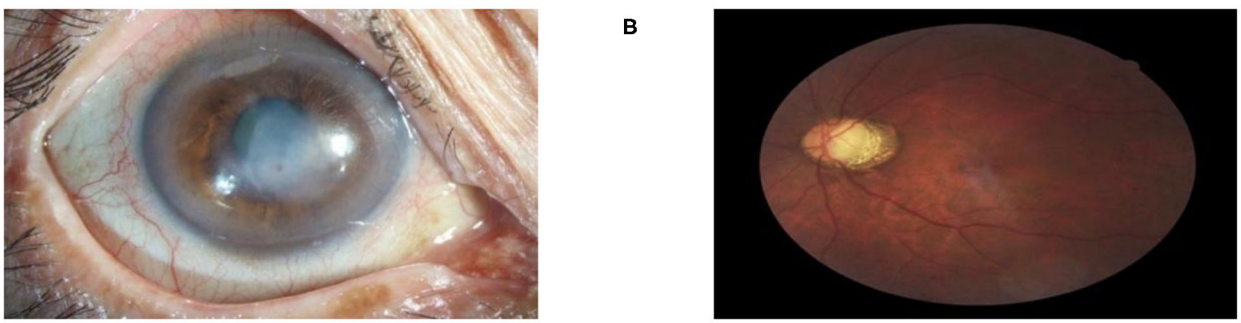

FIGURE 5 | Multi-label diagnostic results. The proposed hierarchical deep learning framework is capable of detecting multiple diseases simultaneously on the same patient: **(A)** cataract with 76.74% and corneal disease with 75.94% confidence and **(B)** glaucoma with 79.04% and retinopathy with 51.01% confidence.

surface disease (5.0%), neoplasm (2.0%), and normal surface images (2.0%). From these results, we can conclude that it is easy to confuse the normal surface with cataract because of appearance similarities, whereas cataract has more appearance diversity, which can also be confused with other ocular surface diseases and neoplasms. Similar results can be found in level 2 confusion matrices.

Because of the multi-task & multi-label property of the proposed framework, the trained model is capable of detecting multiple diseases simultaneously on the same patient, reflecting

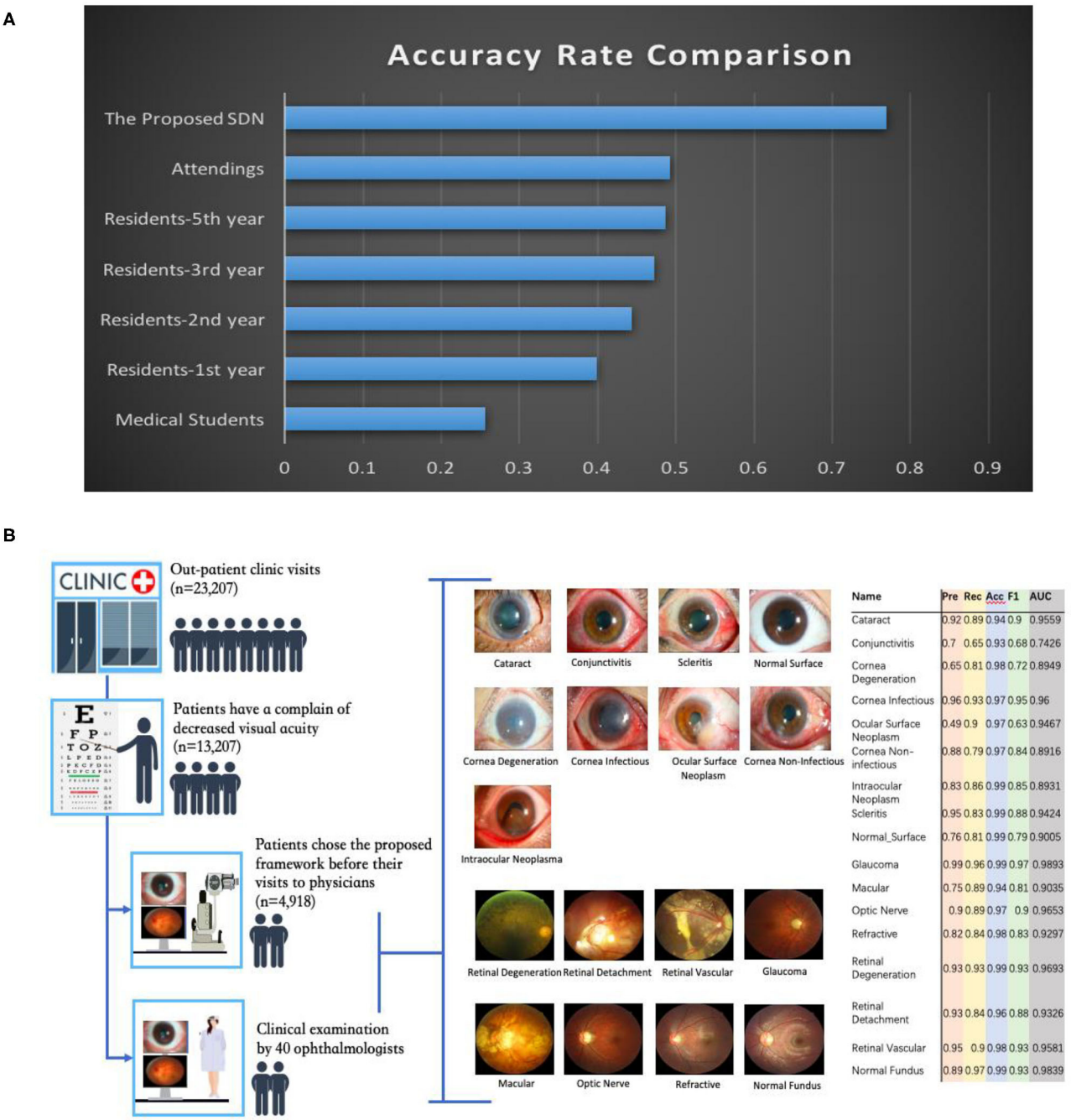

FIGURE 6 | Eye disease classification performance of the proposed hierarchical deep learning framework and ophthalmologists. **(A)** The proposed hierarchical deep learning framework was tested against 40 board-certified ophthalmologists in diagnosing the clinical cases of 13 patients in a real-world setting. For each image, the ophthalmologists were asked to make three diagnoses. The proposed hierarchical deep learning framework outperformed all levels of board-certified ophthalmologists for all cases. **(B)** Clinical application of the proposed hierarchical deep learning framework for visual impairment diseases in a tertiary eye center. Discrepancies between manual grades and the proposed hierarchical deep learning framework results were sent to an independent panel of senior specialists for arbitration.

true clinical cases. As illustrated in **Figure 5**, both cataract and corneal disease were detected simultaneously within a single ocular surface image with 76.74 and 75.94% confidence, respectively. Similarly, both glaucoma and retinopathy were also detected within one retinal image with 79.04 and 51.01% confidence, respectively. It needs to be mentioned here that in this study, if the prediction score was > 50%, the system considered the screening output of the patient with the corresponding disease. In a real-world setting, if the screening output of the patient has one of the diseases listed above, the patient would be referred to a specialist for further diagnosis. Physicians need to consider not only the screening result but also the diagnostic severity of the disease to make clinical decisions for a patient. This was beyond the scope of our study. Our goal was to provide a fast and cost-effective screening tool for patients with visual impairment.

Comparison Tests

To both quantitatively and qualitatively demonstrate the effectiveness of the proposed framework, we also compared it with 40 board-certified ophthalmologists in diagnosing clinical cases. The comparison tests used 20 images from 13 patients.

TABLE 1 | Computational cost comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

Computational cost	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Training (hours)	12.5	11.2	10.0	11.4	11.0
Inference (seconds)	0.097	0.083	0.069	0.075	0.106

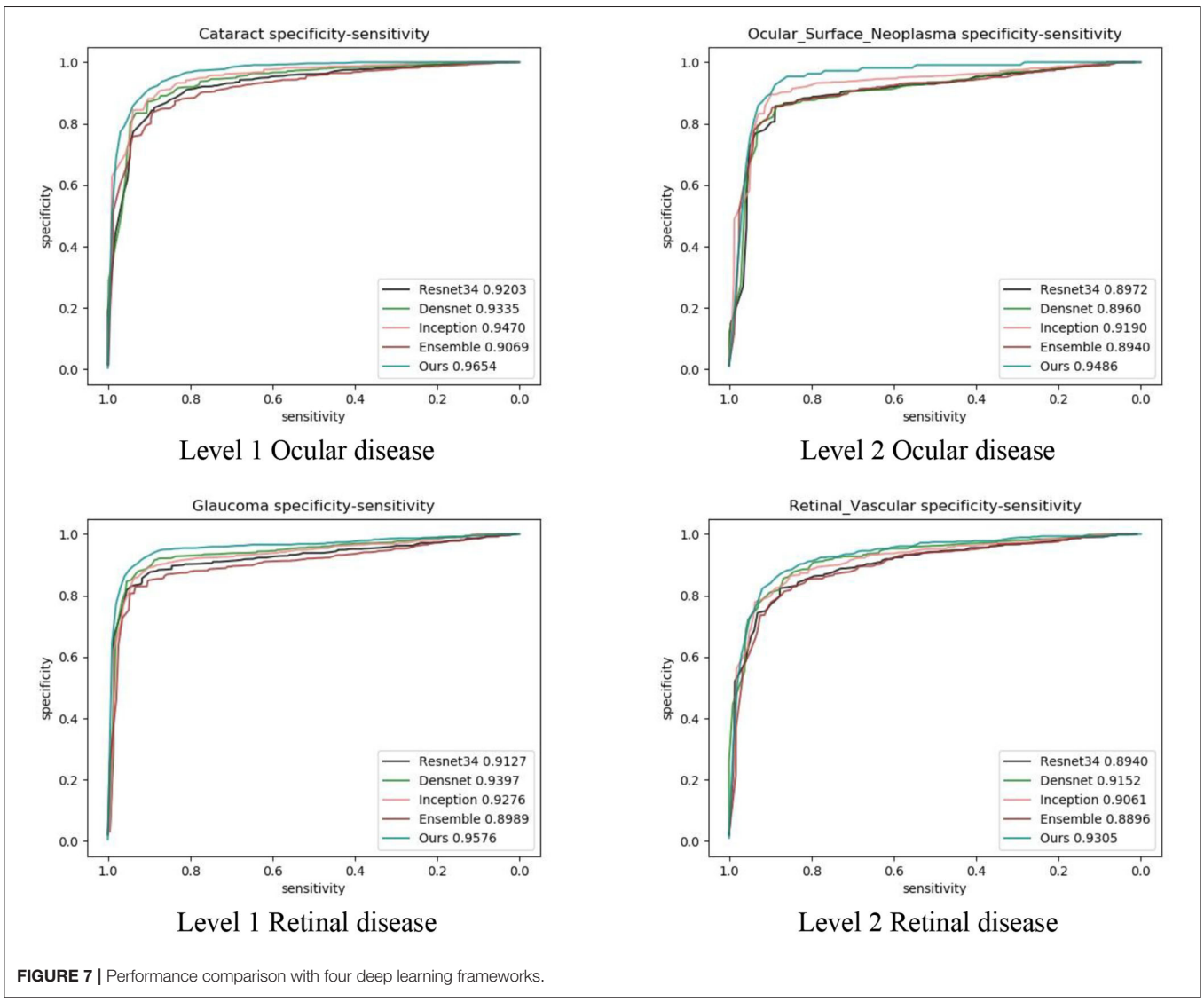

FIGURE 7 | Performance comparison with four deep learning frameworks.

TABLE 2 | AUC comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.96**	0.94	0.92	0.93	0.91
Ocular surface (*n* = 2,018)	**0.95**	0.94	0.91	0.93	0.90
Ocular neoplasm (*n* = 251)	**0.93**	0.89	0.89	0.91	0.88
Normal surface (*n* = 205)	**0.93**	**0.93**	**0.93**	0.92	0.90
Weighted average	**0.95**	0.93	0.91	0.93	0.90
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.97**	0.95	0.92	0.93	0.91
Conjunctivitis (*n* = 372)	**0.83**	0.82	0.81	**0.83**	0.81
Cornea degeneration (*n* = 137)	**0.89**	0.86	0.85	**0.89**	0.83
Cornea infectious (*n* = 1,098)	**0.96**	0.95	0.93	0.94	0.91
Intraocular neoplasma (*n* = 107)	**0.95**	0.92	0.90	0.90	0.89
Cornea non-infectious (*n* = 297)	0.91	0.89	**0.93**	0.88	0.86
Ocular surface neoplasm (*n* = 144)	**0.90**	0.88	0.86	0.87	0.85
Scleritis (*n* = 114)	**0.94**	0.93	0.93	0.93	0.90
Normal surface (*n* = 205)	**0.94**	0.93	**0.94**	0.93	0.91
Weighted average	**0.94**	0.92	0.91	0.91	0.89
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.94	0.91	0.92	0.90
Vitreoretinal disease (*n* = 2,283)	**0.97**	0.95	0.93	0.94	0.92
Normal fundus (*n* = 323)	**0.96**	**0.96**	0.94	0.94	0.92
Weighted average	**0.97**	0.95	0.93	0.94	0.91
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.94	0.91	0.93	0.90
Macular disease (*n* = 480)	**0.89**	0.88	0.85	0.86	0.85
Optic nerve disease (*n* = 467)	**0.94**	**0.94**	0.90	0.91	0.89
Refractive error (*n* = 156)	**0.91**	0.90	0.89	0.89	0.89
Retinal degeneration (*n* = 138)	0.96	**0.97**	0.93	0.96	0.92
Retinal detachment (*n* = 584)	**0.90**	0.89	0.87	0.88	0.85
Retinal vascular disease (*n* = 458)	**0.93**	0.92	0.89	0.91	0.89
Normal fundus (*n* = 323)	0.96	**0.97**	0.93	0.94	0.92
Weighted average	**0.93**	0.92	0.89	0.91	0.88

Bold value means "Best performance".

The tested diseases include allergic conjunctivitis, dry eye, bacterial conjunctivitis, Mooren's corneal ulcer, keratoconus, fungal keratitis, viral keratitis, scleritis, age-related macular degeneration, cataract, primary angle closure glaucoma, myopia, diabetic retinopathy, and retinal detachment. For this study, each ophthalmologist was asked for the three most likely diagnoses of the patient. This choice of question reflects the actual in-clinic task in which ophthalmologists would decide whether or not to request further examinations. For a fair comparison, the proposed hierarchical deep learning framework also outputs the top three diagnoses with probability/confidence scores. The outcome was considered "correct" when one of the three diagnoses made by the proposed hierarchical deep learning framework or an ophthalmologist included the real diagnosis for the case. Remarkably, the proposed hierarchical deep learning framework outperformed all levels of board-certified ophthalmologists in every case, as shown in **Figure 6A** ($P < 0.05$ in t-test).

In addition, we performed an observational diagnostic assessment comparison between the proposed framework and human graders in a tertiary eye center to determine whether or not the proposed framework can be introduced into visual impairment disease screening. As demonstrated in **Figure 6B**, 4,670 consecutive patients visiting the Shanghai Eye and ENT Hospital were invited to get their slit-lamp photographs taken before they were checked by their physicians. Discrepancies between manual grades and the proposed hierarchical deep learning framework results were sent to a panel of senior

TABLE 3 | Accuracy comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.93**	0.92	0.9	0.91	0.89
Ocular surface (*n* = 2,018)	**0.92**	0.9	0.88	0.89	0.87
Ocular neoplasm (*n* = 251)	0.96	**0.97**	0.96	0.96	0.95
Normal surface (*n* = 205)	0.98	0.98	**0.99**	0.98	0.98
weighted average	**0.93**	0.92	0.90	0.91	0.89
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.94**	0.93	0.91	0.92	0.9
Conjunctivitis (*n* = 372)	0.93	0.93	0.93	**0.94**	0.92
Cornea degeneration (*n* = 137)	0.97	0.97	0.97	**0.98**	0.97
Cornea infectious (*n* = 1,098)	**0.97**	0.96	0.94	0.95	0.93
Intraocular neoplasma (*n* = 107)	**0.99**	0.98	0.98	0.98	0.98
Cornea non-infectious (*n* = 297)	0.97	**0.98**	**0.98**	0.97	0.97
Ocular surface neoplasm (*n* = 144)	0.98	**0.99**	0.98	0.98	0.98
Scleritis (*n* = 114)	**0.99**	0.98	0.98	0.98	0.07
Normal surface (*n* = 205)	0.98	0.98	0.98	0.98	**0.99**
Weighted average	**0.96**	0.95	0.94	0.95	**0.90**
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.95	0.93	0.94	0.93
Vitreoretinal disease (*n* = 2,283)	**0.97**	0.96	0.93	0.95	0.92
Normal fundus (*n* = 323)	0.97	**0.98**	0.97	0.97	0.97
Weighted average	**0.97**	0.96	0.93	0.95	0.93
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.97**	0.96	0.94	0.94	0.93
Macular disease (*n* = 480)	**0.93**	0.92	0.91	0.91	0.9
Optic nerve disease (*n* = 467)	**0.96**	**0.96**	0.95	0.95	0.95
Refractive error (*n* = 156)	0.98	**0.99**	0.98	0.98	0.98
Retinal degeneration (*n* = 138)	**0.99**	**0.99**	0.98	0.98	0.98
Retinal detachment (*n* = 584)	**0.96**	0.95	0.94	0.94	0.93
Retinal vascular disease (*n* = 458)	**0.97**	0.96	0.96	0.96	0.96
Normal fundus (*n* = 323)	**0.99**	0.98	0.98	0.98	0.98
Weighted average	**0.97**	0.96	0.95	0.95	0.94

Bold value means "Best performance".

ophthalmologists for arbitration. Our data showed that the proposed hierarchical deep learning framework achieved an acceptable detection accuracy rate for visual impairment disease screening when compared with that of human graders in a clinical setting. The detection AUC of the proposed hierarchical deep learning framework for 17 subclasses in level 2 of visual impairment diseases ranged from 0.743 to 0.989.

We also compared our algorithm performance with four previously reported methods, namely Inception-v3 (8), ResNet (26), DenseNet (27), and Ensemble (28). The Ensemble model combined all backbone features extracted from the other three models and applied a tree-based classifier for the final classification. To have a fair comparison, all the networks above were also trained as multi-task & multi-label networks but without the proposed hierarchical architecture. To be more specific, the last layers of these networks were replaced with a set of binary classifiers with a flat architecture for each level of the disease classification. As shown in **Table 1**, the computational costs for both the training and the inference stage were comparable for all models. However, with the proposed hierarchical architecture, our algorithm outperformed all four existing methods in most of the diseases. For example, as shown in **Figure 7**, for level 1 disease identification—such as glaucoma—our framework achieved AUC 0.958, whereas ResNet, DenseNet, Inception-v3, and Ensemble methods achieved AUC 0.913, 0.940, 0.928, and 0.899, respectively. Similarly, for level 2 disease identification, such as ocular surface neoplasm, our framework achieved

TABLE 4 | Recall comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.91**	0.9	0.88	0.89	0.86
Ocular surface (*n* = 2,018)	**0.9**	0.89	0.86	0.88	0.85
Ocular neoplasm (*n* = 251)	**0.86**	0.82	0.8	0.84	0.8
Normal surface (*n* = 205)	**0.81**	0.75	0.8	0.75	0.8
Weighted average	**0.90**	0.88	0.86	0.87	0.85
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.9**	0.89	0.86	0.88	0.85
Conjunctivitis (*n* = 372)	**0.75**	0.73	0.73	0.74	0.73
Cornea degeneration (*n* = 137)	0.78	0.78	0.78	**0.79**	0.77
Cornea infectious (*n* = 1,098)	**0.94**	0.92	0.89	0.9	0.86
Intraocular neoplasma (*n* = 107)	**0.95**	0.9	0.82	0.82	0.86
Cornea non-infectious (*n* = 297)	0.78	0.8	**0.82**	0.8	0.76
Ocular surface neoplasm (*n* = 144)	**0.86**	0.79	0.73	0.76	0.76
Scleritis (*n* = 114)	**0.91**	0.87	0.87	0.87	0.87
Normal surface (*n* = 205)	0.81	0.8	0.81	0.8	**0.82**
Weighted average	**0.88**	0.86	0.84	0.85	0.83
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.91**	0.89	0.87	0.88	0.86
Vitreoretinal disease (*n* = 2,283)	**0.98**	0.97	0.95	0.96	0.95
Normal fundus (*n* = 323)	0.91	**0.92**	0.88	0.89	0.86
Weighted average	**0.96**	0.94	0.92	0.93	0.92
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.92**	0.9	0.87	0.88	0.84
Macular disease (*n* = 480)	**0.85**	0.82	0.8	0.79	0.79
Optic nerve disease (*n* = 467)	0.86	**0.88**	0.84	0.84	0.84
Refractive error (*n* = 156)	**0.83**	0.81	0.8	0.77	0.81
Retinal degeneration (*n* = 138)	0.82	**0.86**	0.79	0.82	0.79
Retinal detachment (*n* = 584)	**0.83**	0.79	0.77	0.78	0.75
Retinal vascular disease (*n* = 458)	**0.86**	0.84	0.82	0.84	0.8
Normal fundus (*n* = 323)	0.89	**0.91**	0.88	0.89	0.86
Weighted average	**0.87**	0.86	0.83	0.83	0.81

Bold value means "Best performance".

AUC 0.949, whereas ResNet, DenseNet, Inception-v3, and Ensemble methods achieved AUC 0.897, 0.896, 0.919, and 0.894, respectively. More detailed comparison results can be found in **Tables 2**–**5**.

Saliency Maps

To show the interpretation of the proposed framework, we also created heatmaps via the gradient-weighted class activation mapping (Grad-CAM) algorithm (29), which can produce visual explanations for CNN-based deep learning models. Grad-CAM uses the gradient information flowing into the last convolutional layer to understand the importance of each neuron for a decision of interest, thereby highlighting the important regions in the image for prediction. It first computes the gradient of the score for a given class with respect to feature maps of a convolutional layer. Then, these gradients are average-pooled to obtain the neuron importance weights. Finally, the coarse heatmap for a given class is generated via a weighted combination of forward activation maps followed by a ReLU function. As illustrated in **Figure 8**, the generated heatmaps helped indicate the potential corneal lesion regions for further examination, thereby establishing prediction trust and interpretation for physicians.

DISCUSSION

In this study, we demonstrated the effectiveness of the proposed hierarchical deep learning framework in

TABLE 5 | Precision comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.88**	0.85	0.81	0.84	0.8
Ocular surface (*n* = 2,018)	**0.96**	0.94	0.93	0.93	0.92
Ocular neoplasm (*n* = 251)	**0.7**	0.68	0.67	**0.7**	0.63
Normal surface (*n* = 205)	0.59	**0.75**	0.71	**0.75**	0.71
Weighted average	**0.90**	0.88	0.86	0.88	0.85
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.91**	0.89	0.86	0.87	0.85
Conjunctivitis (*n* = 372)	**0.67**	0.66	0.63	0.65	0.61
Cornea degeneration (*n* = 137)	0.64	0.64	0.6	**0.7**	0.61
Cornea infectious (*n* = 1,098)	**0.95**	0.94	0.92	0.93	0.92
Intraocular neoplasma (*n* = 107)	**0.68**	0.62	0.62	0.62	0.58
Cornea non-infectious (*n* = 297)	0.88	0.9	**0.91**	0.89	0.87
Ocular surface neoplasm (*n* = 144)	**0.99**	0.98	0.98	**0.99**	0.97
Scleritis (*n* = 114)	**0.98**	0.97	0.96	**0.98**	0.95
Normal surface (*n* = 205)	0.87	0.92	**0.93**	0.92	**0.93**
Weighted average	**0.88**	0.87	0.85	0.86	0.84
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.94**	0.91	0.86	0.88	0.86
Vitreoretinal disease (*n* = 2,283)	**0.98**	0.96	0.95	0.96	0.94
Normal fundus (*n* = 323)	0.91	0.87	0.93	0.91	**0.95**
Weighted average	**0.96**	0.94	0.93	0.93	0.92
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.95**	0.93	0.89	0.9	0.88
Macular disease (*n* = 480)	**0.7**	0.66	0.62	0.64	0.62
Optic nerve disease (*n* = 467)	0.82	**0.85**	0.8	0.81	0.79
Refractive error (*n* = 156)	0.96	0.96	**0.99**	0.96	0.96
Retinal degeneration (*n* = 138)	0.82	**0.86**	0.79	0.79	0.76
Retinal detachment (*n* = 584)	**0.9**	0.87	0.84	0.87	0.81
Retinal vascular disease (*n* = 458)	**0.93**	0.89	0.89	0.89	0.86
Normal fundus (*n* = 323)	0.91	0.92	**0.93**	0.92	0.92
Weighted average	**0.88**	0.86	0.84	0.85	0.82

Bold value means "Best performance".

identifying most causes of visual impairment diseases worldwide. Training the proposed hierarchical deep learning framework on eye images captured using commonly available equipment, we outperformed the performance of 40 board-certified ophthalmologists on 13 clinical cases. Further assessment of 4,670 cases in a tertiary eye center also demonstrated that the proposed framework achieved a high identification accuracy rate for different visual impairment diseases compared with that of human graders in a clinical setting.

Although we acknowledge that the clinical impression and diagnosis by an ophthalmologist are based on contextual factors beyond the visual inspection of the eye, the ability to classify eye images with the accuracy of a board-certified ophthalmologist has the potential to profoundly expand access to vital medical care. It has the potential to aid the delivery of eye disease screening in developed and developing countries in a manner that is inexpensive, efficient, and easily accessible. It can also be used to provide eye care guiding services in communities and assist doctors in diagnosing visual impairment diseases.

To validate this technique across the full distribution and spectrum of visual impairment diseases encountered in a clinical setting, further research is necessary to evaluate performance in a large community screening setting. This method is primarily constrained by data and can be validated for more visual conditions if sufficient training examples are provided.

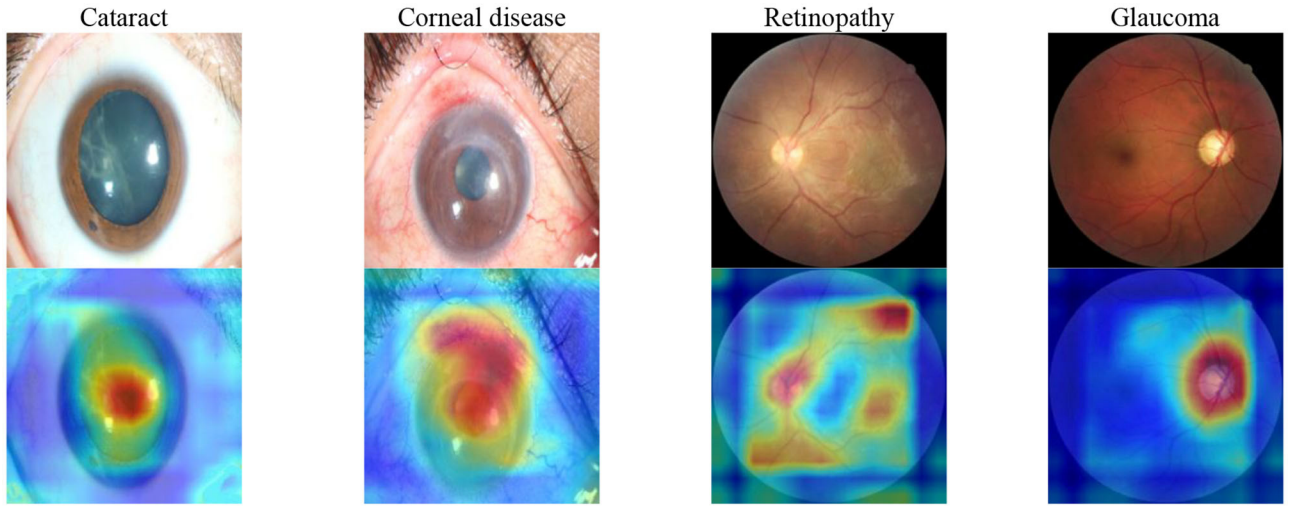

FIGURE 8 | Saliency maps for images with various common visual impairment diseases. These visualizations are generated automatically, locating regions for closer examination after a patient is seen by a consultant ophthalmologist. The bluer the color, the lower the attention of the model; the redder the color, the higher the attention of the model. Visualization maps are generated from deep learning features.

In this study, we applied multiple train–test splits via a 5-fold cross-validation where we randomly divided the entire image dataset into five subsets. Splitting data with respect to patients instead of images is indeed a better strategy; however, the dataset we had did not contain user identification information after data anonymization. We added this as a limitation of our study and would maybe explore it as future work. We would also conduct further experiments with publicly available datasets (such as EyePACS; Kaggle) as one of the future works. In the future, it may also be important to investigate different types of common patient metadata, such as genetic factors, patient history, and other clinical data that may influence a patient's risk of visual impairment diseases. Adding this information to the classification model may yield insightful information outside of strictly imaging information, potentially enhancing the diagnostic accuracy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Shanghai Eye and EENT Hospital (EENTIRB20170607). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and JH: conception and design. XL, JH, and JC: administrative support. JH, LG, JX, YL, and XS: provision of study materials or patients. JH, XL, HG, ZY, and JL: collection and assembly of data. YG: data analysis and interpretation. All authors: manuscript writing and final approval of manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (81970766 and 8217040684), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shanghai Innovation Development Program (2020-RGZN-02033), the Shanghai Key Clinical Research Program (SHDC2020CR3052B); LG was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), EXC 2026, Cardio-Pulmonary Institute (CPI), Project ID 390649896; and Guizhou Science and Technology Program (GZWJKJ2018-1-003).

ACKNOWLEDGMENTS

We thank the residents and faculties of the Department of Ophthalmology at the Shanghai Eye and ENT Hospital and the Affiliated Hospital of Guizhou Medical University for participating in our tests. We also thank Drs. Xiaobo Yu, Rui Jiang, Jingyi Cheng, and Lijia Tian from Fudan University for their help with project coordination. We thank the support from China National GeneBank.

REFERENCES

1. Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob Health*. (2017) 5:e888–97. doi: 10.1016/S2214-109X(17)30293-0
2. Xu L, Wang Y, Li Y, Wang Y, Cui T, Li J, et al. Causes of blindness and visual impairment in urban and rural areas in Beijing: the Beijing eye study. *Ophthalmology*. (2006) 113:1134.e1-11. doi: 10.1016/j.ophtha.2006.01.035
3. Zhao J, Xu X, Ellwein LB, Cai N, Guan H, He M, et al. Prevalence of vision impairment in older adults in rural china in 2014 and comparisons with the 2006 china nine-province survey. *Am J Ophthalmol.* (2018) 185:81–93. doi: 10.1016/j.ajo.2017.10.016
4. Rosenblatt TR, Vail D, Saroj N, Boucher N, Moshfeghi DM, Moshfeghi AA. Increasing incidence and prevalence of common retinal diseases in retina practices across the United States. *Ophthalmic Surg Lasers Imaging Retina.* (2021) 52:29–36. doi: 10.3928/23258160-20201223-06
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
6. Kermany D S, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. (2018) 172:1122–31.e9. doi: 10.1016/j.cell.2018.02.010
7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
8. Szegedy C, Vanhoucke V, Loffe S, Shlens J,Wojna Z. Rethinking the inception architecture for computer vision, 2016. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV (2016).
9. Menegola A, Fornaciali M, Pires R, Avila S, Valle E, et al. Towards automated melanoma screening: exploring transfer learning schemes. *arXiv[Prerpint].arXiv:1609.01228*. (2016). Available online at: https://www.researchgate.net/publication/307636270_Towards_Automated_Melanoma_Screening_Exploring_Transfer_Learning_Schemes
10. Setio AA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans Med Imaging.* (2016) 35:1160–9. doi: 10.1109/TMI.2016.2536809
11. Nie D, Zhang H, Adeli E, Liu L, Shen D. 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. *Med Image Comput Comput Assist Interv*. (2016) 9901:212–20. doi: 10.1007/978-3-319-46723-8_25
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. (2016) 316:2402–10. doi: 10.1001/jama.2016.17216
13. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. (2017) 124:962–9. doi: 10.1016/j.ophtha.2017.02.008
14. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. (2018) 125:1199–206. doi: 10.1016/j.ophtha.2018.01.023
15. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* (2017) 135:11706. doi: 10.1001/jamaophthalmol.2017.3782
16. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip AM, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. (2018) 125:549–58. doi: 10.1016/j.ophtha.2017.10.031
17. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol*. (2018) 256:259–65. doi: 10.1007/s00417-017-3850-3
18. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng*. (2017) 1:24. doi: 10.1038/s41551-016-0024
19. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic population with diabetes. *JAMA*. (2017) 318:2211–23. doi: 10.1001/jama.2017.18152
20. Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* (2018) 24:1342–50. doi: 10.1038/s41591-018-0107-6
21. Li W, Yang Y, Zhang K, Long E, He L, Zhang L, et al. Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. *Nature Bio Eng*. (2020) 4:767–77. doi: 10.1038/s41551-020-0577-y
22. Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605. Available online at: https://search.ebscohost.com/login.aspx?direct=true&db=asr&AN=36099312&lang=zh-cn&site=ehost-live
23. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell.* (2020) 42:318–27. doi: 10.1109/TPAMI.2018.2858826
24. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *IJCV*. (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y
25. Paszke A, Gross S, Chintala S, Chanan G, Yang E, Devito Z, et al. Automatic differeçntiation in PyTorch. *NIPS*. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach (2017). Available online at: https://openreview.net/forum?id=BJJsrmfCZ
26. 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
27. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). doi: 10.1109/CVPR.2017.243
28. van Veen HJ, Nguyen L, Dat T, Segnini A. *Kaggle Ensembling Guide*. (2015). Available online at: https://mlwave.com/kaggle-ensembling-guide/ (accessed February 6, 2018).
29. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *International Conference on Computer Vision (ICCV)*. Venice (2017). p. 618–26. doi: 10.1109/ICCV.2017.74

Conflict of Interest: BP and JC are employed by the company Complete Genomics 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 Digital Health

ORIGINAL RESEARCH
published: 02 July 2021
doi: 10.3389/fdgth.2021.677043

Seasonal Sleep Variations and Their Association With Meteorological Factors: A Japanese Population Study Using Large-Scale Body Acceleration Data

Li Li[1,2†], *Toru Nakamura*[1*†], *Junichiro Hayano*[3*] *and Yoshiharu Yamamoto*[4*]

[1] *Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, Japan,* [2] *Intasect Communications, Inc., Tokyo, Japan,* [3] *Graduate School of Medical Sciences, Nagoya City University, Nagoya, Japan,* [4] *Graduate School of Education, The University of Tokyo, Tokyo, Japan*

OPEN ACCESS

Edited by:
Liang Zhang,
Xidian University, China

Reviewed by:
Jian Guo,
RIKEN Center for Computational Science, Japan
Limin Hou,
Shanghai University, China

****Correspondence:***
Toru Nakamura
t-nakamura@sangaku.es.osaka-u.ac.jp
Junichiro Hayano
hayano@med.nagoya-cu.ac.jp
Yoshiharu Yamamoto
yamamoto@p.u-tokyo.ac.jp

†These authors have contributed equally to this work

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

Received: *07 March 2021*
Accepted: *02 June 2021*
Published: *02 July 2021*

Citation:
Li L, Nakamura T, Hayano J and Yamamoto Y (2021) Seasonal Sleep Variations and Their Association With Meteorological Factors: A Japanese Population Study Using Large-Scale Body Acceleration Data. Front. Digit. Health 3:677043. doi: 10.3389/fdgth.2021.677043

Seasonal changes in meteorological factors [e.g., ambient temperature (*Ta*), humidity, and sunlight] could significantly influence a person's sleep, possibly resulting in the seasonality of sleep properties (timing and quality). However, population-based studies on sleep seasonality or its association with meteorological factors remain limited, especially those using objective sleep data. Japan has clear seasonality with distinctive changes in meteorological variables among seasons, thereby suitable for examining sleep seasonality and the effects of meteorological factors. This study aimed to investigate seasonal variations in sleep properties in a Japanese population (68,604 individuals) and further identify meteorological factors contributing to sleep seasonality. Here we used large-scale objective sleep data estimated from body accelerations by machine learning. Sleep parameters such as total sleep time, sleep latency, sleep efficiency, and wake time after sleep onset demonstrated significant seasonal variations, showing that sleep quality in summer was worse than that in other seasons. While bedtime did not show clear seasonality, get-up time varied seasonally, with a nadir during summer, and positively correlated with the sunrise time. Estimated by the abovementioned sleep parameters, *Ta* had a practically meaningful association with sleep quality, indicating that sleep quality worsened with the increase of *Ta*. This association would partly explain seasonal variations in sleep quality among seasons. In conclusion, *Ta* had a principal role for seasonality in sleep quality, and the sunrise time chiefly determined the get-up time.

Keywords: sleep seasonality, meteorological factors, big data, acceleration data, Japanese

INTRODUCTION

Several meteorological factors, such as ambient temperature (*Ta*), humidity, and sunlight, have significant influences on human biological rhythms, including endogenous circadian rhythms (e.g., rectal temperature and melatonin rhythms) and sleep–wake cycles (1–3). Especially, seasonal climatic changes act as rhythmic external cues or perturbations on biological systems that regulate homeostatic and endogenous processes (1, 4, 5). The response of the systems to these seasonal inputs results in seasonal variations of biological variables, such as those of sleep properties.

Seasonal variations in sleep quality or prevalence of insomnia has been well-studied in terms of associations with characteristic seasonal changes in sunlight durations, such as the midnight sun in summer and the dark period in midwinter, especially among Nordic populations. In the epidemiological survey on Norwegian sleep using questionnaires, insomnia was more frequent in winter than in other seasons of the year (6). Other Nordic interview surveys demonstrated that the prevalence of reported insomnia, particularly sleep onset problems, increased from summer to winter in northern Norway but decreased in the southern regions (7). Meanwhile, in a general population in Finland, the prevalence of sleep dissatisfaction increased during summer (8).

Those sleep seasonality are often explained by an entrainment of the circadian time-keeping system to photoperiodic changes (5, 9, 10). However, interestingly, people living in areas with limited daylight variations had significant sleep seasonality (11, 12). For example, in a survey conducted among young Africans living in a dry tropical area, the number of awakenings increased during hot season (11). Furthermore, polysomnography (PSG) revealed that European expatriates living in a similar tropical climate showed seasonal differences in sleep quality and that sleep quality was significantly associated with *Ta* (12). Hence, seasonal sleep variations could not be fully explained by the sole basis of photoperiodic changes among seasons, and sleep seasonality is probably affected by the modulation of thermoregulatory processes passively induced by climatic temperature alterations (5, 11, 12).

Indeed, both laboratory and real-life settings have shown significant *Ta* effects on sleep; a study conducted under a temperature-controlled laboratory reported that *Tas* outside a thermoneutral zone were destructive to sleep (13). Further, subtle manipulations of skin temperature improved sleep latency (SL) in the elderly (14), while sleep depth enhanced in young adults (15), without causing alterations in core body temperatures. Even in a field-based study participated by the elderly with actigraphy, sleep disturbances were significantly related to skin temperature fluctuations (16). Therefore, ambient climate (e.g., bedroom climate), which possibly affects skin temperature, has strong modulating effects on sleep quality (17).

Despite that sleep seasonality and its relationship with meteorological factors have been extensively reported (6–8, 11–13, 17–20), large-scale population studies remain limited. Besides, almost all population studies largely relied on subjective sleep assessments. Though sleep seasonality has been intensively examined in high-latitude countries (e.g., Nordic countries) or low-latitude areas (e.g., tropical areas close to the equator), that in middle-latitude countries (e.g., temperate zone areas) has not been well-elucidated. Furthermore, most studies examined the effects of only a single meteorological factor on sleep, without considering the comprehensive effects of various meteorological factors (e.g., sunrise time, *Ta*, and humidity).

In examining the seasonal influences of meteorological factors on sleep, Japan is the best location because it is situated in a temperate zone with four distinctive, meteorologically separated seasons (spring, summer, autumn, and winter). Meteorological variables such as *Ta*, humidity, and day length change remarkably among seasons; for instance, in Tokyo, the monthly-based mean atmospheric air temperature varies from a few degrees to roughly 30° throughout a year, and the sunrise time changes from 4:30 AM to 7:00 AM approximately (**Figure 1**).

Although sleep seasonality is poorly investigated using objective measures in Japanese populations, two distinctive studies have been published (16, 19). An actigraphic study in the elderly reported the decrease of total sleep time (TST) and sleep efficiency (SE) and the increase of SL and wake time after sleep onset (WASO) in summer in comparison with those in winter (16), although the sample size is small. Another sleep study using a contactless biomotion sensor also reported the significant increase of WASO and decrease of SE in summer (19). However, these two previous studies had some inconsistencies in sleep parameter values. For example, the SE in the former study declined ∼10% from winter to summer (winter: 91%, summer: 81%), but that in the latter declined slightly (winter: 88%, summer: 86%).

Very recently, we examined the effects of age and gender on sleep among Japanese individuals by using a large-scale trunk acceleration data recorded from around 80,000 Japan residents (21, 22) (**Figure 2**). In that study, we developed an algorithm to determine sleep–wake states from the acceleration data using machine learning approaches and then obtained objective sleep parameters (e.g., sleep duration and SE). The present study aimed to examine the seasonal variations of sleep parameters in a Japanese population by using large-scale objective sleep data and to identity which meteorological factor significantly contributed to seasonal variations in each sleep parameter, if they exist, by multiple regression analysis combined with a bootstrapping method. In other words, this study is a comprehensive sleep research that used objective sleep measures to examine the effects of various ambient meteorological factors on Japanese habitual sleep at the population level.

MATERIALS AND METHODS

Acceleration Database—ALLSTAR Research Project

We used a database constructed by the ALLSTAR research project (23–25). The ALLSTAR database has been thoroughly explained elsewhere (23–25). Briefly, the database stores 24-h electrocardiography (ECG) data and tri-axial acceleration data measured by Holter recorders (Cardy Series; SUZUKEN Co., Ltd.) for clinical purposes by medical facilities all over Japan (47 prefectures in total). Since November 2007, the database has stored more than 300,000 analyzable ECG data (sampling frequency, 250 Hz) and ∼80,000 acceleration data simultaneously measured with ECG (sampling frequency, 31.25 Hz), with accompanying information, including the patient's age and gender, the recording date and time, and location (the medical facility's postal code). Considering that Holter monitoring is generally conducted in natural daily circumstances, not in laboratory settings, over 24 h without any restrictions affecting the patient's daily activities, we can access the patient's physiological data (e.g., acceleration data) during habitual sleep.

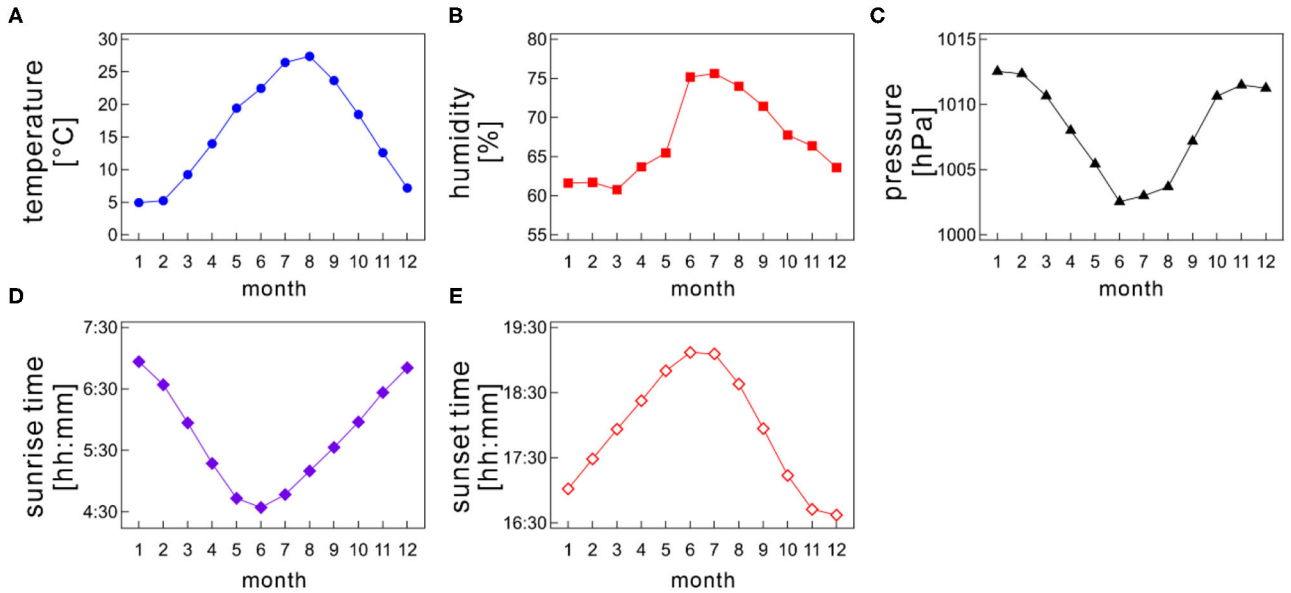

FIGURE 1 | Monthly variations of meteorological variables; **(A)** atmospheric air temperature, **(B)** humidity, **(C)** barometric pressure, **(D)** sunrise time, and **(E)** sunset time as a function of month. The monthly mean values were calculated for each month by averaging the daily meteorological data recorded at a prefectural capital of each medical facility on each Holter recording date.

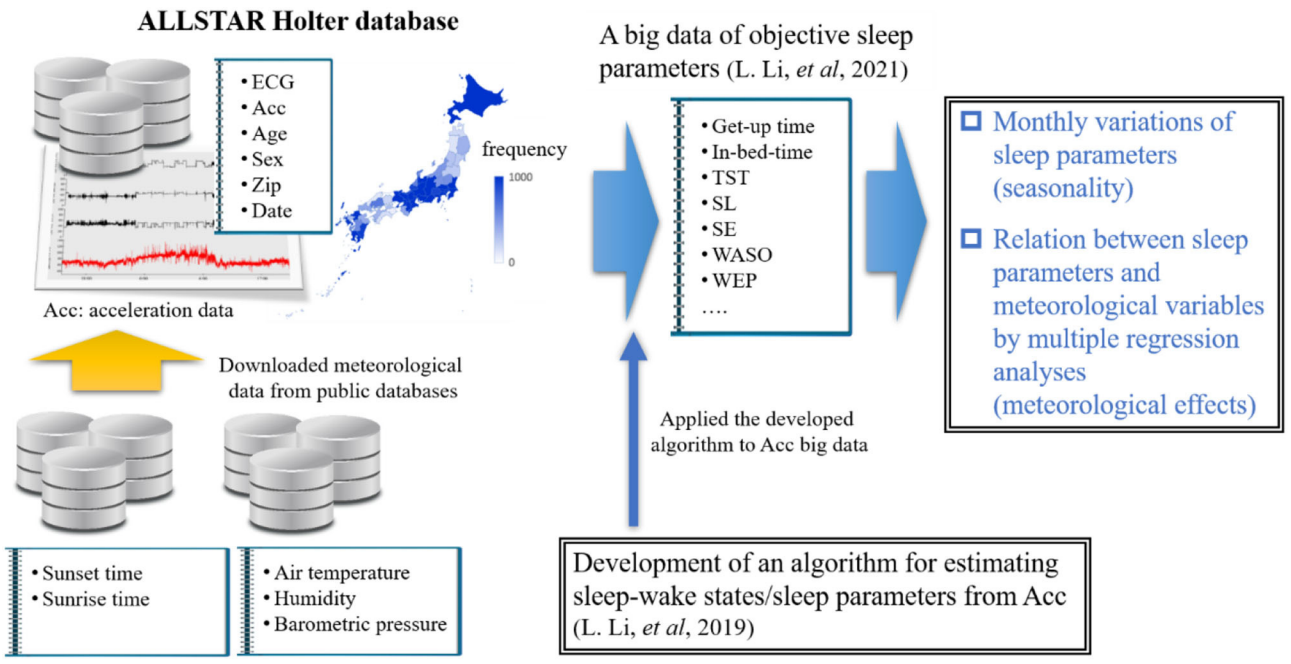

FIGURE 2 | A schematic flow chart of whole research.

Samples

The dataset we used is the same as that reported in our previous study (22). We utilized 68,604 individual acceleration data (30,485 males, 37,951 females, and 168 individuals with unknown gender; age range: 10–89 years old; data length > 20 h) gathered from 2010 to 2016 across Japan. These data were recorded by more than 1,500 medical facilities in 47 prefectures in Japan. **Table 1** summarizes age and monthly distributions of the samples. Further, **Table 2** shows the mean subjects' age (± standard deviation) stratified by month. The ethics committee of Osaka University approved our study, which conformed to the Declaration of Helsinki.

TABLE 1 | Number of samples stratified by month and age group.

Age group	Sample size	Month											
		1	2	3	4	5	6	7	8	9	10	11	12
10s	1,314	70	64	101	103	169	180	160	135	92	84	63	93
20s	1,421	111	109	139	114	133	134	144	95	112	121	89	120
30s	2,816	269	208	234	206	216	245	268	186	204	263	245	272
40s	5,448	507	435	508	411	447	452	454	350	376	478	491	539
50s	7,361	694	674	688	559	596	580	640	458	506	682	668	616
60s	14,727	1,305	1,250	1,457	1,264	1,179	1,219	1,085	858	1,067	1,434	1,363	1,246
70s	21,710	1,856	1,783	2,160	1,824	1,780	1,846	1,698	1,242	1,667	2,172	1,939	1,743
80s	13,806	1,157	1,081	1,283	1,197	1,156	1,227	1,084	856	1,100	1,371	1,233	1,061
Total	68,603	5,969	5,604	6,570	5,678	5,676	5,883	5,533	4,180	5,124	6,605	6,091	5,690

Note that the recording of month was missed for one subject.

TABLE 2 | Sleep parameter values by month.

Month	Sample number	Age mean ± SD	TST [min]	In-bed time [hh:mm]	Get-up time [hh:mm]	SL [min]	SE [%]	WASO [min]	WEP [count]
1	5,969	66.0 ± 15.7	437.8 ± 1.8	22:19 ± 0:01	6:27 ± 0:01	14.1 ± 0.3	92.4 ± 0.1	36.7 ± 0.6	3.78 ± 0.05
2	5,604	66.3 ± 15.4	430.6 ± 1.8	22:26 ± 0:01	6:27 ± 0:01	13.4 ± 0.3	92.3 ± 0.1	36.7 ± 0.7	3.81 ± 0.05
3	6,570	66.4 ± 15.8	426.2 ± 1.7	22:20 ± 0:01	6:18 ± 0:01	13.8 ± 0.3	92.0 ± 0.1	37.8 ± 0.6	3.94 ± 0.05
4	5,678	66.7 ± 16.1	421.0 ± 1.8	22:17 ± 0:01	5:59 ± 0:01	14.7 ± 0.3	91.7 ± 0.1	39.2 ± 0.7	4.19 ± 0.06
5	5,676	65.6 ± 17.2	409.1 ± 1.8	22:22 ± 0:01	6:08 ± 0:01	15.0 ± 0.3	90.7 ± 0.1	42.5 ± 0.7	4.81 ± 0.06
6	5,883	65.6 ± 17.2	402.6 ± 1.7	22:17 ± 0:01	5:59 ± 0:01	16.0 ± 0.3	90.4 ± 0.1	43.5 ± 0.7	5.24 ± 0.06
7	5,533	64.7 ± 17.5	398.6 ± 1.9	22:23 ± 0:01	6:06 ± 0:01	16.0 ± 0.3	89.3 ± 0.2	48.0 ± 0.7	5.9 ± 0.07
8	4,180	64.9 ± 17.6	404.4 ± 2.2	22:13 ± 0:02	6:02 ± 0:02	15.8 ± 0.3	89.4 ± 0.2	49.0 ± 0.9	6.07 ± 0.08
9	5,124	66.7 ± 16.3	404.7 ± 1.8	22:13 ± 0:01	6:01 ± 0:01	16.4 ± 0.4	89.6 ± 0.2	47.5 ± 0.7	5.67 ± 0.07
10	6,605	66.9 ± 15.6	417.2 ± 1.6	22:15 ± 0:01	6:10 ± 0:01	14.8 ± 0.3	90.7 ± 0.1	43.3 ± 0.7	4.73 ± 0.06
11	6,091	66.7 ± 15.3	425.4 ± 1.7	22:19 ± 0:01	6:17 ± 0:01	14.1 ± 0.3	91.9 ± 0.1	39.0 ± 0.7	4.05 ± 0.05
12	5,690	65.3 ± 16.3	429.7 ± 1.8	22:25 ± 0:01	6:23 ± 0:01	13.4 ± 0.3	92.6 ± 0.1	35.3 ± 0.7	3.69 ± 0.05

Sleep parameter values are represented as mean ± SEM. SD, standard deviation.

Sleep–Wake Inference From the Acceleration Data Using Machine Learning

Sleep and wake states are often inferred according to body movements measured by wearable devices (26–29). Following these approaches, we recently developed algorithms to accurately estimate minute-by-minute sleep–wake states, as well as sleep parameters, from trunk acceleration data measured by the Holter recorder. In this study, we utilized the sleep parameter values calculated by our algorithms in our previous work (21, 22). Our algorithms are summarized below.

Using a support vector machine (SVM), we constructed a sleep–wake classifier (30, 31) that converted tri-axial trunk acceleration data into a sequence of "sleep" and "wake" labels, with 1-min time resolution using the statistical features extracted from the acceleration data. More specifically, we used upper-body tilt angles and local variances of trunk acceleration data as input vectors to the machine. Our method was validated by comparing the outputs of a watch-type sleep monitor (referred to as an actigraph) manufactured by Ambulatory Monitoring Inc. (AMI, Ardsley, NY). An AMI actigraph correctly distinguishes sleep from wakefulness with high accuracy (>90%) (32, 33) and high sensitivity (>95%) (33, 34) compared with PSG, which is the gold standard for sleep assessment. Therefore, the actigraph has been widely used in sleep studies as a PSG substitute (26, 29, 32). Our validation study demonstrated that our SVM-based method was consistent with the AMI actigraph (accuracy = 94.4% ± 3.8%, specificity = 94.2% ± 5.2%, sensitivity = 94.8% ± 3.9%, and F1-score = 92.0 ± 4.5) (21, 22). Note that while we used a classical machine learning approach for the sleep–wake classification, state-of-the-art methods, such as ensemble tree-based algorithms [e.g., extreme gradient boosting (XGBoost) (35), or light gradient boosting machine (LightGBM) (36, 37)], or deep neural networks [e.g., long short-term memory (38–40)], may improve classification performance significantly.

Sleep Parameters

We examined seasonality of the following seven sleep parameters (22, 28, 29): in-bed time, get-up time, SL, WASO, wake episodes (WEP), TST, and SE. In-bed time is the clock time when a patient gets into bed to sleep and then switches the light off, while get-up

time is when a patient finally wakes up in the morning. In-bed time and get-up time are often ascertained by using data from the event marker of an actigraph, sleep diary, or ambient light sensor (29). However, such data were unavailable in the database; hence, we determined those timings from the acceleration data (21, 22). Moreover, SL refers to the time it took a patient to fall asleep; it is the number of minutes between in-bed time and sleep onset, where sleep onset is the time at the start of the first 10 consecutive minutes of sleep after in-bed time. WASO is the sum of the awakening minutes from sleep onset to the get-up time. WEP refers to the number of awakenings between sleep onset and get-up time. TST is the number of minutes asleep between sleep onset and get-up time; it can be calculated by subtracting SL and WASO from time in bed (practically, time in bed was defined by the period between in-bed time and get-up time). Lastly, SE is the ratio of TST to time in bed multiplied by 100. Note that these sleep parameters were strongly related with dynamics in sleep structures commonly assessed by PSG.

Meteorological Variables

Japan locates in the northern hemisphere, and its climate is separated into four seasons, namely, spring, summer, autumn, and winter. Many meteorological variables, such as *Ta* and photoperiod length, distinctively change among seasons (**Figure 1**). Each season generally lasts 3 months. Monthly average of atmospheric air temperature is highest during summer (June–August) and lowest during winter (December–February). Meanwhile, spring (March–May) and autumn (September–November) bridge a gap between summer and winter (**Figure 1A**). Therefore, atmospheric data show an annual sinusoidal pattern. Japan experiences a short rainy season, which generally lasts from the beginning of June to the middle of July, making the area dampish (**Figure 1B**).

The sunrise time and sunset time also varies between seasons (**Figures 1D,E**). In summer, the sun rises earlier and sets later, causing a longer daytime; conversely, the sun rises later and sets earlier in winter, resulting in a shorter daytime. Thus, the difference in the daytime length between summer and winter is ∼3.5 h. Of note, the daylight-saving time system has not yet been introduced in Japan.

We downloaded daily meteorological data (mean *Ta* [°C], humidity degree (%), and barometric pressure [hPa]) measured in the prefectural capital of each medical facility on each Holter recording date from the open public database of Japan Meteorological Agency (41). Considering that Holter recordings were performed over 2 consecutive days to obtain continuous 24-h data, we used the average *Ta*, humidity, and barometric pressure values over the recording days.

The sunset/sunrise time on the start/end day of the Holter recording was downloaded from the public database of the National Astronomical Observatory of Japan (42).

Statistics

Seasonality of Sleep Parameters

The seasonality (specifically, monthly variations) in each sleep parameter was examined using a generalized linear model (GLM). In fitting a GLM, the month of Holter recording was the categorical variable. Patient's gender and age were also included into the GLM because the gender and age effects were significant in all sleep parameters (21, 22). The age was categorized into eight groups by 10-year intervals (**Table 1**).

In addition to the main effects of these categorical variables (i.e., age group, gender, and month), the interaction term between age and gender was considered as a possible factor affecting the sleep parameter values. When the interaction term was not significant, a separate GLM without it was created and then fitted to the data again. If the interaction term was significant, we stratified the data by gender or age and then tested simple main effects (i.e., pairwise comparisons) with Bonferroni correction for multiple comparisons. In fitting GLMs, in-bed time and get-up time values were represented as an elapsed time (in minutes) counted from 0:00 on the start day of a Holter recording; hence, the values ranged from 0 to 2,880 (1,440 min × 2 days). Similarly, the sunrise time and sunset time were represented as an elapsed time (in minutes) counted from 0:00 of the start and end day of the recording. The sleep parameter values between July and other months were compared.

All statistical data were analyzed using SAS software version 9.04 (SAS Institute, Cary, NC, USA). In addition, *p*-values were adjusted by Bonferroni adjustment correction for multiple comparisons. To avoid potential inferential biases caused by a large sample size, we considered $p < 0.01$ statistically significant (43, 44). The results were expressed as the mean and the standard error of the mean (SEM) except for the coefficient values in multiple regression analysis explained below.

Multiple Linear Regression Analysis

To identify which meteorological variable (i.e., *Ta*, humidity, barometric pressure, sunset time, and sunrise time) contributed to seasonal variations in each sleep parameter, we further evaluated multiple linear regression models in which each sleep parameter was a response variable and meteorological variables were the explanatory variables.

Several meteorological variables highly correlated with each other (e.g., Pearson's correlation was $r = 0.70$ between *Ta* and sunrise time). To avoid the variance inflation caused by high multicollinearity in the regression analysis, we used a shrinkage-based variable selection method, which allowed the exclusion of redundant variables from a regression model. We also combined a model averaging method based on a bootstrapping algorithm [PROC GLMSELECT, ModelAverage (45), in SAS software] to search for a robust and parsimonious model. Each step was explained below in detail.

***Variable selection step*:** Least Absolute Shrinkage and Selection Operator (LASSO) (46), which is a popular method for selecting shrinkage variables, can effectively select important explanatory variables from a set of candidates potentially correlated with a response variable, and then estimate the coefficient values of regressors simultaneously. LASSO belongs to a particular class of penalized least square regression with the sum of absolute values of regression coefficients (or L1 norm), making some coefficients estimated to be zero. In this study, we employed the modified standard LASSO called the adaptive LASSO algorithm (47); in forming the LASSO constraint (i.e., penalized

term), weights were applied to each regression coefficient, leading to better performance in identifying a parsimonious model. If all entering explanatory variables were not significant at $p < 0.01$, the selection process was terminated, and from the sequence of models obtained by the selection process, the final model was chosen using the Schwarz Bayesian Criterion (48). Hence, each regression coefficient was ensured to be significant ($p < 0.01$). To adjust both the gender and age effects, we also included the categorical variables of gender and age in the regression models.

***Model averaging step*:** We employed a model averaging method based on a bootstrap method (45, 49) to perform more stable inferences of models. Model selections by the adaptive LASSO regression were repeated on bootstrap samples. The model selected by variable selection possibly varies from sample to sample; therefore, the importance of each explanatory variable was scored by using the number of times it was incorporated in the selected model. Considering that frequently selected explanatory variables were regarded as true underlying regressors, we constructed a final model that used merely the variables above the selection frequency cutoff value. In model averaging, we calculated the ensemble average of each coefficient value estimated by fitting the model to each bootstrap sample. In each bootstrap analysis, 5,000 samples were randomly resampled from the entire dataset. The frequency cutoff value in this study was 70%. Effects of the choice of cutoff values were also examined.

RESULTS

Monthly Variations of Sleep Parameters

Figure 3 shows the monthly average values of each sleep parameter (TST, in-bed time, get-up time, SL, SE, WASO, and WEP) as a function of month. The mean TST showed a clear annual cycle with shorter durations during summer and longer durations during winter (**Figure 3A**). Specifically, it was shortest in July (6.64 ± 0.03 h) and longest in January (7.30 ± 0.03 h), showing a difference of ∼40 min monthly.

Seasonal variations were similar in both get-up time (**Figure 3C**) and SE (**Figure 3E**), with a nadir in summer. The mean get-up time was earliest in June (5:59 AM) and latest in January (6:27 AM). The mean difference of get-up time between summer and winter was ∼24 min (overall mean get-up time: 6:02 AM during summer and 6:26 AM during winter). The mean SE decreased slightly but significantly by ∼2.7% in summer compared with that in winter (overall mean SE: 89.7 ± 0.1% during summer and 92.4 ± 0.1% during winter).

The mean in-bed time was almost constant across months; any significant monthly difference was not found between July and other months (**Figure 3B**). Overall mean in-bed time was 22:19 in our samples.

The monthly average of SL peaked in summer in an annual cycle, although the amplitude of differences among months was subtle (**Figure 3D**); the mean SL varied between 14.0 and 16.1 min.

Seasonality was noticeable in WASO (**Figure 3F**). The amount of time of WASO exceedingly increased during summer, with the longest duration of 49.0 ± 0.9 min in August. Conversely, the shortest duration of 35.3 ± 0.7 min was observed in December. Similarly, the number of wake episodes slightly, but significantly, increased during summer compared with the remaining seasons (**Figure 3G**).

Meteorological Effects on Sleep Seasonality

According to multiple regression analysis, three meteorological variables, namely, humidity, barometric pressure, and the sunset time, did not significantly contribute to the seasonality of any sleep parameter. **Table 3** summarizes the coefficient values of the final averaged model for each sleep parameter. **Figure 4** shows the scatter plots between the sleep parameter values and the meteorological variables shown in **Table 3**. When the frequency cutoff value was changed from 65 to 80%, the final averaged model consistently selected regressors shown in **Table 3**.

The *Ta* was selected as a significant regressor in the final averaged model for all sleep parameters, excepting get-up time. The *Ta* was negatively associated with TST and SE (coefficient value: −1.58 ± 0.04 for TST and −0.111 ± 0.004 for SE; **Figures 4A,C**) but positively correlated with SL, WASO, and WEP (coefficient value: 0.11 ± 0.01 for SL, 0.48 ± 0.01 for WASO, and 0.082 ± 0.002 for WEP; **Figures 4B,D**). The linear relations were considerably clear above 5°C (**Figures 4A–E**). Thus, sleep quality worsened as the *Ta* increased; this result possibly explained the worsening of sleep quality during summer. However, we also found a declining tendency in SE and WASO below 5°C (**Figures 4C,D**). These suggested a U-shaped correlation of these sleep parameters with *Ta*.

The sunrise time positively associated with get-up time (coefficient value: 0.182 ± 0.006; **Figure 4F**); this result probably explained the delay of get-up time in winter and the early get-up time in summer. The sunrise time also significantly correlated with SE, though the absolute magnitude of the regression coefficient was practically small (coefficient value: 0.005 ± 0.001); the effect size was below 1% even when the sunrise time changed from 4:30 AM to 7:00 AM. Hence, the sunrise time had practically no influence on SE. As well, the significant, but subtle negative relation was confirmed between sunrise time and WEP. The influence of sunrise time on sleep quality is thought to be limited.

DISCUSSION

The current study aimed (1) to examine seasonality in various sleep parameters (TST, in-bed time, get-up time, SL, SE, WASO, and WEP) by using a large-scale objective sleep data of a Japanese population and (2) to identify meteorological factors statistically associated with sleep seasonality. This study is the largest population-based research that used objective sleep data in real-life settings to examine sleep seasonality and its association with climatic factors in Japan.

Seasonality in Sleep Parameters

We found clear seasonal variations with an annual cycle in all sleep parameters, excluding in-bed time. Average monthly values of TST, get-up time, and SE showed a sinusoidal functional form with a nadir in summer, while mean SL, WASO, and WEP peaked during summer. Thus, sleep quality worsened

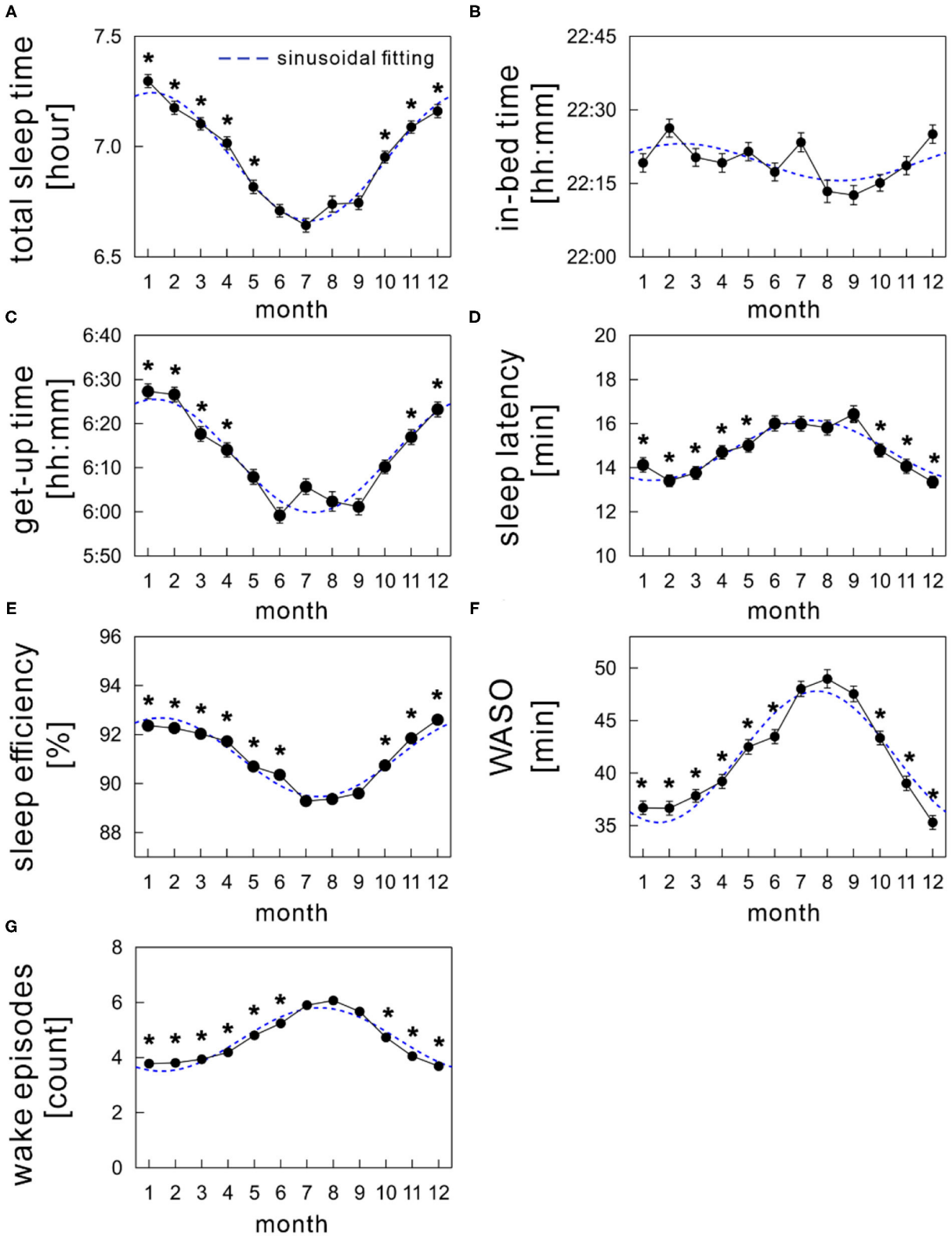

FIGURE 3 | Monthly variations of sleep parameters; **(A)** total sleep time, **(B)** in-bed time, **(C)** get-up time, **(D)** sleep latency, **(E)** sleep efficiency, **(F)** wake time after sleep onset (WASO), and **(G)** wake episodes. The mean values of each sleep parameter are shown as a function of month (solid black circles). The sinusoidal functional curve with 1-year period was fitted to the mean values of each sleep parameter (broken blue curve). The error bars are the standard error of mean. *indicates a significant difference from July ($p < 0.01$).

TABLE 3 | Coefficient values of the selected significant regressor by multiple regression analysis.

Sleep parameter	Meteorological variable (regressor)	
	Ambient temperature [°C]	Sunrise time (elapsed time) [min]
Get-up time [min]	-	0.18 ± 0.01
TST [min]	−1.58 ± 0.04	-
SL [min]	0.11 ± 0.01	-
SE [%]	−0.111 ± 0.004	0.005 ± 0.001
WASO [min]	0.48 ± 0.01	-
WEP [count]	0.082 ± 0.002	−0.0022 ± 0.0003

Values are represented as mean ± SD. The humidity, barometric pressure, and sunset time were not selected as significant regressors in any regression model. Therefore, the cells for those regressors are not shown. The bar (-) in a cell indicates that the corresponding regressor was not selected in the final regression model.
SE, sleep efficiency; SL, sleep latency; TST, total sleep time; WASO, wake time after sleep onset; WEP, wake episodes.

from winter to summer but then improved from summer to winter. These are partly comparable with previous research using objective sleep measures (16, 19), while there are some inconsistencies in sleep parameter values, such as magnitudes of seasonal differences or absolute values of SE. These discrepancies could be influenced by numerous factors, including differences in measurement devices, patients' age and gender distributions, and local climates. Furthermore, the incased frequency in WEP during summer could be related with increased prevalence of self-reported insomnia, especially difficulty in maintaining sleep, in a Japanese population in summer (20).

In our study, seasonal variations were not confirmed in the in-bed time compared with those in the get-up time. Under well-controlled laboratory conditions, both sleep and wake-up times in summer were significantly advanced (5). However, other studies that objectively assessed sleep in real-life settings could not find any seasonal shift in bedtime but wake-up time was significantly advanced during summer (5, 19). Therefore, bedtimes were less influenced by seasonal climate changes in real-life settings. We hypothesized that sociocultural factors (e.g., lifestyle, work, social role, and family) have a large impact on bedtimes in habitual sleep.

Meteorological Effects on Sleep Seasonality

The most noticeable finding of our study was the identification of meteorological factors contributing to seasonal variations in sleep parameters, using the robust multiple regression analysis. The analysis revealed that *Ta* chiefly determined seasonal variations in sleep quality (TST, SL, SE, WASO, and WEP) in real-life settings in the Japanese population. It would be valuable to address effects of a choice of different classes of sparse regressions. We tested a ridge regression (L2 penalty) (50, 51) and Elastic net (a combination of L1 and L2 penalties) (52). Both methods selected the identical regressors to those of LASSO in the final averaged models at the selection frequencies ranging from 65 to 80%. This indicates the robustness of our results.

The seasonal differences in sleep–wake cycles or sleep quality are commonly interpreted as a consequence of the entrainment of circadian rhythm to photoperiodic changes among seasons (4–7, 53–55). However, interestingly, meaningful contributions of sunlight durations to sleep quality were not detected in our study.

Our results indicated that sleep quality worsened as the *Ta* increases, suggesting the principal role of *Ta* for the seasonality in sleep quality. Further, SE and WASO exhibited a deteriorating trend at colder *Ta* (below 5°C), indicating that sleep quality worsened at colder or hotter *Ta*. These are supported by the results of previous studies based on actigraphy or contactless biomotion sensor (16, 19). Although the functional link between *Ta* and sleep has remained poorly understood, the contribution of a feedback system of skin temperature to sleep-regulating brain areas (preoptic area/anterior hypothalamus) has been suggested as a possible mechanism (56). Indeed, a direct manipulation of skin temperature revealed a notable effect on sleep propensity in the elderly with and without sleep insomnia (14). Without alternating the core temperature, the induction of a small increase (0.4°C) in skin temperature suppressed nocturnal wakefulness and shifted sleep to deeper stages in healthy young and elderly, as well as in patients with insomnia (15). These findings support the interpretation that seasonality in sleep quality was caused by the modulation of skin temperature induced by seasonal changes in *Ta*.

The get-up time did not correlate with *Ta*. This is explained by the difference in the timing of a peak or a nadir in annual cycle of get-up time and *Ta*; the mean get-up time was earliest in June, while the *Ta* was highest in Aug. Meanwhile, the sunrise time had a nadir in June, similar to get-up time. The results of the regression analysis reflect such phase differences between sleep parameters and meteorological variables.

Limitations

This study has several limitations. The first originates from an ambulatory monitoring in real-life circumstances. Behavioral thermoregulation, such as the use of air conditioning, clothing, and bedspreads, might affect our results because it likely changes both the actual skin and core body temperature. In addition, we did not consider the duration and intensity of light exposure. This limitation could be related to the lack of association between photoperiodic changes and sleep quality. We also did not control the regional differences. Considering that Japan covers several degrees of latitude (from 20 to 46° north), the meteorological variables largely differ between southern and northern areas; for example, the sun rises earlier in northern areas than in southern areas, and monthly *Ta*s are usually lower in northern areas than in southern areas.

Other significant limitations are related to the database. As discussed in our previous study (22), the database probably included selection biases because Holter recordings were usually obtained from patients suspected of having some form of a cardiovascular disease (57). In addition, other clinical conditions (e.g., sleep problems and depression) were not controlled because of the unavailability of such information. The effects of imbalanced age distribution of the samples would be remained. Further population studies controlled those factors might be

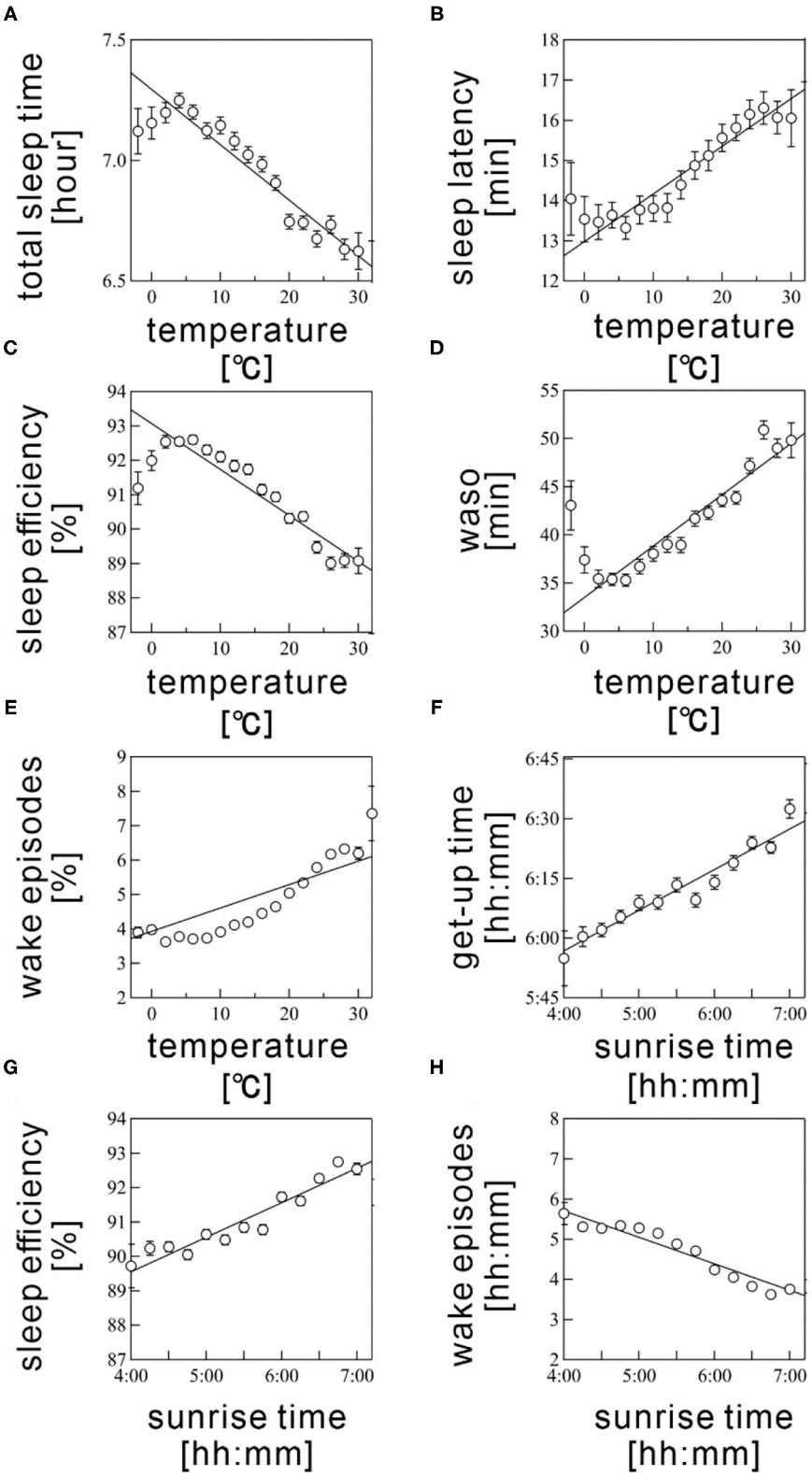

FIGURE 4 | Scatter plots between sleep parameters and meteorological variables. **(A)** total sleep time, **(B)** sleep latency, **(C)** sleep efficiency, **(D)** wake time after sleep onset (WASO), **(E)** wake episodes are shown as a function of *Ta*. **(F)** get-up time, **(G)** sleep efficiency, and **(H)** wake episodes are plotted as a function of sunrise time. Sleep parameter values were averaged every 5°C for *Ta* and 10 min for sunrise time. The error bars are the standard error of mean. The straight line represents a linear regression fit.

important. In addition, assessments of sleep structures (e.g., sleep stages) might provide more deeper insights into seasonal influence on nocturnal sleep. Nevertheless, our findings on sleep seasonality derived from the largest Japanese population are scientifically important and informative.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The database is available for academic research under owners concent. Requests to access these datasets should be directed to http://www.med.nagoya-cu.ac.jp/mededu.dir/allstar/.

AUTHOR CONTRIBUTIONS

LL and TN analyzed the data. LL, TN, JH, and YY contributed to manuscript preparation and revision. All authors contributed to data interpretation.

FUNDING

This work was supported in part by Grant-in-Aid for Scientific Research (B) [15H03095] (to TN) and Grant-in-Aid for Scientific Research (A) [17H00878, 20H00569] (to YY) from the Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

1. Sollberget A. Significance of biological rhythm study for human biometeorology. *Int J Biometeorol.* (1963) 7:193–220. doi: 10.1007/BF02184898
2. Okamoto-Mizuno K, Mizuno K, Michie S, Maeda A, Iizuka S. Effects of humid heat exposure on human sleep stages and body temperature. *Sleep.* (1999) 22:767–73.
3. Pandey J, Grandner M, Crittenden C, Smith MT, Perlis ML. Meteorologic factors and subjective sleep continuity: a preliminary evaluation. *Int J Biometeorol.* (2005) 49:152–5. doi: 10.1007/s00484-004-0227-1
4. Honma K, Honma S, Wada T. Phase-dependent shift of free-running human circadian rhythms in response to a single bright pulse. *Experientia.* (1987) 43:1205–7. doi: 10.1007/BF01945525
5. Kohsaka M, Fukuda N, Honma K, Honma S, Morita N. Seasonality in human sleep. *Experientia.* (1992) 48:231–3. doi: 10.1007/BF01930461
6. Husby R, Lingjaerde O. Prevalence of reported sleeplessness in northern Norway in relation to sex, age and season. *Acta Psychiatr Scand.* (1990) 81:542–7. doi: 10.1111/j.1600-0447.1990.tb05009.x
7. Pallesen S, Nordhus IH, Nielsen GH, Havik OE, Kvale G, Johnsen BH, et al. Prevalence of insomnia in the adult Norwegian population. *Sleep.* (2001) 24:771–9. doi: 10.1093/sleep/24.7.771
8. Ohayon MM, Partinen M. Insomnia and global sleep dissatisfaction in Finland. *J Sleep Res.* (2002) 11:339–46. doi: 10.1046/j.1365-2869.2002.00317.x
9. Binkley S, Tome MB, Crawford D, Mosher K. Human daily rhythms measured for one year. *Physiol Behav.* (1990) 48:293–8. doi: 10.1016/0031-9384(90)90316-V
10. Bliwise DL. Sleep in normal aging and dementia. *Sleep.* (1993) 16:40–81. doi: 10.1093/sleep/16.1.40
11. Buguet A, Hankourao O, Gati R. Self-estimates of sleep in african students in a dry tropical climate. *J Environ Psychol.* (1990) 10:363–9. doi: 10.1016/S0272-4944(05)80035-0
12. Montmayeur A, Buguet A. Sleep patterns of European expatriates in a dry tropical climate. *J Sleep Res.* (1992) 1:191–6. doi: 10.1111/j.1365-2869.1992.tb00037.x
13. Haskell EH, Palca JW, Walker JM, Berger RJ, Heller HC. The effects of high and low ambient temperatures on human sleep stages. *Electroencephalogr Clin Neurophysiol.* (1981) 51:494–501. doi: 10.1016/0013-4694(81)90226-1
14. Raymann RJ, Van Someren EJ. Diminished capability to recognize the optimal temperature for sleep initiation may contribute to poor sleep in elderly people. *Sleep.* (2008) 31:1301–9. doi: 10.5665/sleep/31.9.1301
15. Raymann RJ, Swaab DF, Van Someren EJ. Skin deep: enhanced sleep depth by cutaneous temperature manipulation. *Brain.* (2008) 131:500–13. doi: 10.1093/brain/awm315
16. Okamoto-Mizuno K, Tsuzuki K. Effects of season on sleep and skin temperature in the elderly. *Int J Biometeorol.* (2010) 54:401–9. doi: 10.1007/s00484-009-0291-7
17. Muzet A, Libert JP, Candas V. Ambient temperature and human sleep. *Experientia.* (1984) 40:425–9. doi: 10.1007/BF01952376
18. Friborg O, Bjorvatn B, Amponsah B, Pallesen S. Associations between seasonal variations in day length (photoperiod), sleep timing, sleep quality and mood: a comparison between Ghana (5 degrees) and Norway (69 degrees). *J Sleep Res.* (2012) 21:176–84. doi: 10.1111/j.1365-2869.2011.00982.x
19. Hashizaki M, Nakajima H, Shiga T, Tsutsumi M, Kume K. A longitudinal large-scale objective sleep data analysis revealed a seasonal sleep variation in the Japanese population. *Chronobiol Int.* (2018) 35:933–45. doi: 10.1080/07420528.2018.1443118
20. Suzuki M, Taniguchi T, Furihata R, Yoshita K, Arai Y, Yoshiike N, et al. Seasonal changes in sleep duration and sleep problems: a prospective study in Japanese community residents. *PLoS ONE.* (2019) 14:e0215345. doi: 10.1371/journal.pone.0215345
21. Li L, Nakamura T. An epidemiological sleep study based on a large-scale physical activity database. In: *The 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech2019)*. Osaka. (2019). p. 292–3. doi: 10.1109/LifeTech.2019.8883989
22. Li L, Nakamura T, Hayano J, Yamamoto Y. Age and gender differences in objective sleep properties using large-scale body acceleration data in a Japanese population. *Sci Rep.* (2021) 11:9970. doi: 10.1038/s41598-021-89341-x
23. Hayano J, Kiyono K, Yuda E, Yamamoto Y, Kodama I. Holter ecg big data project: allostatic state mapping by ambulatory ecg repository (allstar). *Int J Inform Res Rev.* (2018) 5:5617–24.
24. Hayano J, Ohashi K, Yoshida Y, Yuda E, Nakamura T, Kiyono K, et al. Increase in random component of heart rate variability coinciding with developmental and degenerative stages of life. *Physiol Meas.* (2018) 39:054004. doi: 10.1088/1361-6579/aac007
25. Hayano J, Kisohara M, Yoshida Y, Sakano H, Yuda E. Association of heart rate variability with regional difference in senility death ratio: ALLSTAR big data analysis. *SAGE Open Med.* (2019) 19:2050312119852259. doi: 10.1177/2050312119852259
26. Teicher MH. Actigraphy and motion analysis: new tools for psychiatry. *Harv Rev Psychiatry.* (1995) 3:18–35. doi: 10.3109/10673229509017161
27. Martin JL, Hakim AD. Wrist actigraphy. *Chest.* (2011) 139:1514–27. doi: 10.1378/chest.10-1872
28. Ancoli-Israel S, Martin JL, Blackwell T, Buenaver L, Liu L, Meltzer LJ, et al. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav Sleep Med.* (2015) 13(Suppl.1):S4–38. doi: 10.1080/15402002.2015.1046356
29. Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-based assessment of sleep parameters. *Ann Work Expo Health.* (2020) 64:350–67. doi: 10.1093/annweh/wxaa007
30. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018
31. Bishop CM. *Pattern Recognition and Machine Learning.* New York, NY: Springer. (2006).
32. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep.* (1994) 17:201–7. doi: 10.1093/sleep/17.3.201

33. De Souza L, Benedito-Silva AA, Pires MLN, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. *Sleep*. (2003) 26:81–5. doi: 10.1093/sleep/26.1.81
34. Jean-Louis G, Kripke DF, Mason WJ, Elliott JA, Youngstedt SD. Sleep estimation from wrist movement quantified by different actigraphic modalities. *J Neurosci Methods*. (2001) 105:185–91. doi: 10.1016/S0165-0270(00)00364-2
35. Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System*. (2016). Available online at: https://ui.adsabs.harvard.edu/abs/2016arXiv160302754C (accessed March 01, 2016). doi: 10.1145/2939672.2939785
36. Ke GL, Meng Q, Finley T, Wang TF, Chen W, Ma WD, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inform Process Syst*. (2017) 30.
37. Zhao X, Sun G. A multi-class automatic sleep staging method based on photoplethysmography signals. *Entropy*. (2021) 23:e23010116. doi: 10.3390/e23010116
38. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735
39. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. (2017) 25:1998–2008. doi: 10.1109/TNSRE.2017.2721116
40. Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, et al. Automatic human sleep stage scoring using deep neural networks. *Front Neurosci*. (2018) 12:781. doi: 10.3389/fnins.2018.00781
41. Japan Meteorological Agency. *Japan Meteorological Agency*. (2021). Available online at: https://www.jma.go.jp/jma/indexe.html (accessed February 24, 2021).
42. National Astronomical Observatory of Japan. *National Astronomical Observatory of Japan*. (2021). Available online at: https://www.nao.ac.jp/en/ (accessed February 24, 2021).
43. Lin M, Lucas Jr HC, Shmueli G. Too big to fail: large samples and the p-value problem. *Inform Syst Res*. (2013) 24:906–17. doi: 10.1287/isre.2013.0480
44. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci*. (2014) 7:342–6. doi: 10.1111/cts.12178
45. Burnham KP, Anderson DR, Burnham KP. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer (2002).
46. Tibshirani R. Regression shrinkage and selection *via* the Lasso. *J Royal Statist Soc Ser B Methodol*. (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x
47. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. (2006) 101:1418–29. doi: 10.1198/016214506000000735
48. Judge GG, Hill RC, William GE, Lutkepohl H, Lee TC. *The Theory and Practice of Econometrics*. New York, NY: John Wiley & Sons, Inc. (1985).
49. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall (1993). doi: 10.1007/978-1-4899-4541-9
50. Hoerl AE, Kennard RW. Ridge regression - applications to nonorthogonal problems. *Technometrics*. (1970) 12:69. doi: 10.1080/00401706.1970.10488635
51. Hoerl AE, Kennard RW. Ridge regression - biased estimation for nonorthogonal problems. *Technometrics*. (1970) 12:55. doi: 10.1080/00401706.1970.10488634
52. Zou H, Hastie T. Regularization and variable selection *via* the elastic net. *J Royal Statist Soc Ser B*. (2005) 67:768–8. doi: 10.1111/j.1467-9868.2005.00527.x
53. Winfree AT. *The Timing of Biological Clocks*. New York, NY: Scientific American Library : Distributed by W.H. Freeman. (1987).
54. Czeisler CA, Kronauer RE, Allan JS, Duffy JF, Jewett ME, Brown EN, et al. Bright light induction of strong (type 0) resetting of the human circadian pacemaker. *Science*. (1989) 244:1328–33. doi: 10.1126/science.2734611
55. Kantermann T, Juda M, Merrow M, Roenneberg T. The human circadian clock's seasonal adjustment is disrupted by daylight saving time. *Curr Biol*. (2007) 17:1996–2000. doi: 10.1016/j.cub.2007.10.025
56. Van Someren EJ. More than a marker: interaction between the circadian regulation of temperature and sleep, age-related changes, and treatment possibilities. *Chronobiol Int*. (2000) 17:313–54. doi: 10.1081/CBI-100101050
57. Yuda E, Ueda N, Kisohara M, Hayano J. Redundancy among risk predictors derived from heart rate variability and dynamics: ALLSTAR big data analysis. *Ann Noninvas Electrocardiol*. (2021) 26:e12790. doi: 10.1111/anec.12790

Conflict of Interest: LL was employed by the company Intersect communications 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 Medicine

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fmed.2021.661309

Deep Learning for Identification of Acute Illness and Facial Cues of Illness

Castela Forte[1,2,3†], Andrei Voinea[3], Malina Chichirau[3], Galiya Yeshmagambetova[3], Lea M. Albrecht[2], Chiara Erfurt[2], Liliane A. Freundt[2], Luisa Oliveira e Carmo[2], Robert H. Henning[1†], Iwan C. C. van der Horst[4†], Tina Sundelin[5,6†], Marco A. Wiering[3], John Axelsson[5,6†] and Anne H. Epema[2]*

[1] Department of Clinical Pharmacy and Pharmacology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands, [2] Department of Anesthesiology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands, [3] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, Netherlands, [4] Department of Intensive Care Medicine, Maastricht University Medical Centre+, University Maastricht, Maastricht, Netherlands, [5] Department of Psychology, Stress Research Institute, Stockholm University, Stockholm, Sweden, [6] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

OPEN ACCESS

Edited by:
Juan Liu,
Huazhong University of Science and Technology, China

Reviewed by:
Mohammad Shahid,
Children's National Hospital, United States
Juan Song,
Xidian University, China

***Correspondence:**
Castela Forte
j.n.alves.castela.cardoso
forte@umcg.nl

†ORCID:
Castela Forte
orcid.org/0000-0001-9273-0702
Robert H. Henning
orcid.org/0000-0002-5135-4621
Iwan C.C. van der Horst
orcid.org/0000-0003-3891-8522
Tina Sundelin
orcid.org/0000-0002-7590-0826
John Axelsson
orcid.org/0000-0003-3932-7310

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: 30 January 2021
Accepted: 30 June 2021
Published: 26 July 2021

Citation:
Forte C, Voinea A, Chichirau M, Yeshmagambetova G, Albrecht LM, Erfurt C, Freundt LA, Carmo LOe, Henning RH, Horst ICCvd, Sundelin T, Wiering MA, Axelsson J and Epema AH (2021) Deep Learning for Identification of Acute Illness and Facial Cues of Illness. Front. Med. 8:661309. doi: 10.3389/fmed.2021.661309

Background: The inclusion of facial and bodily cues (clinical gestalt) in machine learning (ML) models improves the assessment of patients' health status, as shown in genetic syndromes and acute coronary syndrome. It is unknown if the inclusion of clinical gestalt improves ML-based classification of acutely ill patients. As in previous research in ML analysis of medical images, simulated or augmented data may be used to assess the usability of clinical gestalt.

Objective: To assess whether a deep learning algorithm trained on a dataset of simulated and augmented facial photographs reflecting acutely ill patients can distinguish between healthy and LPS-infused, acutely ill individuals.

Methods: Photographs from twenty-six volunteers whose facial features were manipulated to resemble a state of acute illness were used to extract features of illness and generate a synthetic dataset of acutely ill photographs, using a neural transfer convolutional neural network (NT-CNN) for data augmentation. Then, four distinct CNNs were trained on different parts of the facial photographs and concatenated into one final, stacked CNN which classified individuals as healthy or acutely ill. Finally, the stacked CNN was validated in an external dataset of volunteers injected with lipopolysaccharide (LPS).

Results: In the external validation set, the four individual feature models distinguished acutely ill patients with sensitivities ranging from 10.5% (95% CI, 1.3–33.1% for the skin model) to 89.4% (66.9–98.7%, for the nose model). Specificity ranged from 42.1% (20.3–66.5%) for the nose model and 94.7% (73.9–99.9%) for skin. The stacked model combining all four facial features achieved an area under the receiver characteristic operating curve (AUROC) of 0.67 (0.62–0.71) and distinguished acutely ill patients with a sensitivity of 100% (82.35–100.00%) and specificity of 42.11% (20.25–66.50%).

Conclusion: A deep learning algorithm trained on a synthetic, augmented dataset of facial photographs distinguished between healthy and simulated acutely ill individuals,

demonstrating that synthetically generated data can be used to develop algorithms for health conditions in which large datasets are difficult to obtain. These results support the potential of facial feature analysis algorithms to support the diagnosis of acute illness.

Keywords: gestalt, deep learning, facial analysis, synthetic data, acute illness

INTRODUCTION

It is estimated that patients with sepsis alone account for as much as 6% of all hospital admissions and that while case-fatality rates are declining, the incidence of sepsis keeps increasing (1, 2). Early recognition of acute illness is critical for timely initiation of treatment (1). However, patients admitted to the emergency department (ED) or intensive care unit (ICU) with critical conditions such as sepsis often present with heterogeneous signs and symptoms, making detection and diagnosis challenging (3). Numerous risk scores based on laboratory variables and vital signs have been developed in an attempt to tackle this, but these achieved variable performance or were inferior to clinicians' informed judgment, also known as the clinical gestalt (4–7).

The clinical gestalt theory states that healthcare practitioners can actively organize clinical perceptions into coherent constructs or heuristics to reduce decision complexity, for example, by analyzing patients' facial and bodily cues, to estimate their functional status (8, 9). The value of the clinical gestalt as a diagnostic tool has been studied in different health conditions (10–13). In acute coronary syndrome, heart failure, pneumonia, and COVID-19, the clinical gestalt registered by doctors was comparable to clinical scores in "ruling in" or "ruling out" patients with certain symptoms presenting to the ED (10–14). For sepsis, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) advocates clinicians should, in addition to systemic inflammatory response syndrome (SIRS) criteria, use clinical gestalt in screening, treating and risk-stratifying patients with infection (15).

The clinical gestalt is also increasingly used as the basis for building deep learning models, with facial pictures being used to identify different genetic syndromes (16), as well as to detect coronary artery disease in an emergency setting (17). However, despite a growing number of studies reporting good results of deep learning models trained with a variety of clinical measurements to predict or detect early sepsis, no model has yet included clinical gestalt or facial feature analysis (18, 19). One major challenge to the development of a well-performing deep learning algorithm for facial analysis is the datasets' size and quality of the images (20, 21). With small datasets, deep neural networks will inevitably overfit, i.e., perfectly model the training data but lack generalizability and therefore perform poorly in a different validation dataset (21). However, there is substantial difficulty in obtaining a large gestalt dataset when privacy concerns associated with collecting facial photographic data exist, and especially in the emergency setting (22, 23). The use of simulated or synthetic data and augmenting existing data may solve this problem, as previously demonstrated for medical imaging and electronic medical record data (24–27). Moreover, there is vast literature, including recent studies, highlighting several key features of acute illness – including "a tired appearance," "pale skin and/or lips," "swollen face," and "hanging eyelids" – which can accurately be simulated (28–31).

Thus, to get insight into the usability of gestalt data in categorizing sick individuals, we used facial photographs of volunteers simulating these features to represent persons with and without acute illness. We trained a deep learning algorithm on facial photographs of simulated acute illness and a dataset of augmented facial photographs using a style transfer algorithm. Then, a concatenated model with multiple convolutional neural networks was validated on an external dataset of photographs of otherwise healthy volunteers injected with lipopolysaccharide (LPS).

METHODS

Dataset

An overview of the different steps of this study is provided in **Figure 1**. Three different data sources were used. The training dataset was created through combining two sets of photographs. First, a set of "simulated" sick faces, where the facial features of healthy volunteers had been manipulated using make-up, and second, a set of synthetically generated data resulting from the transfer of these features onto photographs from an open-source faces database (32). The validation dataset used data from a third set of photographs, which consisted of facial photographs from a previous study of individuals before and after they were administered LPS to experimentally induce acute illness (33).

Dataset With Simulated Sick Facial Features on Healthy Volunteers

Facial features characteristic of acute illness were simulated using make-up on 26 individuals (11 female). These characteristics of early acute illness included changes in skin color (pallor) due to vasoconstriction, drooping of mouth corners, and eye closure, often due to altered mental status (28–31). In total, seven facial features were simulated: paler skin tone, pale lips, redness around the eyes, sunken eyes, redness around the nasal alae, droopy mouth, and more opaque skin. The standard protocol followed for the make-up application is shown in **Supplementary Table 1** and **Supplementary Figures 1–3**. Two photographs of each participant were selected and included in the study, one without any make-up to represent the "healthy" control state, and another to represent the "acutely ill" state. A standardized environment with a gray background and LED light was used, and photographs were taken with an iPhone 8 camera (4,032 × 3,024 pixels) with standardized settings (ISO 22, RAW, AF, S1/40, MF: 0.9 and AWB in the Halide app). White balance of the complete set of photographs was standardized by a professional photographer using Adobe Photoshop (CC 2019).

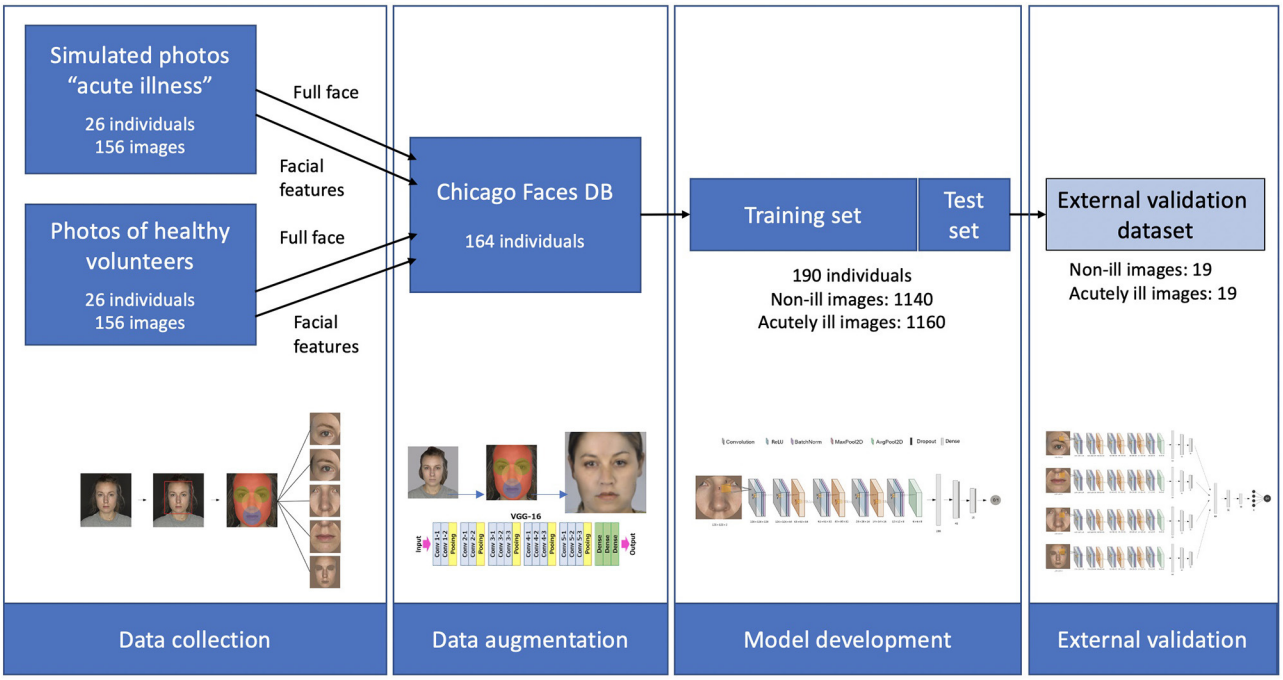

FIGURE 1 | Schematic overview of the three datasets, and of the different steps in the study. Starting with data collection from 26 healthy individuals and simulated acute illness. These features were then extracted and transferred to 164 faces from the Chicago Faces database, resulting in a total of 2,300 images of healthy and acutely ill individuals. After training the deep learning algorithm, validation was carried out on an external dataset of 19 individuals using a stacked CNN combining 4 individual networks.

Data Augmentation to Expand Training Dataset

To expand the dataset, one hundred sixty-four distinct faces from the Chicago Face Database (CFD) were retrieved and taken to represent "non-sick" individuals (32). In addition, photographs mimicking acute illness were generated using the same individual faces from the CFD and a neural algorithm of artistic style transfer. This algorithm transferred the make-up style representing acute illness to healthy individuals from the CFD. A VGG19 deep convolutional network was trained so that it got exposed to each image for 1,500 steps. Male and female participants were separated to ensure appropriate transfer of features and lower artifact creation. The one image per subject visually assessed by two researchers (JCF and AV) to represent the best acute illness was selected.

Validation Dataset of Individuals With LPS-Induced Illness

The external validation dataset consisted of the photographs of 22 individuals before (placebo, healthy) and 2 h after being injected with LPS. These individuals were mostly male (9 female) and of a similar age (mean 23.4). Camera resolution settings used were similar to those described before, and an equally standardized procedure was followed using a studio set-up. Additional details of these data are provided elsewhere (33).

Ethics

The study was exempt from ethical approval from the Medical Ethical Committee of the University Medical Centre Groningen. For the healthy volunteers, consent was obtained from all volunteers, including for the use of certain images for publication. A license for the use of the CFD was obtained by the study's authors (JCF and AV). Lastly, consent for collection and use of the photographs in the validation set was obtained previously, with the original study being approved by the regional ethical review board of Stockholm, Sweden (Registration number 2015/1415-32) and registered in ClinicalTrials.gov (NCT02529592) (33).

Data Pre-Processing

The simulated photographs and the validation photographs differed in certain aspects. In the simulated data, the features of acute illness were more accentuated than in the LPS group. In addition, the lighting was brighter in the validation data set, with somewhat dimmer light and more pronounced shadows and contrasts in the simulated dataset. To correct for this, all photographs in the simulated set were brightened (gamma = 1.3). All photographs were then resized to 128 × 128 pixels, and the four facial features (eyes, nose, mouth, and skin) were extracted separately using computer vision algorithms, as shown in **Figure 2**. A Haar cascade facial classifier was used to identify the entire face region in an image (34, 35). The

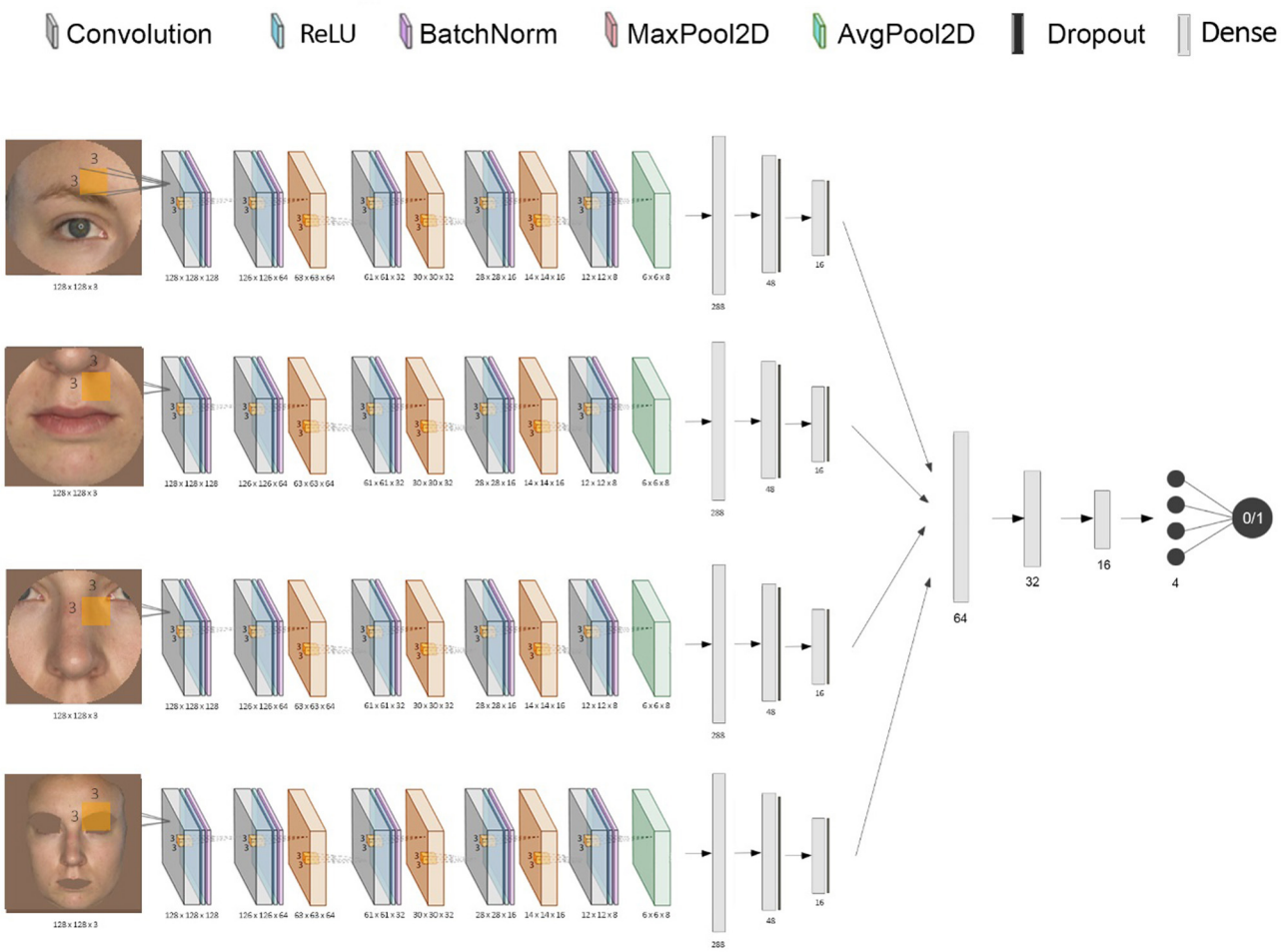

FIGURE 2 | Diagram of the stacked CNN. This shows the combination of each CNN's inputs and outputs into one final binary classification of "acutely ill" or "healthy".

facial landmark detector identified the face features, obtained by training a shape predictor on a labeled dataset (36, 37). The eyes, nose, and lips were extracted by calculating the minimum circle enclosing the 2D set of points representing each feature (given by the facial landmark detector). Finally, the skin area was extracted by removing the eyes and lips regions and everything outside the jaw region. Any other background and hair were removed by thresholding out certain color ranges (between HEX #000000 and #646464; #a0a0a0 and #aaaaaa were selected based on observation). The removed regions were replaced with the dominant color calculated from each face region, ensuring no other noise is passed down through the CNNs.

Deep Learning Algorithm

A CNN was trained for each facial feature using Keras with a Tensorflow backend. The individual networks input is represented by a 128 by 128 pixels RGB image, which is convolved with a convolution kernel of size (3, 3) after adding padding, using 128 filters. We use a rectified linear unit (ReLU) as an activation function, the output being normalized and scaled through a layer of batch normalization. The subsequent layers progressively down-sample the image data through groups of convolution layers (without padding), batch normalization, and max pooling layers with a pool size of (2, 2). Then, the final down-sampling layer uses an average pooling layer (with the same pooling size) to smooth the resulting filters. Finally, the output is flattened, resulting in a tensor of length 288. This is passed through two other fully-connected layers, each having a drop-out layer. The final layer is fully-connected with the output unit that uses a sigmoidal activation function, which generates an output value between 0 and 1 representing the probability of being classified as "ill."

To build the stacked ensemble combining all the previously mentioned CNNs, the final layer of all individual networks was removed, and each vector representation of size 16 was concatenated, resulting in a vector of size 64 (**Figure 2**). The data was then again gradually down-sampled through four fully-connected layers using ReLU (of size 32, 16, 4, and 1, respectively). The final activation function for the output is again the sigmoid function to ensure a value between 0 and 1. Both the CNNs and the stacked network use an Adam optimizer (adaptive moment estimation) with an initial learning rate of 0.001 and

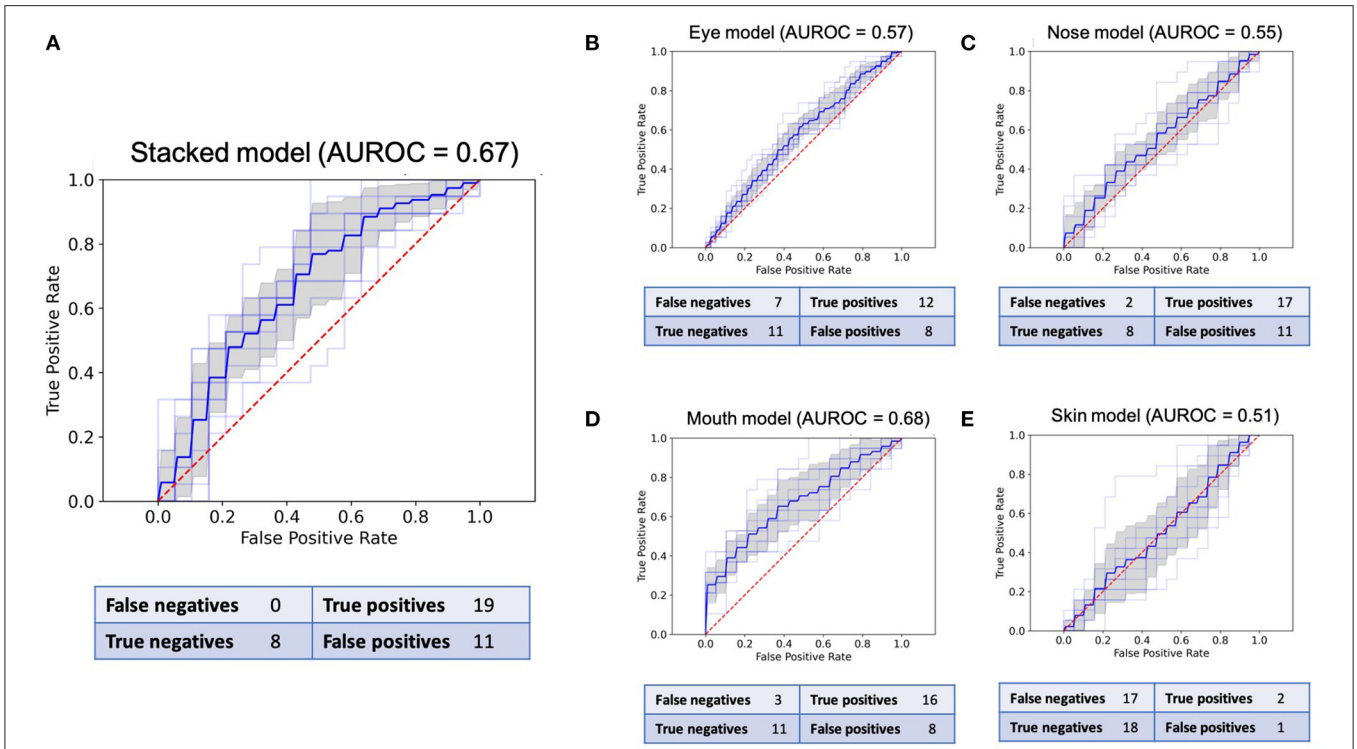

FIGURE 3 | Receiver operating characteristic (ROC) curves and confusion matrices for the final model of all five CNNs in the validation set of 38 images. **(A)**: ROC for the stacked model. **(B)**: ROC for the eyes. **(C)**: ROC for the nose. **(D)**: ROC for the mouth. **(E)**: ROC for the skin.

values for beta1 = 0.9, beta2 = 0.999, and epsilon = 10^{-8}. All models used a binary cross-entropy loss function. In order to minimize overfitting, early stopping and model checkpoints were used to save the model with the best testing F1 score during training.

Statistical Analysis

Each CNN was trained using 10-fold cross-validation. The best model with regard to testing accuracy across all folds was used to make predictions on the validation data. The different CNN models' performance is reported as the respective area under the receiver characteristic operating curve (AUROC), sensitivity, specificity, and negative and positive predictive values on the external validation data (38). Box-and-whisker plots were used to represent the median and interquartile ranges (25–75%) of all model AUROCs. All results are presented with a 95% confidence interval. Confusion matrices aggregating the predictions made by the final models are provided in **Figure 3**.

RESULTS

After data augmentation, the training dataset included photographs from 190 distinct individuals, adding up to a total of 1,140 healthy images and 1,160 images representing a state of acute illness for different facial regions, as well as for the complete face.

The sensitivity and specificity reported for each model pertain to the best models in the binary classification task and are based on the confusion matrices presented in **Figure 3**. The stacked CNN achieved an AUROC in the validation dataset of 0.67 (95% CI 0.61–0.72), with a sensitivity of 100% (82.4–100.0%) and specificity of 42.1% (20.3–66.5%). With regard to the four CNNs trained on individual features, the network with the best performance at distinguishing between healthy and ill individuals was the mouth CNN, with an AUROC of 0.68 (0.62–0.74) and sensitivity of 84.2% (60.4–96.6%) and specificity of 57.9% (33.5–79.8%). All other CNNs achieved AUROCs between 0.51 and 0.57, with sensitivities between 10.5% (1.3–33.1%) and 89.4% (66.9–98.7%), and specificities between 42.1% (20.3–66.5%) and 94.7% (73.9–99.9%). The positive predictive values (PPV) for individual models ranged between 60 and 66.7% for the nose and mouth models, respectively (**Table 1**). The negative predictive values (NPV) ranged between 51.4% for the skin model and 80% for the nose model. For the stacked model, PPV was 63% (54.1–71.7%) and the NPV was 100%.

The variation in performance of the individual and stacked models in the validation set across the different folds can be seen in **Figure 4**. Despite the marginally higher AUROC of the best mouth model compared to the stacked model, the stacked model was the most stable across all folds.

DISCUSSION

In this study, we developed a deep learning algorithm combining multiple convolutional neural networks to distinguish between healthy and acutely ill individuals based on facial feature analysis.

TABLE 1 | Performance of the best models for each feature and the stacked model on the validation set.

Trained on CFD augmented with simulated acute illness photographs					
Model	**AUROC**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**
Mouth	0.68 (0.62–0.74)	84.2 (60.4–96.6)	57.9 (33.5–79.8)	66.7 (53.3–77.8)	78.6 (54.8–91.7)
Nose	0.55 (0.50–0.60)	89.4 (66.9–98.7)	42.1 (20.3–66.5)	60.7 (50.6– 70.0)	80.0 (49.3–94.3)
Skin	0.51 (0.43–0.59)	10.5 (1.3–33.1)	94.7 (73.9–99.9)	66.7 (16.5–95.3)	51.4 (46.8–56.1)
Eye	0.57 (0.55–0.59)	63.2 (38.4–83.7)	57.9 (33.5–79.8)	60.0 (44.4–73.8)	61.1 (43.8–76.0)
Stacked	0.67 (0.61–0.72)	100 (82.4–100.0)	42.1 (20.3–66.5)	63.3 (54.1–71.7)	100.00

Values are presented as the area under the curve (AUROC), sensitivity, specificity, and positive and negative predictive values for 50% disease prevalence with 95% confidence intervals. PPV, positive predictive value, NPV, negative predictive value.

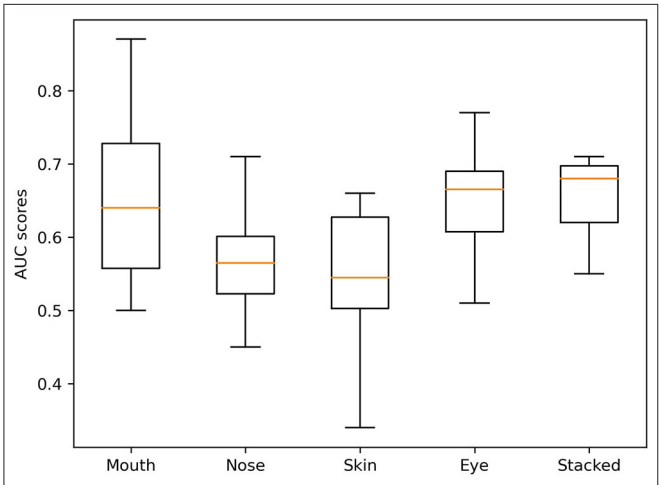

FIGURE 4 | Box-and-whiskers plot of AUC scores of the final models.

We showed that an algorithm trained on augmented facial data of simulated acute illness can successfully generalize predictions on an external dataset of individuals injected with LPS. The final, stacked model combining eyes, mouth, skin, and nose distinguished healthy and ill participants with a sensitivity of 100% (95% CI 82.4–100.0), specificity of 42.1% (20.3–66.5), and AUROC 0.67 (0.61–0.72).

The aim of this study was to investigate how a deep learning algorithm trained on augmented, facial data of simulated acute illness would perform in distinguishing between acutely ill and not ill individuals from an external set of photographs of real individuals with LPS-induced illness. While clinicians or other algorithms' baseline discriminatory ability for acute illness is not established, previous studies on the identification of acute illness based on facial features reported an AUROC of 0.62 (0.60–0.63), with sensitivity and specificity of 52 and 70%, respectively (33). These results were somewhat improved by the stacked model. However, both previous studies on the detection of different acute pathologies by trained physicians, as well as of clinical scores in sepsis detection, have found better results (7, 12, 13). For pneumonia and acute rhinosinusitis, the clinical gestalt achieved AUROCs of between 0.77 and 0.84 (12). Similarly, for acute heart failure, a specific combination of physical cues was converted into a score and achieved AUROCs above 0.90, diagnosing up to 88% of heart failure patients (13). Therefore, we can say this deep learning algorithm trained on simulated "gestalt" data distinguished between photographs of acutely ill and healthy people above chance level, surpassing the performance of non-experts, but fell below the performance of trained clinicians in other studies of different health conditions. This has several potential clinical implications. Firstly, it supports further research on the use of clinical gestalt for detection of acute illness in the ED and ICU, alone or possibly in combination with other clinical parameters. Combining "gestalt" and the modified SIRS score has already been shown to achieve good predictive performance for 24-h mortality in children (39). In adults, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) support the idea of combining the adult SIRS criteria and clinical gestalt to screen, triage, and treat patients with infection (15). And secondly, it suggests that adding "gestalt" to other machine learning algorithms for sepsis or septic shock detection may be of value, as these have traditionally focused on vital signs and electronic health record information (40, 41).

In addition, our study reached some technically interesting conclusions related to the feasibility of using synthetic data for deep learning. It is known that the generalizability of deep learning is lower, and the chance of over-fitting conversely higher, in small datasets. This is especially true for imaging data. Therefore, it was an interesting challenge to test whether synthetic data generation and data augmentation could be valid methodologies to address the problem of data availability for certain health conditions in a research setting, be it due to legal-ethical and privacy concerns or to low prevalence of disease (21, 22). We found scarce examples in literature of studies simulating a specific disease-state using techniques such as facial manipulation with moulage or make-up. One other study took photographs of volunteers before and after application of moulage designed to simulate traumatic facial injuries, and found that upon examination of these photographs by a facial analysis software, between 39 and 90% of photographs of injured patients were identified correctly (42). Clearly, synthetic and augmented datasets have the potential to enable researchers to "tailor" data to a specific context, but their generation and use is not without challenges. One immediate challenge is that a definitive measure for the quality of synthetic data is currently lacking (43). Here, we attempted to achieve as great a similarity as possible between

training and test data by using a widely validated methodology for feature detection and extraction, and then manually selecting the photographs to be included in the training set (36). Yet, we found that both the deep learning algorithms identified "healthy" individuals with higher accuracy. This was also the case for the non-expert raters in Axelsson et al. 's study, and could be due to an inherently greater degree of similarity between the facial features of healthy individuals than those of the acutely ill ones (33). However, we cannot rule out the possibility that it could also be a reflection of the features of acute illness in the validation dataset being less prominent than in the simulated training data. Because the risk of dissimilarity between training and testing data increases as the size of the dataset increases, and manual verification would not be possible for millions of images, the development of methodologies and standards to measure the quality of synthetic data is necessary before it can be used more widely.

Limitations of this study include the relatively small size of the training dataset, despite the data augmentation process, if compared to established clinical image databases for other diseases (44–46). This prevented us from further tuning the models' hyper-parameters on a holdout subset of the data and may have led to some overfitting. Second, there is a chance the data are inherently biased regarding the illness features and the ethnicity of participants. Despite the standardized, literature-based procedure for acute illness simulation in healthy volunteers, it is possible that individuals whose sick features are naturally more discrete were underrepresented. Equally, both the training and validation datasets included mostly Caucasian individuals, limiting the generalizability of the model to other ethnicities. Further tuning of the model on more ethnically diverse data and testing on a multi-ethnic dataset is warranted (47). Lastly, the potential for implementation of the algorithm can only truly be assessed in a dataset of real ICU or emergency department patients. While LPS produces physical symptoms similar to sepsis and is a well-acknowledged model to study sepsis in humans (48), real patient photographs collected in the ICU or emergency department would bring different challenges than photographs taken in a simulated setting. This could be due to noisy data from different lighting, wires, respirator tubes, and lower standardization of data.

In conclusion, a deep learning algorithm trained on synthetic data representing the clinical gestalt of acute illness was able to distinguish moderately well-between healthy and acutely ill individuals in an external dataset of individuals with LPS-induced acute illness. These results support the value of clinical gestalt as a diagnostic tool for acute illness. Additionally, synthetically generated data seem to be a valid alternative methodology to develop models for health conditions in which large datasets are difficult to obtain.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession numbers can be found below: https://github.com/J1C4F8/deep_learning_acute_illness.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Regional ethical review board of Stockholm, Sweden (Registration number 2015/1415-32). For generation of the training data, no ethical approval from the Medical Ethical Committee of the University Medical Centre Groningen was needed. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

CF: main contributor to all aspects of the manuscript. AV and MC: artificial intelligence student, significant contributor to the methods and results sections of the manuscript, and designer of figures of software architecture. GY: artificial intelligence student, significant contributor to the methods and results sections of the manuscript. LA, CE, LF, and LC: medical students responsible for data collection for the dataset of simulated features of illness. RH: significant contributor to the methods and discussion sections of the manuscript. IH: co-supervisor in the clinical aspects of the manuscript and original ideation. MW: supervisor of the model development and significant contributor to the methods and discussion aspects of the manuscript. TS and JA: significant contributors to the methods sections and responsible for creating and providing the validation data. AE: main supervisor in the clinical aspects of the manuscript, original ideation, and significant contributions to the introduction and discussion. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors wish to thank Mrs. Romée Stapel for assistance in make-up of volunteers and Mr. Marco Wiegers for creating high quality and standardized photographs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.661309/full#supplementary-material

REFERENCES

1. Husabø G, Nilsen RM, Flaatten H, Solligård E, Frich JC, Bondevik GT, et al. Early diagnosis of sepsis in emergency departments, time to treatment, and association with mortality: an observational study. *PLoS ONE.* (2020) 15:e0227652. doi: 10.1371/journal.pone.0227652
2. Lagu T, Rothberg MB, Shieh MS, Pekow PS, Steingrub JS, Lindenauer PK. Hospitalizations, costs, and outcomes of severe sepsis in the United States 2003 to 2007. *Crit Care Med.* (2012) 40:754–61 doi: 10.1097/CCM.0b013e318232db65
3. Morr M, Lukasz A, Rübig E, Pavenstädt H, Kümpers P. Sepsis recognition in the emergency department–impact on quality of care and outcome? *BMC Emerg Med.* (2017) 17:11. doi: 10.1186/s12873-017-0122-9
4. Singer M, Deutschman CD, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA.* (2016) 315:801–10. doi: 10.1001/jama.2016.0287
5. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5
6. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Int Care Med.* (2020) 46:383–400. doi: 10.1007/s00134-019-05872-y
7. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open.* (2018) 8:e017833. doi: 10.1136/bmjopen-2017-017833
8. Griffiths F, Svantesson M, Bassford C, Dale J, Blake C, McCreedy A, et al. Decision-making around admission to intensive care in the UK pre-COVID-19: a multicentre ethnographic study. *Anaesthesia.* (2020) 76:489–99. doi: 10.1111/anae.15272
9. Cook C. Is clinical gestalt good enough? *J Man Manip Ther.* (2009) 17:6–7. doi: 10.1179/106698109790818223
10. Oliver G, Reynard C, Morris N, Body R. Can emergency physician gestalt "Rule In" or "Rule Out" acute coronary syndrome: validation in a multicenter prospective diagnostic cohort study. *Acad Emerg Med.* (2020) 27:24–30. doi: 10.1111/acem.13836
11. Visser A, Wolthuis A, Breedveld R, ter Avest E. HEART score and clinical gestalt have similar diagnostic accuracy for diagnosing ACS in an unselected population of patients with chest pain presenting in the ED. *Emerg Med J.* (2015) 32:595–600. doi: 10.1136/emermed-2014-203798
12. Dale AP, Marchello C, Ebell MH. Clinical gestalt to diagnose pneumonia, sinusitis, and pharyngitis: a meta-analysis. *Br J Gen Pract.* (2019) 69:e444–53. doi: 10.3399/bjgp19X704297
13. Roncalli J, Picard F, Delarche N, Faure I, Pradeau C, Thicoipe M, et al. Predictive criteria for acute heart failure in emergency department patients with acute dyspnoea: the PREDICA study. *Eur J Emerg Med.* (2019) 26:400–4. doi: 10.1097/MEJ.0000000000000622
14. Soto-Mota A, Marfil-Garza BA, de Obeso SC, Martínez E, Carrillo-Vázquez DA, Tadeo-Espinoza H, et al. Prospective predictive performance comparison between Clinical Gestalt and validated COVID-19 mortality scores. *medRxiv*. (2021). doi: 10.1101/2021.04.16.21255647
15. Fernando SM, Rochwerg B, Seely AJE. Clinical implications of the third international consensus definitions for sepsis and septic shock (Sepsis-3). *CMAJ*. (2018) 190:E1058–9. doi: 10.1503/cmaj.170149
16. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med.* (2019) 25:60–4. doi: 10.1038/s41591-018-0279-0
17. Lin S, Li Z, Fu B, Chen S, Li X, Wang Y, et al. Feasibility of using deep learning to detect coronary artery disease based on facial photo. *Eur Heart J.* (2020) 00:1–12 doi: 10.1093/eurheartj/ehaa640
18. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Front Med.* (2021) 8:607952. doi: 10.3389/fmed.2021.607952
19. Giacobbe DR, Signori A, Del Puente F, Mora S, Carmisciano L, Briano F, et al. Early detection of sepsis with machine learning techniques: a brief clinical perspective. *Front Med.* (2021) 8:617486. doi: 10.3389/fmed.2021.617486
20. Shorten C., Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. (2019) 6:60. doi: 10.1186/s40537-019-0197-0
21. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst.* (2009) 24:8–12. doi: 10.1109/MIS.2009.36
22. de Sitter A, Visser M, Brouwer I, Cover KS, van Schijndel RA, Eijgelaar RS, et al. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *Eur Radiol.* (2020) 30:1062–74. doi: 10.1007/s00330-019-06459-3
23. Martinez-Martin N. What are important ethical implications of using facial recognition technology in health care? *AMA J Ethics.* (2019) 21:E180–7. doi: 10.1001/amajethics.2019.180
24. Iqbal T, Ali H. Generative adversarial network for medical images (MI-GAN). *J Med Syst.* (2018) 42:231. doi: 10.1007/s10916-018-1072-9
25. Kazuhiro K, Werner RA, Toriumi F, Javadi MS, Pomper MG, Solnes LB, et al. Generative adversarial networks for the creation of realistic artificial brain magnetic. *Tomography.* (2018) 4:159–63. doi: 10.18383/j.tom.2018.00042
26. Buczak A, Babin S, Moniz L. Data-driven approach for creating synthetic electronic medical records. *BMC Med Inform Decis Mak.* (2010) 10:59. doi: 10.1186/1472-6947-10-59
27. Choi E, Siddharth B, Bradley M, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. *PMLR.* (2017) 68:286–305.
28. Henderson AJ, Lasselin J, Lekander M, Olsson MS, Powis SJ, Axelsson J, et al. Skin colour changes during experimentally-induced sickness. *Brain Behav Immun.* (2017) 60:312–8. doi: 10.1016/j.bbi.2016.11.008
29. Harris RL, Musher DM, Bloom K, Gathe J, Rice L, Sugarman B, et al. Manifestations of sepsis. *Arch Intern Med.* (1987) 147:1895–906. doi: 10.1001/archinte.1987.00370110023003
30. Heffernan AJ, Denny KJ. Host diagnostic biomarkers of infection in the ICU: where are we and where are we going? *Curr Infect Dis Rep.* (2021) 23:4. doi: 10.1007/s11908-021-00747-0
31. Filbin MR, Lynch J, Gillingham TD, Thorsen JE, Pasakarnis CL, Nepal S, et al. Presenting symptoms independently predict mortality in septic shock: importance of a previously unmeasured confounder. *Crit Care Med.* (2018) 46:1592–9. doi: 10.1097/CCM.0000000000003260
32. Ma DA, Correll J, Wittenbrink B. The chicago face database: a free stimulus set of faces and norming data. *Behav Res.* (2015) 47:1122–35. doi: 10.3758/s13428-014-0532-5
33. Axelsson J, Sundelin T, Olsson MJ, Sorjonen K, Axelsson C, Lasselin J, et al. Identification of acutely sick people and facial cues of sickness. *Proc R Soc B.* (2018) 285:20172430. doi: 10.1098/rspb.2017.2430
34. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Kauai, HI (2001).
35. Bradski G, Kaehler A. *Learning OpenCV, Computer Vision with the OpenCV Library*. Sebastopol, CA: O'Reilly (2008).
36. King DE. Dlibml: a machine learning toolkit. *J Mach Learn Res.* (2009) 10:1755–8. doi: 10.1145/1577069.1755843
37. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 faces in-the-wild challenge: database and results. *Image Vision Comp.* (2016) 47:3–18. doi: 10.1016/j.imavis.2016.01.002
38. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Int Res.* (2016) 18:e323. doi: 10.2196/jmir.5870
39. Nariadhara MR, Sawe HR, Runyon MS, Mwafongo V, Murray BL. Modified systemic inflammatory response syndrome and provider gestalt predicting adverse outcomes in children under 5 years presenting to an urban emergency department of a tertiary hospital in Tanzania. *Trop Med Health.* (2019) 47:13. doi: 10.1186/s41182-019-0136-y
40. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif Intell Med.* (2020) 104:101820. doi: 10.1016/j.artmed.2020.101820
41. Fagerström J, Bång M, Wilhelms D, Chew MS. LiSep LSTM: a machine learning algorithm for early detection of septic shock. *Sci Rep*. (2019) 9:15132. doi: 10.1038/s41598-019-51219-4

42. Broach J, Yong R, Manuell M-E, Nichols C. Use of facial recognition software to identify disaster victims with facial injuries. *Disaster Med Public Health Prep*. (2017) 11:568–72. doi: 10.1017/dmp.2016.207
43. Jordon J, Yoon J, van der Schaar M. Measuring the quality of synthetic data for use in competitions. *arXiv preprint* arXiv:1806.11345 (2018).
44. Meng T, Guo X, Lian W, Deng K, Gao L, Wang Z, et al. Identifying facial features and predicting patients of acromegaly using three-dimensional imaging techniques and machine learning. *Front Endocrinol.* (2020) 11:492. doi: 10.3389/fendo.2020.00492
45. Thomsen K, Christensen AL, Iversen L, Lomholt HB, Winther O. Deep learning for diagnostic binary classification of multiple-lesion skin diseases. *Front Med.* (2020) 7:574329. doi: 10.3389/fmed.2020.574329
46. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med.* (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3
47. Nagpal S, SinghM, Singh R, VatsaM. Deep learning for face recognition: pride or prejudiced? *arXiv preprint* arXiv:1904.01219 (2019).
48. Fiuza C, Suffredini AF. Human models of innate immunity: local and systemic inflammatory responses. *J Endotoxin Res.* (2001) 7:385–8. doi: 10.1177/09680519010070050701

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 Medicine

BRIEF RESEARCH REPORT
published: 05 August 2021
doi: 10.3389/fmed.2021.655084

Unsupervised Phonocardiogram Analysis With Distribution Density Based Variational Auto-Encoders

Shengchen Li[1*†] *and Ke Tian*[2†]

[1] Department of Interlligent Science, Xi'an Jiaotong-Liverpool University, Suzhou, China, [2] College of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

This paper proposes an unsupervised way for Phonocardiogram (PCG) analysis, which uses a revised auto encoder based on distribution density estimation in the latent space. Auto encoders especially Variational Auto-Encoders (VAEs) and its variant $\beta-$VAE are considered as one of the state-of-the-art methodologies for PCG analysis. VAE based models for PCG analysis assume that normal PCG signals can be represented by latent vectors that obey a normal Gaussian Model, which may not be necessary true in PCG analysis. This paper proposes two methods DBVAE and DBAE that are based on estimating the density of latent vectors in latent space to improve the performance of VAE based PCG analysis systems. Examining the system performance with PCG data from the a single domain and multiple domains, the proposed systems outperform the VAE based methods. The representation of normal PCG signals in the latent space is also investigated by calculating the kurtosis and skewness where DBAE introduces normal PCG representation following Gaussian-like models but DBVAE does not introduce normal PCG representation following Gaussian-like models.

Keywords: phonocardiogram analysis, auto-encoder, data density, unsupervised learning, abnormality detection

OPEN ACCESS

Edited by:
Liang Zhang,
Xidian University, China

Reviewed by:
Guangming Zhu,
Xidian University, China
Juan Song,
Xidian University, China

****Correspondence:***
Shengchen Li
shengchen.li@xjtlu.edu.cn

†These authors share co-first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *18 January 2021*
Accepted: *15 June 2021*
Published: *05 August 2021*

Citation:
Li S and Tian K (2021) Unsupervised Phonocardiogram Analysis With Distribution Density Based Variational Auto-Encoders. Front. Med. 8:655084. doi: 10.3389/fmed.2021.655084

1. INTRODUCTION

Phonocardiogram (PCG) analysis is a popular way for portable heart surveillance, which makes use of the heart sound to identify possible anomaly of heart statues. Existing PCG analysis methods use supervised methods which demands a labor expensive process of labeling. The paper proposes an unsupervised way of PCG analysis, which identifies abnormal PCG signals based on PCG analysis with normal signals only.

The main task of the proposed system is to characterize normal PCG signals in an unsupervised way and then identify abnormal PCG signals as outliers despite the existence of background noise and sound from other resources. In recent year, many attempts have been made to analyse PCG signals including the PhysioNet and CinC (Computing in Cardiology Challenge) data Challenge (1), which contains multiple sets of PCG data where both normal and abnormal PCG signals are presented and labeled.

With labels of normal and anomaly PCG signals, the PCG analysis can be considered as a classification problem. Classical machine learning techniques such as Support Vector Machine (SVM) (2), i-vector based dictionary learning method (3) and solutions based on Markov models (4) are used to solve the proposed problem besides deep learning algorithms (5, 6). However, as a supervised problem, PCG data collected needs to cover all types of PCG abnormality, which is labor expensive.

Inspired by the Anomalous sound detection (ASD) of Detection and Classification of Acoustic Scenes and Events (DCASE) data challenge 2020 (7, 8), the PCG analysis could also be considered as an unsupervised problem where only normal PCG signals are analyzed for the identification of anomaly PCG signals, which is considered as an outlier detection problem. This solution avoids PCG data collection problem as there is not need to collect all types of anomaly PCG signals for training.

The outlier detection of high-dimensional data is not a new research problem. Aggarwal and Yu (9) proposed to use sparse representation to find outliers. Pang et al. (10) using homophily couplings to identify outlier with noise. With the development of deep learning, Variational Automatic Encoder (VAE) (11) and a variant of VAE: $\beta-$VAE (12) are used for outlier detection in the PCG analysis, where the anomaly score of a PCG signal could be calculated by the features exacted from latent space of the VAE (13) or the reconstruction loss of $\beta-$VAE (14).

The PCG analysis based on VAE systems is based on an assumption that normal PCG signals can be represented by via latent vectors that obey a normal Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$. However, as normal PCG signals could be different from each other, the representation in latent space obeying a normal Gaussian distribution may not be the best feature representing PCG signals. For example, if the PCG collected from different sources, the PCG features could follow a Gaussian Mixture Model (GMM) due to different background noise and recording devices. In extreme cases, the resulting VAE may serve as a denoise VAE that converts anomaly PCG signals to normal PCG signals. As a result, this paper proposes two different ways to model normal PCG signals in a latent space.

The novelty of this paper is the use of sample density in latent space during the training process, which removes the assumption that normal PCG signals can be represented by latent vectors obeying a normal Gaussian distribution. At the same time, the KL divergence between latent vector distribution and normal Gaussian distribution is removed from the loss function, which potentially removes the assumption that the latent vectors must follow a normal Gaussian distribution.

Besides, the paper compares the system with and without the introduction of sampling process in the latent space during the training process. The proposed system with the sampling process in latent space follows the procedure that a VAE system is trained hence is named as Density based $\beta-$VAE system (DBVAE). The proposed system without the sampling process in the latent space likes a more traditional auto-encoder hence is named as Density based $\beta-$Auto-Encoder (DBAE) system. Both systems are compared with a $\beta-$VAE system, which is a more classical way for outlier identification.

The proposed method is tested with the Physio/CinC Heart Sound Dataset (1). There are six subsets of data collected, where each subset is collected in roughly the same way but from different places. This paper proposes two experiments to examine the performance of the proposed system. Firstly, the training data used is from the same subset. The resulting systems are evaluated by data from both the same subset and other subsets. Then data from different subsets are combined as the data used for training. The performance of the proposed systems are tested by the Receiver Operator Characteristic (ROC) test with Area Under Curve (AUC) values, which avoids the introduction of thresholds.

Theoretically speaking, the normal PCG representation in the latent space should follow a Gaussian-like model as there is a sampling process from Gaussian model during training. For the proposed DBAE, the resulting normal PCG representation in the latent space may not follow a Gaussian-like model due to the removal of sampling process from a Gaussian model. To examining the resulting normal PCG representation in the latent space, the kurtosis and skewness of the latent vectors are measured.

The paper is organized in the following way. Firstly, the proposed system is introduced. Then we present the results of the proposed experiments followed by the discussion to conclude this paper.

2. METHODS

The proposed system is formed by three stages: pre-processing of the PCG signal, the training of the revised VAE system and the post-processing stage to produce the anomaly score, which is then evaluated by a Receiver Operator Characteristic (ROC) test for Area Under Curve (AUC) values.

2.1. Pre-processing

The Physio/CinC Heart Sound Dataset contains the audio of heart sound ranges from 5 to 120 s, effectively contains 6–13 cardiac cycles. For easier processing during the training process, a standardized 6-s length is used for all samples where longer samples are truncated and shorter samples are padded in a recurrent way.

As a common way to extract features, the Mel Spectrogram is calculated with the following configuration engaged: a window length of 1,024 with a hope length of 512. There are 14 Mel filters are used. As the sampling rate of the heart sound audio is 2 kHz, each frame engaged in the Mel Spectrogram lasts 0.51 s.

For data bias removal, the resulting coefficients in the Mel Spectrogram is standardized according to each row. Given a Mel Spectrogram $\mathbf{S}_{M\times N} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M]^T$, the standardized row $\hat{S}_i$ in a Mel Spectrogram can be written as

$$\hat{S}_i = \frac{S_i - mean(S_i)}{std(S_i)}. \tag{1}$$

The standardized Mel Spectrogram can be written as $\hat{S} = [\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_M]^T$.

Each five frames of the standardized Mel Spectrogram then forms a super-frame, which is considered as a data sample in the training dataset. The starting frame of each super frame is selected in a rolling manner i.e., there are $L - 4$ super-frames for a piece of audio with L frames. Each super-frame lasts about 3 s, which should contain at least one complete cardiac cycle.

2.2. Proposed Systems

The motivation of the proposed system is to relax the assumption that the use of VAE introduced in PCG analysis: there is a way to represent normal PCG signals whose representation in the latent space obeys a standardized Gaussian distribution. The assumption may cause two types of problems: (1) As VAE is commonly used as a de-noise system, the resulting VAE system could serve as a de-noise system for PCG signals which converts anomaly PCG signals to normal ones; (2) If the PCG signals are collected from multiple sources, the latent representation of PCG signals is unlikely to follow a single Gaussian model but a Gaussian Mixture model. As a result, there are two models proposed in this paper to solve the potential problems.

The first model is named "Density $\beta-$VAE" (DBVAE) that attempts to avoid the resulting latent representation of the normal PCG signals follows a normal Gaussian distribution if unnecessary. The DBVAE adopts a VAE system whose loss function is formed by the combination of reconstruction loss and the density of samples in the latent space. Adopted from the VAE framework, there is a re-sampling procedure from a Gaussian model in the latent space, which makes the representation of PCG signals in the latent space may potentially follow a Gaussian distribution. As a result, the DBVAE expects the PCG signals can be represented by latent vectors following a single-component Gaussian model.

If the training data collected is from multiple sources, latent representations for normal PCG signals resulted from DBVAE may not necessarily follow a single-component Gaussian model hence the "Density $\beta-$Auto Encoder" (DBAE) is introduced. By removing the re-sampling process in the latent space, the representation of normal PCG signals in latent space no longer follows a Gaussian distribution compulsory. The DBAE uses the same loss function with DBVAE, which pursues a high density distribution in the latent space. With the proposed loss function, DBAE could avoid overfitting in the latent space, which overcomes the problem of auto-encoders may have. **Figure 1** gives a more intuitive explanation of the two methods.

The novel point of the proposed systems is to introduce a sample density based loss function term in the latent space. We now describe how sample density is estimated in the proposed systems.

In this paper, the sample density in latent space is defined as the average distance between each individual sample and the centroid point of the dataset. The centroid point of the dataset $\mathbf{C} = (c_1, c_2, \ldots, c_M)$ is formed by the centroid point of each dimension, where

$$c_i = \frac{max(z_{i1}, z_{i2}, \ldots, z_{iN}) + min(z_{i1}, z_{i2}, \ldots, z_{iN})}{2}. \quad (2)$$

The representation of all samples in the latent space is represented by $\mathbf{Z}_{M\times N}$ whose ith dimension for the jth sample is represented as z_{ij}. Using $\mathbf{Z}_j$ to represent the latent vector for sample j, The density measurement for all samples is then proposed as

$$\mathcal{D} = \frac{1}{N}\sum_{j=1}^{N} ||\mathbf{Z}_j - \mathbf{C}||^2. \quad (3)$$

Given $\mathcal{L}_r$ to represent the reconstruction loss measure by Mean Squared Error (MSE), the overall loss functions for both DBVAE and DBAE are

$$\mathcal{L} = \mathcal{L}_r + \beta\mathcal{D}. \quad (4)$$

2.3. Post-processing

The anomaly score for a PCG signal is based on the reconstruction error of the proposed systems. For each super-frame (five consecutive frames) in Mel Spectrogram, the MSE between original Mel Spectrogram and the recovered Mel Spectrogram is considered as the anomaly score (a_i) for this particular super-frame. The overall anomaly score (a) for a PCG signal with N frames is

$$a = \frac{1}{N-4}\sum_{i=1}^{N-4} a_i. \quad (5)$$

3. RESULTS

We firstly test the performance of the proposed systems with each single subset. Then we test the performance of the proposed system when how the subsets are combined. The baseline system selected is a β-VAE based system (14), which follows the extract experiment design in this paper.

There are six subset of data in the Physio/CinC dataset labeled as "a," "b," "c," "d," "e," "f." Given the fact that there are only a few samples in the subset "c," the results for subset "c" is omitted when only a single subset is used as the data source for training. Besides the single subset tests, this paper also presents the experiments that use the combination of multiple subsets as the training data source. Specifically, the subsets with most data are tested (e.g., 'a' & "e," "e," and "f") and the case of all subsets used is also tests (subset "c" inclusive). In all cases, 90% normal PCG data is used for training and the remaining 10% normal PCG data and all anomaly PCG data are used for testing. In addition, in order to make the experiment more credible, this paper introduces an additional data set called Michigan (15). The experimental results are labeled "Michigan" with the same training proportion.

As discussed by Higgins et al. (12), in general $\beta > 1$ is necessary to achieve good disentanglement. However, as reported by Li et al. (14), a smaller β value may help the performance of PCG analysis. As a result, this paper sets the β values to wider range: 0.01, 0.1, 1, 10, and 100 to test how the value of β effects the performance of proposed systems.

As a summary, **Table 1** shows the best and worst performed model for each type of candidate model with different settings of β values.

From **Table 2**, the proposed DBAE and DBVAE systems generally outperform the BVAE system if the value of β is

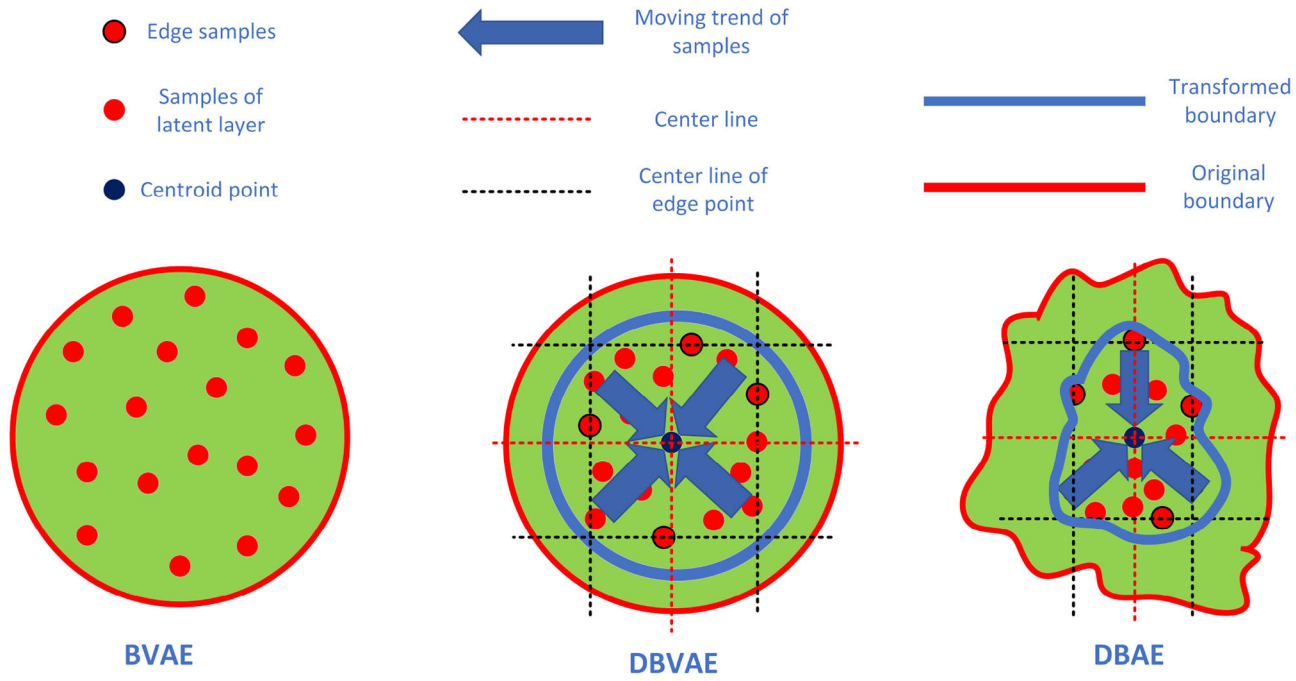

FIGURE 1 | In the BVAE method, the latent space only ensures that the sample points roughly obey the standard normal distribution, but there is no specific requirement for the density of the sample points. In the DBVAE method, the latent space not only retains the characteristic that the sample points obey the standard normal, but also makes the samples more aggregated by increasing the sample density. The whole process is shown in the figure, taking two-dimensional plane as an example. Firstly, the edge sample is determined, and the center line of the edge point is made according to the center of the edge sample on each dimension. Secondly, the average of the center lines of the edge points is calculated to get the center line, and the intersection of the center lines of different dimensions is defined as the centroid point. Finally, all the sample points shrink to the centroid. In the method of DBAE, the distribution assumption of sample points is canceled and only the sample density is required. The samples shown in the figure are normal sample points during the training process.

TABLE 1 | Best and worst performed system for all tests in terms of AUC values (at the first line of each row).

Best	**a**	**b**	**d**	**e**	**f**	**ae**	**ef**	**ALL**	**Michigan**
BVAE (AUC)	0.825	0.559	0.691	0.923	0.846	0.822	0.899	0.786	0.966
when β:	*0.01*	*1.00*	*100*	*0.01*	*100*	*0.01*	*0.01*	*0.01*	*1.00*
DBVAE (AUC)	0.862	0.642	0.845	0.924	0.831	0.861	0.914	0.803	0.966
when β:	*0.1*	*0.01*	*0.1*	*0.01*	*10*	*0.01*	*0.01*	*0.01*	*1.00*
DBAE (AUC)	0.842	0.614	0.940	0.928	0.842	0.887	0.929	0.808	0.944
when β:	*0.1*	*1.00*	*0.1*	*0.1*	*10*	*0.1*	*100*	*1.00*	*0.01*
Worst	**a**	**b**	**d**	**e**	**f**	**ae**	**ef**	**ALL**	**Michigan**
BVAE (AUC)	0.798	0.551	0.583	0.881	0.801	0.765	0.836	0.644	0.725
when β:	*1.00*	*0.1*	*0.1*	*1.00*	*10*	*100*	*10*	*100*	*0.1*
DBVAE (AUC)	0.763	0.523	0.726	0.844	0.787	0.793	0.870	0.719	0.731
when β:	*100*	*0.1*	*100*	*100*	*1.00*	*100*	*100*	*10*	*0.1*
DBAE (AUC)	0.762	0.527	0.75	0.918	0.765	0.851	0.895	0.761	0.616
when β:	*10*	*100*	*0.01*	*1.00*	*100*	*10*	*10*	*100*	*1.00*

The configuration of β is set as the second line of each row. The subsets used for training is labeled as the title of each column.

properly set. Specifically, when a single subset serves as the data source for training, the DBVAE has a comparable performance with DBAE in general whereas when multiple subsets are used as the data source for training, the DBAE in general outperforms the DBVAE and DBVAE is better than BVAE baseline.

Moreover, in the experiment presented, the results reveal that the effects of β differ from the candidate systems. Assuming the best performed β configuration is β_b and the worst performed β configuration is β_w, **Table 2** shows the value of $\delta = \frac{\beta_b}{\beta_w} - 1$ for all experiments presented, which effectively measures how much performance be can gained by adjusting the value of β in extreme cases.

From results of δ, the effects of β value selection can be summarized as the following: (1) using multiple subsets generally reduce the effects on β value; (2) BVAE systems are more stable than DBVAE and DBAE when data from single subset is used; (3) DBAE improves the stability of system performance when multiple subsets are used for training.

4. DISCUSSION

The proposed systems pursues different regulations on the distribution of latent vectors. To show how the PCG signals is presented in the latent space, the kurtosis and skewness are measured for the distribution of normal PCG signals. The definition of kurtosis and skewness is represented as follows.

TABLE 2 | The ratio δ between the models with best β settings and the worst β setting in all experiments.

	a	b	d	e	f	ae	ef	ALL	Michigan
BVAE	0.034	0.018	0.185	0.048	0.051	0.074	0.076	0.220	0.332
DBVAE	0.131	0.217	0.164	0.095	0.056	0.085	0.050	0.118	0.321
DBAE	0.105	0.165	0.254	0.011	0.101	0.042	0.037	0.062	0.532

A smaller number indicates less effects of β on system performance.

TABLE 3 | The average skewness and kurtosis for all resulting models in all experiments.

	Skewness (γ_1)	Kurtosis (γ_2)
$\mathcal{N}(0, 1)$	0	0
BVAE	0.115 (±0.093)	1.368 (±0.884)
DBVAE	2.674 (±8.128)	610.499 (±1669.592)
DBAE	−0.088 (±0.317)	3.369 (±4.413)

The sign "±" represents then standard deviation of the data. $\mathcal{N}(0, 1)$ represents normal Gaussian distribution.

Given a representation of PCG signal in the latent space $[\mathbf{Z}_j = (z_{1j}, z_{2j}, \ldots, z_{Mj})]$ and the mean value of all latent vectors ($\overline{\mathbf{Z}}$), the skewness (γ_1) and kurtosis (γ_2) of N samples in the latent space can be calculated as:

$$\gamma_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_j - \overline{\mathbf{Z}})^3}{(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_j - \overline{\mathbf{Z}})^2)^{3/2}} \tag{6}$$

$$\gamma_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_j - \overline{\mathbf{Z}})^4}{(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_j - \overline{\mathbf{Z}})^2)^2} - 3. \tag{7}$$

Table 3 shows the average value and standard deviation of the skewness and kurtosis of the distribution in the latent space. For a normal Gaussian distribution, the skewness and kurtosis is expected to be 0. A larger kurtosis value indicates the distribution of latent vectors is more dense. A skewness value with higher absolute value is considered as more different with a normal Gaussian distribution.

It is not surprising to find that BVAE systems produce a latent vector distribution that is similar with the normal Gaussian distribution. For DBAE, the resulting latent vectors in the latent space also follow a unbiased distribution with gentle variations on kurtosis in most cases, which suggests the resulting latent vectors follow a Gaussian-like model. Given the fact that for training data from multiple subsets should follow and mixture of models, it is interesting to find that the latent vectors as PCG normal signal representation follow a Gaussian-like model rather than a mixture of models. Moreover, it is surprising to find that the DBVAE results to heavily biased and high dense distribution despite a sampling process from Gaussian distribution, which suggests the resulting latent representation for DBVAE model is not following a Gaussian-like model. As a result, the normal PCG representation in the latent space needs further investigation in the future.

The motivation of proposing the DBVAE is to relax the assumption of the latent representation for normal PCG signals should follow a normal Gaussian distribution. The motivation of proposing DBAE is to relax the assumption of the latent representation for normal PCG signals should follow a Gaussian-like distribution. Both proposed system are expected to introduce an improvement of the system performance compared with VAE systems. Moreover, the DBAE is expected to outperform DBVAE when multiple subsets are used for training.

The final results confirm that both DBVAE and DBAE introduce an improvement on performance. DBAE introduces a small improvement compared with DBVAE when single subset is used as the source of training data. When multiple subsets are used for training, DBAE introduces a larger improvement compared with DBVAE. However, the investigation on the kurtosis and skewness of the distribution of PCG normal representation in latent space does not confirm the assumption this paper made where the DBVAE introduces a normal PCG representation in the latent space does not follow a Gaussian-like model but the DBAE introduces a normal PCG representation in the latent space that follows the a Gaussian-like model which are not expected.

As a quick conclusion, the introduction of density based auto-encoder systems, DBAE and DBVAE, improves the performance of PCG analysis however the latent representation of the proposed systems for normal PCG signals need investigation in the future for further improvements. The introduction of multiple subsets stabilizes the performance of the systems especially for DBAE, which reduces the efforts of tuning the value of β in the proposed systems.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the datasets analyzed for this study can be found in the PhysioNet at https://physionet.org/content/challenge-2016/1.0.0/.

AUTHOR CONTRIBUTIONS

SL composes the manuscript and designs the experiment proposed in the manuscript. KT implements the experiment for results with essential experiment design. All authors consider this piece of work as a full scale of collaboration.

FUNDING

SL was funded by National Neural Science Foundation of China (NSFC) for project Acoustic Scenes Classification based on Domain Adoptation Methods (62001038). KT was supported by the Fundamental Research Funds for the Central Universities (2019XD-A05).

REFERENCES

1. Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. *Physiol Meas*. (2016) 37:2181–213. doi: 10.1088/0967-3334/37/12/2181
2. Zabihi M, Rad AB, Kiranyaz S, Gabbouj M, Katsaggelos AK. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: *2016 Computing in Cardiology Conference (CinC)*. Vancouver, BC (2016). doi: 10.22489/CinC.2016.180-213
3. Adiban M, BabaAli B, Shehnepoor S. Statistical feature embedding for heart sound classification. *J Electric Eng*. (2019) 70:259–72. doi: 10.2478/jee-2019-0056
4. Grzegorczyk I, Solinski M, Lepek M, Perka A, Rosinski J, Rymko J, et al. PCG classification using a neural network approach. In: *2016 Computing in Cardiology Conference (CinC)*. Vancouver, BC (2016). p. 1129–32. doi: 10.22489/CinC.2016.323-252
5. Koike T, Qian K, Kong Q, Plumbley MD, Schuller BW, Yamamoto Y. Audio for audio is better? An investigation on transfer learning models for heart sound classification. In: *The 42nd International Engineering in Medicine and Biology Conference*. Montréal, QC (2020). p. 74–7. doi: 10.1109/EMBC44109.2020.9175450
6. Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Recognizing abnormal heart sounds using deep learning. *arXiv [Preprint]. arXiv:1707.04642*. (2017).
7. Koizumi Y, Saito S, Uematsu H, Harada N, Imoto K. ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY: IEEE (2019). doi: 10.1109/WASPAA.2019.8937164
8. Purohit H, Tanabe R, Ichige T, Endo T, Nikaido Y, Suefusa K, et al. MIMII dataset: sound dataset for malfunctioning industrial machine investigation and inspection. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events*. New York, NY: New York University (2019). doi: 10.33682/m76f-d618
9. Aggarwal CC, Yu PS. Outlier detection for high dimensional data. In: *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. Santa Barbara, CA (2001). p. 37–46. doi: 10.1145/376284.375668
10. Pang G, Cao L, Chen L, Liu H. Learning homophily couplings from Non-IID data for joint feature selection and noise-resilient outlier detection. In: *Proceeding of International Joint Conferences on Artificial Intelligence*. Melbourne (2017). p. 2585–91. doi: 10.24963/ijcai.2017/360
11. Kingma DP, Welling M. An introduction to variational autoencoders. *Found Trends Mach Learn*. (2019) 12:307–92. doi: 10.1561/2200000056
12. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In: *5th International Conference on Learning Representations, ICLR*. Toulon (2017).
13. Banerjee R, Ghose A. A semi-supervised approach for identifying abnormal heart sounds using variational autoencoder. In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (2020). p. 1249–53. doi: 10.1109/ICASSP40776.2020.9054632
14. Li S, Tian K, Wang R. Unsupervised heart abnormality detection based on phonocardiogram analysis with beta variational auto-encoders. *arXiv [Preprint]. arXiv:2101.05443*. (2021) doi: 10.1109/ICASSP39728.2021.9414165
15. Richard D, Judge MD, FACC. (2015). Available online at: https://open.umich.edu/find/open-educational-resources/medical/heart-sound-murmur-library (accessed July 14, 2021).

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