2023 5th International Academic Exchange Conference on Science and Technology Innovation (IAECST) | 979-8-3503-5773-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IAECST60924.2023.10502998

Deep Learning-based Decision-tree Classifier for Tuberculosis Diagnosis

Zhixiang Lu
Wisdom Lake Academy of Pharmacy
Xi'an Jiaotong-Liverpool University
Suzhou, China
Zhixiang.Lu22@student.xjtlu.edu.cn

Tenglong Li*
Wisdom Lake Academy of Pharmacy
Xi'an Jiaotong-Liverpool University
Suzhou, China
* Corresponding author: Tenglong.Li@xjtlu.edu.cn

Mingming Chen
Wisdom Lake Academy of Pharmacy
Xi'an Jiaotong-Liverpool University
Suzhou, China
Mingming.Chen22@student.xjtlu.edu.cn

***Abstract*-In recent years, medical disease-assisted diagnosis has been increasingly used. Prior to the COVID-19 epidemic, tuberculosis was the leading cause of death in the single infectious disease that dominated the global epidemic, and approximately 40% of tuberculosis patients were undiagnosed. Thus, making the development of a low-cost, non-invasive digital screening tool important for improving diagnosis in this area. In this paper, based on clinical and demographic data from 1105 patients collected from clinics in seven countries, and cough records from 1082 of these patients combined with convolutional neural networks and light gradient boosting machine to construct a model for the diagnosis of tuberculosis, with the final model achieving an AUC of 0.792 on the test set. This model is therefore a good reference for the auxiliary diagnosis of tuberculosis.**

Keywords-component; Deep learning; Gradient boosting; Tuberculosis diagnosis; Acoustic classification

I. Introduction

Tuberculosis (TB), a contagious ailment triggered by the bacterium Mycobacterium tuberculosis, represents a significant global health issue. In the year 2020, it is projected that around 9.9 million individuals globally will contract TB, with approximately 1.3 million succumbing to the disease. Prior to the advent of the COVID-19 pandemic, TB held the grim distinction of being the deadliest single infectious disease worldwide, surpassing even AIDS in its mortality rate [1]. However, about 40% of people with TB are not diagnosed or reported to public health because they have difficulty accessing health care or fail to receive testing or treatment when they do. The development of low-cost, non-invasive digital screening tools may improve some of the gaps in diagnosis.

Cough is a common symptom of TB and according to some previous studies, cough sounds can be used to screen for TB cases [2-4]. Although these studies were carried out in small samples or in limited settings. The field necessitates ongoing advancements and assessments for progress. In numerous studies [5,6], the effective use of machine learning algorithms has been demonstrated in analyzing acoustic properties derived from cough sounds. Deep learning has been widely used in the field of cough classification with promising results [7]. However, there is still much scope for exploring how to combine basic demographic information with cough audio data in the diagnostic process of tuberculosis.

In this study, the model was trained based on clinical and demographic data collected from clinics in seven countries for 1105 patients and the recruitment data (cough recordings) for 1082 of these patients. Firstly, the audio information was first enhanced by noise injection, stretching and pitching, and then it was further converted into a spectrogram for feature extraction. All features extracted from the audio information were used as input for the one-dimensional convolutional neural network (1D-CNN) model as stage 1. Secondly, a separate TB diagnosis model was constructed using Light Gradient Boosting Machine (LightGBM) [8] based on clinical and demographic data through feature engineering. Finally, the 1D-CNN model based on audio information and the LightGBM based on clinical and demographic data were ensembled as the stage 2 to obtain the final prediction results.

II. Methods

A. Exploratory Data Analysis

Preliminary analysis of demographic data and TB status of 1105 subjects revealed that 102 out of 517 Females (19.7%) and 195 out of 588 Males (33.2%) were found to be positive.108 out of 621 subjects with no night sweats (17.3%) and 189 out of 484 subjects with night sweats (39%) were found to be positive. 252 of 895 subjects without last week's smoking history were positive (28.2%) and 45 of 210 subjects with last week's smoking history were positive (21.4%). 233 of 957 subjects without hemoptysis were positive (24.3%) and 64 of 148 subjects with hemoptysis were positive (43.2%). 69 of 580 subjects who did not lose weight from the previous week were positive (11.9%) and 228 of 625 subjects who lost weight from the previous week were positive (36.5%). 98 of 608 subjects who did not get fever were positive (16.1%) and 199 of 497 subjects who got fever were positive (40%). Hemoptysis, night sweats and fever seem to have a significant association or effect on TB positivity. Figure 1 presents a comparative distribution of TB across various categorical variables.

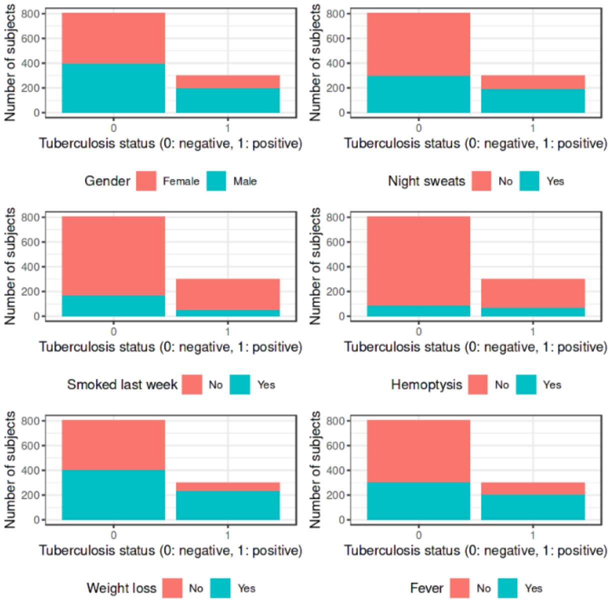

Figure 1. Comparison of the distribution of demographic data with TB status

B. Data Enhancement

One of the biggest challenges in automatic audio recognition is the preparation and enhancement of data. Audio data analysis may involve time or frequency domain, which leads to more complexity than other data formats. In general, data enhancement in audio is tantamount to an enhancement of the time-frequency spectrum (spectrogram) of audio which can be represented by image, making data enhancement in audio very similar to data enhancement in images. Noise augmentation, stretching augmentation and pitch shift augmentation were used in this project.

- Noise Augmentation: Noise augmentation refers to the process of introducing a random noise waveform to the initial audio signal. Let *x(t)* be the original signal and *n(t)* be the noise signal. The noise-augmented signal *y(t)* can be represented as:

$$y(t) = x(t) + \alpha * n(t) \tag{1}$$

where α is the noise scaling factor, determining the intensity of the noise added to the original signal. The corresponding default setting here is 0.035.

- Stretching Augmentation: Stretching augmentation pertains to the modification of the initial audio signal's duration by either expanding or condensing it within the time domain, effectively altering its duration. Let *x(t)* be the original signal and β be the stretching factor. The stretched signal *y(t)* can be represented as:

$$y(t) = x(\beta * t) \tag{2}$$

where $\beta > 1$ results in time stretching (expanding) the signal, and $0 < \beta < 1$ results in time compression (shortening) the signal. Here, 0.8 is the corresponding default setting value.

- Pitch Shift Augmentation: Pitch shift augmentation involves changing the pitch of the audio signal without affecting its duration. This is typically achieved through the use of the Short-Time Fourier Transform (STFT) and inverse STFT. Let X(ω, τ) be the STFT of the original audio signal x(t) and γ be the pitch shift factor in terms of frequency scaling. The pitch-shifted signal Y(ω, τ) in the frequency domain can be represented as:

$$Y(\omega,\tau) = X\left(\frac{\omega}{\gamma},\tau\right) \tag{3}$$

Then, the inverse STFT is applied to obtain the pitch-sifted audio signal y(t) in the time domain.

C. Feature Extraction

Features such as zero crossing rate (ZCR), Mel Frequency Cepstral Coefficients (MFCCs), chromagram from the spectrogram (Chroma STFT), mel-scaled spectrogram and root mean square (RMS) value were extracted from the audio signal [9]. MFCCs have been effectively utilized for audio examination, particularly in the realm of automated speech recognition. [10]. Label coding was used to process demographic data containing string content for multiple categorical variables.

D. Classifier Training

Convolutional Neural Networks (CNNs) are a renowned form of deep neural network architecture, primarily deployed for tasks like image classification [11]. They have also shown commendable performance in distinguishing COVID-19 related coughs [12,13]. The CNN architecture, depicted in Figure 2, includes 1D convolutional layers that have a specific kernel size and employ Rectified Linear Units (ReLUs) for activation. A dropout strategy was integrated, accompanied by max-pooling, and then followed by densely connected layers that also used ReLUs as activation functions. Adam was selected as the optimizer for training, while a learning strategy of ReduceLROnPlateau (RLRP) was adopted. The model accepted inputs with a feature dimension of 162, which were pre-processed using a standard scaler. For the training of the CNN, the data was apportioned into a validation set and a training set at a 2:8 ratio. In contrast, the LightGBM training implemented a 5-fold cross-validation approach. Detailed specifications of the training hyperparameters can be found in Table 1.

TABLE I. CLASSIFIER HYPERPARAMETERS

Hyperparameters	Classifier	Range
Batch size	CNN	128
Number of epochs	CNN	100
Number of convolutional filters	CNN	256, 128, 64
Kernel size	CNN	5
Dropout rate	CNN	0.2, 0.3
Dense layer size	CNN	32
Number of leaves	LightGBM	31
Number of estimators	LightGBM	1500
Colsample bytree	LightGBM	0.8
Learning rate	LightGBM, CNN	0.001

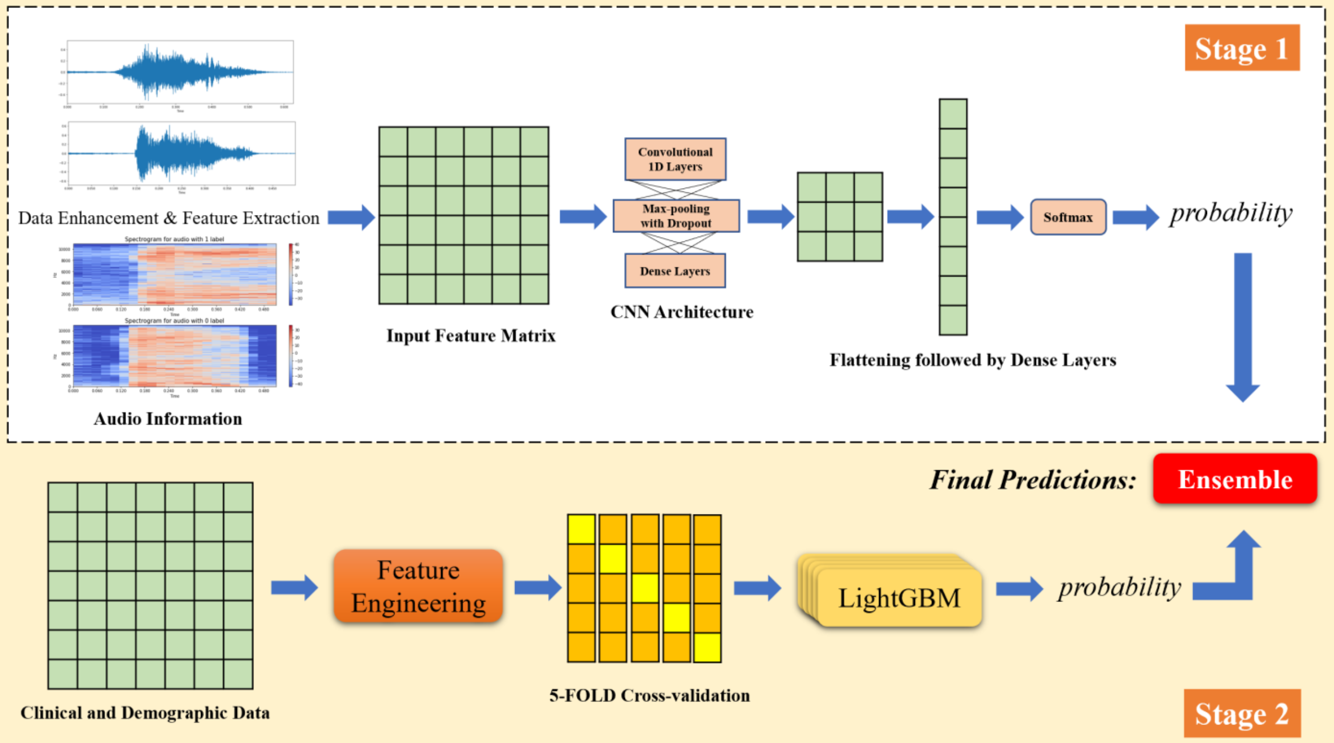

Figure 2. Schematic diagram of the model structure

III. Results

The input features of the 1D-CNN constructed in this study are treated as sequences rather than 2D grids, which is applicable to time-series data such as audio signals, where patterns can be detected across the time axis. Initially, the audio signal undergoes preprocessing and feature extraction, encompassing methods like MFCCs and other time-domain feature representations. These extracted features are then organized into a sequential format along the time axis, with a distinct feature vector corresponding to each time step. In a 1D-CNN, the convolutional layer uses 1D filters that slide over the input sequence to compute dot products with local regions. These filters learn to detect patterns, such as specific vocal tones or speech features, which are associated with audio digits. The convolution layer produces a feature map highlighting the regions where the input matches the filter pattern.

A. Performance evaluation of CNN

In the training process, the CNN model built on audio information had a tendency of overfitting. Because of the early stop setting, the model stopped running after reaching a local optimum after the 34th iteration. The accuracy on the training set displayed a linear increase in relation to the number of iterations, while the accuracy of the model on the validation set stopped to increase after the 10th round of iteration, which indicates that the model had a tendency of overfitting after training for 10 epochs (refer to the top-side graph in Figure 3). The AUC of the CNN on the validation set reached 0.76 (refer to the bottom-side graph in Figure 3), and the AUC of the LightGBM built on the demographic data exceeded 0.75 on each fold of the validation set (see table 2).

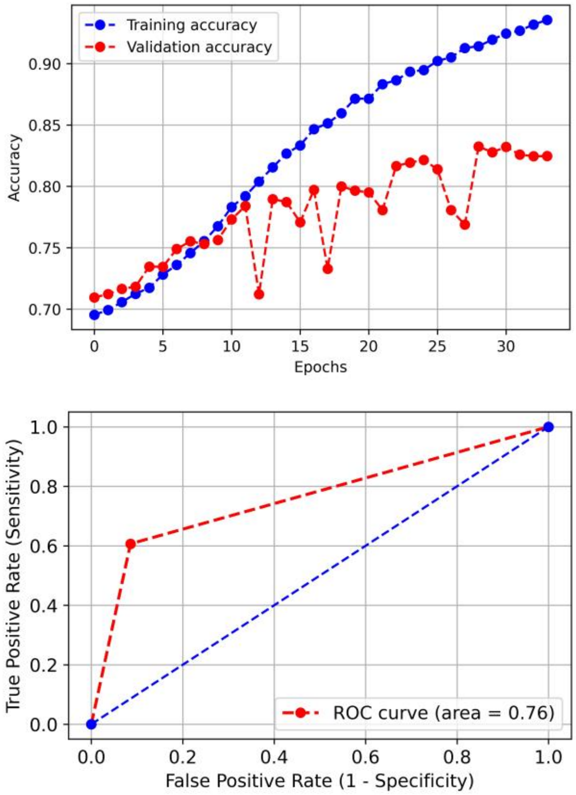

Figure 3. Model prediction performance of CNN in validation set

B. *Cross-validation based on LightGBM*

For the tabular demographic data, LightGBM model was solely used with 5-fold cross-validation to develop the Stage 2 Decision-Tree model. Since the independent patient population data is far less rich compared to the cough audio data of patients, the accuracy limit of this Decision-Tree model is markedly constrained. However, the method in this study, which combines CNN and LightGBM for independent modeling based on audio data and demographic data separately before ensemble, optimizes this issue effectively.

TABLE II. VALIDATION RESULTS OF LIGHTGBM FOR EACH FOLD

Model / Fold	LightGBM			
	Precision	*Recall*	*AUC*	*Accuracy*
FOLD 1	0.531	0.741	0.833	0.756
FOLD 2	0.568	0.712	0.846	0.774
FOLD 3	0.513	0.707	0.791	0.741
FOLD 4	0.441	0.776	0.777	0.676
FOLD 5	0.451	0.638	0.756	0.694
Mean	0.496	0.715	0.780	0.728

The 1D-CNN built on cough recordings and the LightGBM built on demographic data achieved an AUC of 0.76 and 0.78 respectively in the corresponding validation set, through the ensemble of the two models, the final AUC in the test set reached 0.792, which is a significant improvement. Moreover, this model is very efficient, the whole model training process takes only 3 minutes to complete on a P100 16G, and the inference time for a single sample is within 0.8 seconds.

C. *Comparative experiments*

To better evaluate the predictive performance of the model framework proposed in this paper, we use the same randomly selected 20% (221) of subjects as validation set. The performance is compared between different models based on demographic data, audio data, and the combination of demographic and audio data (see table 3). Among them, although TabTransformer [14] shows 10.1% increase in AUC with the addition of audio information compared to using only demographic data, its accuracy is significantly lower than the AUC of 0.792 of the model framework proposed in this paper. At the same time, we also tested the ResNet model [15], which excels in image and speech recognition tasks, but the model trained solely on audio data does not perform as well as the 1D-CNN model built in this paper, with an AUC of only 0.725.

TABLE III. COMPARISON OF MODELS UNDER DEMOGRAPHIC, AUDIO DATA

Model	**Data**	**AUC**
LightGBM	Demographic	0.780
TabTransformer [14]	Demographic	0.636
1D-CNN	Audio	0.760
ResNet [15]	Audio	0.725
TabTransformer	Demographic+Adudio	0.737
Our Model	Demographic+Adudio	0.792

IV. CONCLUSIONS

In this study, a one-dimensional network was chosen over a two-dimensional network due to the paramount importance of capturing time series information from coughing sounds. Significant differences in time series data, such as coughing rates in TB-positive patients compared to the negative population, are challenging to discern using a two-dimensional network. Moreover, the integration of pooling, fully-connected, and recursive layers demonstrated notable performance improvements, contributing to the robustness of the model. Furthermore, when compared to certain large-scale pre-trained models, this model offers a substantial cost advantage. It achieved an impressive AUC of 0.76 for the validation set, requiring only three minutes of training on a single P100 16G GPU.

In conclusion, the model developed in this study exhibits robust predictive performance by innovatively combining the advantages of deep learning models for capturing audio signal variations in audio information and machine learning models for learning probabilistic prior distributions from demographic data to maximise the predictive performance of the model while maintaining cost-effectiveness and efficiency. With a remarkably swift prediction time of less than one second for a single sample, this model facilitates the rapid diagnosis of TB based on known demographic characteristics and cough sounds. This outcome underscores the potential for efficient and accurate tuberculosis screening, contributing significantly to public health efforts in disease detection and control.

ACKNOWLEDGMENT

The datasets used for the analyses described were contributed by Dr. Adithya Cattamanchi at UCSF and Dr. Simon Grandjean Lapierre at University of Montreal and were generated in collaboration with researchers at Stellenbosch University (PI Grant Theron), Walimu (PIs William Worodria and Alfred Andama); De La Salle Medical and Health Sciences Institute (PI Charles Yu), Vietnam National Tuberculosis Program (PI Nguyen Viet Nhung), Christian Medical College (PI DJ Christopher), Centre Infectiologie Charles Merieux Madagascar (PIs Mihaja Raberahona & Rivonirina Rakotoarivelo), and Ifakara Health Institute (PIs Issa Lyimo & Omar Lweno) with funding from the U.S. National Institutes of Health (U01 AI152087), The Patrick J. McGovern Foundation and Global Health Labs. They were obtained as part of the COugh Diagnostic Algorithm for Tuberculosis (CODA TB).

REFERENCES

[1] WHO 2020 Global Tuberculosis Report 2020 (Switzerland: World Health Organization)

[2] Pathri R, Jha S, Tandon S and GangaShetty S 2022 Acoustic epidemiology of pulmonary tuberculosis (TB) & Covid19 leveraging AI/ML J. Pulmonol. Res. Reports 4 pp 2-6

[3] Pahar M, et al. 2021 Automatic cough classification for tuberculosis screening in a real-world environment Physiol. Meas. 42, 105014

[4] Botha G H R, et al. 2018 Detection of tuberculosis by automatic cough sound analysi Physiol. Meas. 39, 045005

[5] Miranda D I, Diacon H A, and Niesler R T 2019 A comparative study of features for acoustic cough detection using deep architectures Proceedings of 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp 2601-2605

[6] Laguarta J, Hueto F and Subirana B 2020 COVID-19 Artificial intelligence diagnosis using only cough recordings IEEE Open Journal of Engineering in Medicine and Biology 1 pp 275-281

[7] Pahar M, Klopper M, Warren R and Niesler T 2021 COVID-19 cough classification using machine learning and global smartphone recordings Computers in Biology and Medicine 135, 104572

[8] Qi M 2017 LightGBM: A highly efficient gradient boosting decision tree Proceedings of 31st Conference on Neural Information Processing Systems (New York: Curran Associates Inc)

[9] Han W, Chan C F, Choy C S and Pun K P 2006 An efficient MFCC extraction method in speech recognition IEEE International Symposium on Circuits and Systems, 9047629

[10] Pahar M and Smith L S 2020 Coding and decoding speech using a biologically inspired coding system 2020 IEEE Symposium Series on Computational Intelligence (SSCI) pp 3025-3032

[11] Krizhevsky A, Sutskever I and Hinton E G 2017 Imagenet classification with deep convolutional neural networks Communications of the ACM 60(6) pp 84-90

[12] Madhurananda P, Klopper M, Warren R and Niesler T 2021 COVID-19 Detection in Cough, Breath and Speech Using Deep Transfer Learning and Bottleneck Features arXiv preprint arXiv:2104.02477

[13] Pathri R, Jha S, Tandon S and Ganga Shetty S 2022 Acoustic epidemiology of pulmonary tuberculosis (TB) & Covid19 leveraging AI/ML J. Pulmonol. Res. Reports 4 pp 2-6

[14] C. Gong and T. Ren, "TabTransformer: Tabular data modeling using contextual embeddings," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, pp. 725-733, 2021.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.