Abstract—Auscultation is an important tool for diagnosing respiratory-related diseases. Unfortunately, the quality of auscultation is limited by the professional level of the doctor and the environment of the auscultation. Some studies have focused on automated auscultation techniques. However, existing approaches suffer from two challenges: (1) the models cannot learn from data distributed among multiple hospitals; and (2) the predictions of the models are difficult to interpret for physicians. To address this issue, this work proposes a novel explainable respiratory sound analysis framework with fuzzy decision tree regularization. This framework develops an ensemble knowledge distillation technique to learn distributed data and achieves good performance in terms of model efficiency and accuracy. Fuzzy decision trees are used to explain the predictions of the model and produce decision rules that can be well accepted by physicians. The effectiveness of this framework is thoroughly validated on Respiratory Sound Database and compared with other existing approaches.

Index Terms—Respiratory sounds, interpretable, fuzzy decision tree, convolutional neural network, knowledge distillation.

I. INTRODUCTION

CHRONIC Respiratory Diseases, such as Chronic Obstructive Pulmonary Disease (COPD) and Asthma, are responsible for a large number of deaths in the world, affecting more than 15% of the world population [1]. According to the prediction of the World Health Organization, COPD will become the third leading cause of death in the world [2]. Up till December of 2021, the outbreak of Coronavirus Disease 2019 (COVID-19) has caused nearly 265 million confirmed cases and globally over 5.2 million deaths due to severe pneumonia [3]. In practice, respiratory sounds are commonly used to diagnose obstructive or restrictive lung diseases [4]. By examining respiratory sounds during auscultation, the medical practitioners can identify adventitious sounds (e.g., crackles or wheezes) during the respiratory cycle. For example, crackles are the earliest sign of idiopathic pulmonary fibrosis and wheezes are usually related to COPD and asthma [5]. However, the traditional way to detect abnormalities in lung sounds by using a stethoscope could be affected by various factors such as the environmental noise, hearing fatigue, and lack of experience among junior doctors. In addition, traditional stethoscopes are not user-friendly when doctors have to wear full Personal Protective Equipment (PPE) in highly infectious environment.

To overcome the limitations of conventional auscultation, the researchers have developed digital signal processing technology for analyzing lung sounds which can be digitized and converted into signals in time domain, frequency domain, or a combination of both [6]. As it is less effective to analyze lung sounds solely in time domain or frequency domain techniques, the time-frequency techniques are more commonly applied to lung sound signals analysis [7], [8]. With the development of automated technology and adventitious sound detection, algorithms such as k-Nearest Neighbors (KNN) method, genetic algorithm, fuzzy logic, wavelet transform, etc., are used to analyze respiratory sounds [9], [10]. In addition, several studies focus on automated adventitious sound detection or the characteristics of lung sounds [11]. Although the traditional machine learning methods have contributed to the development of automated auscultation, their predictive accuracy in lung sound analysis are very limited and need to be further improved.

To achieve better predictive accuracy, deep learning-based methodologies have been proposed for respiratory sounds detection [12]–[15]. Existing studies of respiratory sounds based on deep learning technology mainly convert lung sounds into spectrograms and then analyze them using Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) [15]. As the CNN models are successful in the field of computer vision, image recognition, and audio analysis, similarly, many studies analyze respiratory sound data by first extracting Mel Frequency Cepstral Coefficients [12], Mel spectrograms [13] or local binary pattern from the lung sounds spectrograms [14], and then feed into the CNN models for training and prediction. In addition, RNN is designed to discover temporal patterns and can be used for lung sound analysis. As an advanced extension of RNN, Long short-term memory (LSTM) network has been developed for the detection of respiratory abnormal sounds and chronic/non-chronic
diseases [15]. Although the development of the deep learning models can improve the predictive accuracy of respiratory sound analysis, their real-world deployment is constrained due to lack of interpretability. Intuitively, deep learning models are intrinsically black-box models, and it is difficult for healthcare practitioners to fully trust the predictions if no explanation is available. In fact, as discussed in [16], the interpretability of a model is crucial in the medical and healthcare field.

In order to explain the deep learning models, existing explainable machine learning research [17]–[21] can be categorized into intrinsic interpretable models and post-hoc interpretable models. For intrinsic interpretable models, the key is to construct directly a model that can be explained by their internal structures. Methods in this category include the use of semantic representation constraints [17], rule-based methods [18], attention mechanisms [19], etc. Although the intrinsic interpretability has good explanatory power, it also degrades the discriminatory capability of the model [20]. In contrast, the post-hoc interpretable methods suffer less from the degeneration by learning an additional explainer network [20] or using traditional explainable models (e.g., decision tree) to approximate the original black-box model [21]. However, the above explainable methods cannot be directly applied in healthcare scenarios due to several reasons. Firstly, the well-annotated data are distributed among multiple hospitals and cannot be directly shared due to privacy. If the model only learns from one data source, the features learned from small datasets typically fail to generalize numerous patients [22]. Secondly, these models are still difficult to be understood by the healthcare practitioners.

In this work, we propose a novel framework of ensemble knowledge distillation with fuzzy logic that can achieve both interpretability and predictive accuracy in auscultation scenario where respiratory sound data are distributed (see Fig. 1). To handle the distributed data, the multiple teacher models will be trained on datasets from different sources, then these teacher models will transfer the learned patterns to a student model. Here, each teacher model can learn a large network structure from a single data source. The student model with a smaller network structure then obtains the knowledge of multiple teacher models through knowledge distillation, which can take advantage of the distributed lung sound data and achieve more efficient data analysis. Due to the ensemble knowledge distillation, the student model can integrate and learn knowledge from multiple sources without having to touch directly the datasets of the teacher model. To achieve better explanation, the decision boundaries of the student model can be approximated by a small fuzzy decision tree. Usually, the medical personnel apply vague concepts rather than specific values in reasoning about the condition of the disease in real world situation [23]. As fuzzy logic is good at managing uncertain information [24], fuzzy decision tree is a more suitable method which is easy to understand by people in other fields such as doctors. Different from decision tree with clear decision boundaries, the fuzzy decision trees can handle uncertain information better and are more interpretable [25]. To use fuzzy logic in respiratory sound analysis, respiratory sound data is first converted into the Mel spectrograms which can represent the magnitude of sound energy. Then, the fuzzy decision tree can use the different frequency bands of lung sound as features to approximate the prediction of the student model, which makes the decision of the student model more interpretable in a step-by-step manner.

To sum up, the contributions of this work are as follows:
- To learn an interpretable model from medical data that is distributed among multiple hospitals, this work proposed the ensemble knowledge distillation to learn from multiple sources. The proposed model can overcome the poor generalization due to training from a small single-sourced dataset. The distilled student model with a more compacted network structure can also be more efficient.
- To introduce interpretability of learning model in respiratory sound analysis, the proposed method incorporate a fuzzy decision tree regularization with a CNN model such that each lung sound prediction can be explained by a sequence of fuzzy logic.
- The proposed framework is tested on real datasets and the experimental results show that the proposed method can outperform the state-of-the-art counterparts.

II. RELATED WORK

A. Respiratory Sound Analysis

Limited by traditional lung sound auscultation methods, many machine learning algorithms are currently used in lung sound analysis, such as Artificial Neural Network (ANN), Hidden Markov Model, KNN, and so on [10]. Some research was motivated by cepstral features, using statistical properties of the cepstral coefficients as the feature extraction method and used ANN with Multilayer Perceptron (MLP) as a classifier for normal, wheezes, and crackles - three types of lung sounds [26]. Automated lung sound analysis process needs to be patient-friendly. For this reason, a non-invasive electronic stethoscope system based on support vector machines and CNN machine learning algorithm was constructed [12]. In order to overcome the challenge of labeling large amounts of data, a semi-supervised algorithm has been proposed to train two support vector machine classifiers to identify wheezes and crackles [22]. In addition, some research work proposed an automated lung sound analysis platform to identify abnormal
lung sounds, which does not require the labeled respiratory cycle [27]. However, these methods are proposed under the premise that there is only one source of lung sound data, and it is difficult to train the model from multiple data sources if these data do not contribute to each other. Unlike these studies, we use multiple teacher models to learn datasets from different sources, making the student model more generalized than learning from a single dataset. The use of complex classification technology and multiple computationally-expensive features would limit their use in lightweight devices [1]. Through knowledge distillation, the knowledge of the teacher models is distilled to a student model, which can reduce its parameters while retaining the classification ability of the student model.

B. Knowledge Distillation with Multiple Teachers

The development of deep learning has brought about complex models with huge overhead. The application of these complex models in production requires a lot of inference time. Knowledge distillation based on teacher-student framework as an effective model compression method tries to achieve a trade-off between model accuracy and inference efficiency [28]. To improve the applicability of the knowledge distillation framework and the accuracy of the model, some researchers have proposed to improve the framework of distillation from a single teacher to multiple teachers. For example, multiple pre-trained teacher models are directly assigned fixed weights to integrate their predictions [29]. Some researchers have used different teacher models to learn different types of inputs and then used the weighted average to teach student models [30]. In addition, to accelerate the training of the student model in the word embeddings task, multiple teacher models were used to train a student model by combining their logit values such that the student no longer needs the teachers during decoding [31]. However, these approaches only treat teacher models equally, without taking into account the differences between them. In order to solve the conflicts and competition among all teachers, Shangchen et al. formulated ensemble knowledge distillation as a multi-objective optimization problem and assigned dynamic weights to each teacher model [32]. You et al. proposed to use a voting mechanism to unify multiple relative dissimilarity information, which can be transferred into the student network [33]. These existing methods use multiple teacher models to learn category-fixed training samples. In this paper, we use multiple teachers models to learn local data distributed across multiple hospitals with flexible categories.

C. Explainable Machine Learning

There are some interpretable works based on knowledge distillation in recent years. Some studies describe the features via linear projections and univariate functions based on the additive index model [16]. However, this additive model may be biased towards selecting a few visual concepts rather than all. To overcome the typical bias-interpreting problem, the researchers distilled knowledge from a pre-trained model into an explainable additive model [20]. Sometimes, the interpretability of a network comes at the expense of its power. An explainer network was trained to explain features inside the CNN aim to trade-off between the network interpretability and the network performance [34]. The emergence of knowledge distillation provides an opportunity to explain the variables final learned by the model. Other methods attempt to distill the knowledge of neural networks into tree structures. The Gradient Boosting Trees (GBT) was used to mimic deep learning models and provided interpretable features and decision rules [35]. Some studies learned filters to make hierarchical decisions by training a soft decision tree [36]. Some researchers used model distillation techniques to learn global additive explanations, which describe the relationship between input features and model predictions [37]. Zhang et al. [21] approximated the decision boundaries of deep model via the decision tree. Besides, the decision tree was built to summarize an approximate explanation for CNN predictions at the semantic level [38]. However, it requires each CNN filter to represent a semantic concept, which limits the performance of the network. Compared to previous works [21], our proposed framework can learn multiple sources of data through knowledge distillation while protecting the privacy of the data. Moreover, the fuzzy decision tree in our work is close to the human reasoning process and more interpretable.

III. BACKGROUND

The combination of fuzzy set theory and decision tree produces a fuzzy decision tree that handles uncertainty [39]. Fuzzy ID3 is a top-down algorithm applied to construct fuzzy decision trees. In this paper, the fuzzy decision tree will be constructed by fuzzy ID3 algorithm [25].

For dataset $D$, it has $l$ attributes $A={A_1, \ldots, A_l}$, $n$ classes, and $e$ fuzzy sets $F_{1e}, F_{2e}, \ldots, F_{le}$ for the attribute $A_i$. We can calculate the information entropy by Eq. 1.

$$I(D) = -\sum_{k=1}^{n} (r_k \log_2 r_k)$$  \hspace{1cm} (1)$$

where $r_k$ denotes the proportion of the $k$-th class of samples in the current data set $D$. The fuzzy ID3 algorithm splits the attributes based on fuzzy information gain by Eq. 2 and the pseudocode is shown in Algorithm 1.

$$G(A_i, D) = I(D) - \sum_{j=1}^{e} \frac{|D_{F_{ij}}|}{\sum_{j=1}^{e} |D_{F_{ij}}|} I(D_{F_{ij}})$$  \hspace{1cm} (2)$$

where $\sum_{j=1}^{e} |D_{F_{ij}}|$ denotes the sum of the membership values for the attribute $A_i$.  

IV. OUR APPROACH

In this section, we first formally describe the problem encountered in respiratory sound analysis, and then propose an ensemble knowledge distillation to learn multi-source data and fuzzy decision tree regularization to provide model interpretation.
Algorithm 1 Fuzzy ID3 Algorithm to build decision tree.

Require: training set \( D \), attributes \( \mathcal{A} \).

1: root \( \leftarrow \) Create a root node with \( D \) and \( \mathcal{A} \).
2: \( \text{tree} \leftarrow \text{TREEGENERATE}(\text{root}, D, \mathcal{A}) \).
3: procedure TREEGENERATE(node, \( D \), \( \mathcal{A} \))
4: \( \text{if node} \) satisfies the leaf node condition then
5: \( \text{node} \leftarrow \text{mark node} \) as a leaf.
6: \( \text{else} \)
7: \( \text{for each} \) \( A_i \) in \( \mathcal{A} \) do
8: \( G(A_i, D) \leftarrow \text{Calculate information gains by Eq. 2.} \)
9: \( \text{end for} \)
10: \( \mathcal{A} \leftarrow \text{Select the attribute that maximizes information gains.} \)
11: \( \text{for each} \) \( A_i \) in \( \mathcal{A} \) do
12: \( D_j \leftarrow \text{split the sample subset by} \) \( A_i \).
13: \( \text{if} \) \( D_j \) is null then
14: \( \text{return node.} \)
15: \( \text{end if} \)
16: \( \text{child}_j \leftarrow \text{create new child node by} \) \( D_j \) and \( A_i \).
17: \( \text{node} \leftarrow \text{Connect TREEGENERATE(child}_j, \ D_j, \mathcal{A} \setminus \mathcal{A}_j) \) as branch node.
18: \( \text{end for} \)
19: \( \text{end if} \)
20: \( \text{return node} \)
21: \( \text{end procedure} \)

A. Problem Formulation

Assuming that the well-annotated lung sound is distributed across \( M \) different sources. For the \( i \)-th source, its dataset is defined as \( (X^{T_i}, Y^{T_i}) \), where the element \( x^{T_i}_n \) is composed of the time domain (td) and the frequency domain (fd). In the medical scenario, the lung sound category of each source is different. We assume that the category of \( i \)-th source is \( C_i \in \{1, \ldots, M\} \). For a hospital with dataset \( (X, Y) \), we want to learn a function \( f : \mathbb{R}^{fd \times td} \rightarrow \{0, 1\} \) and thus obtain prediction and some explainable mechanism. One problem is how we can use these distributed data, which cannot be directly shared due to privacy.

For this purpose, we design a framework to guide student model to learn the knowledge of multiple teacher models and distill the knowledge from the student model to an explainable model. We are given \( M \) teacher models \( \{T_i| i \in \{1, \ldots, M\}\} \) and a student model \( S \). In learning task, teacher \( T_i \) will learn dataset \( (X^{T_i}, Y^{T_i}) \), the label of \( x^{T_i}_n \) is \( y^{T_i}_n \in Y^{T_i} \). Let \( x^{S}_n \) denotes an example in dataset \( X \) of student model \( S \), where \( y^{S}_n \) is the label of \( x^{S}_n \). The number of elements in dataset \( X \) is \( N \).

To merge the knowledge of multiple teacher models and construct an explainable strategy require the following four steps: (i) We need to train teacher model and calculate the soft prediction with pre-trained teacher model, and expand the dimension of the soft prediction calculated by multiple teacher models. (ii) Train the student model \( S \) with soft label and fuzzy tree regularization term. (iii) Update the tree regularization term by the surrogate model. (iv) We need to iterate the training process. The details of our framework will be described in the following subsections.

B. Ensemble Knowledge Distillation

Teacher model. The teacher model is a classification network and each teacher model \( T_i \) corresponds to the lung sound category is \( C_i \). We can train the teacher model via the following loss minimization objective:

\[
L_{T_i} = -\frac{1}{N^{T_i}} \sum_{i=1}^{N^{T_i}} y^{T_i}_n \log f^{T_i}(x^{T_i}_n)
\]

where \( N^{T_i} \) is the number of elements in \( X^{T_i}, f^{T_i}(.) \) is the predictive function of \( T_i \).

The different datasets from multiple sources cannot be shared directly, the joint training of teacher model and student model is not applicable in this case [40]. Unlike this, the teacher models have achieved the ability to predict during the training process. We can distill the knowledge of teacher models via the soft prediction, for each element \( x^{S}_n \), the calculation of the soft prediction \( q^{(i)}_k(x^{S}_n) \) at a temperature of \( T \) is given as Eq. 4.

\[
q^{(i)}_k(x^{S}_n) = \frac{\exp\left(\frac{g^{(i)}_k(x^{S}_n)}{T}\right)}{\sum_k \exp\left(\frac{g^{(i)}_k(x^{S}_n)}{T}\right)}
\]

where \( g^{(i)}_k(.) \) is the logits layer output of the teacher model \( T_i \) corresponds to the \( k \)-th (\( k \in C_i \)) category. The larger the value of \( T \), the smoother the soft prediction distribution [28].

Transform soft label for student model. Since the categories of teacher models are not fixed, the dimension of their soft predictions need to be transformed into the same dimension as the student model before knowledge distillation. First, the soft predictions will be extended through Eq. 5.

\[
q^{(T_i)}_j(x^{S}_n) = \begin{cases} 
q^{(i)}_k(x^{S}_n), & j = k, \\
0, & j \neq k
\end{cases}
\]

where \( j \in \{1, \ldots, M\} \) and \( q^{(T_i)}_j(x^{S}_n) \) denotes the transformed soft predictions. After extending the dimension of the soft predictions, these soft predictions will be grouped together as ensemble soft labels \( \pi_j(x^{S}_n) \) for student model with the following formula Eq. 6.

\[
\pi_j(x^{S}_n) = \sum_{i=1}^{M} w_i q^{(T_i)}_j(x^{S}_n)
\]

where \( w_i \in [0, 1] \) denotes weight for \( q^{(T_i)}_j(x^{S}_n) \) and here \( \sum_{i=1}^{M} w_i = 1 \).

Student model. Similar to the teacher model, the soft prediction can be calculated through the student model as follows:

\[
p_j(x^{S}_n) = \frac{\exp\left(h_j(x^{S}_n)/T\right)}{\sum_{j=1}^{M} \exp\left(h_j(x^{S}_n)/T\right)}
\]

where \( h_j(.) \) is the logits layer output of the student model to the \( j \)-th category.
On the one hand, we want to reduce the difference of information between teacher and student by knowledge distillation. On the other hand, we do not want the student model to be biased towards the teacher model. We want the predictions of the student model to be close to the ground truth. The loss of ensemble knowledge distillation is as follows:

\[
L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left\{ -\lambda \, y_n^S \log f(x_n^S, \theta) + (1 - \lambda) T^2 \sum_{j=1}^{M} KL(\pi_j(x_n^S, p_j(x_n^S)) \right\}
\]

(8)

where the \( N \) is the number of elements in the dataset for \( S \), \( f(., \theta) \) is the predictive function of \( S \), and \( \theta \) is the parameters for \( S \). \( \lambda \in [0, 1] \) is the constant parameter calibrating the relative importance between the ground truth and the soft labels. \( KL(.) \) denotes the KL divergence.

C. Fuzzy Tree Regularization for Student Model

The student model will learn the knowledge of multiple teacher models through the technology of ensemble knowledge distillation. We still have a problem to solve, i.e., how to explain the student model whose predictions cannot be easily explained. We are inspired by [21], distilling the knowledge from the student model to the decision tree via tree regularization term \( \Omega(\theta) \). The tree regularization will penalize the student model and requires the behavior of the student model can be simulated through the decision tree which is a step-by-step manner.

Since the fuzzy decision tree approximates the prediction of the student model, its target is \( y_n^S \) the output of the student model for \( x_n^S \). When the feature dimension of \( x_n^S \) is relatively high, the fuzzy decision tree directly trained on this is too complex to simulate. Thus, we replace \( x_n^S \) with low-dimensional feature by using a mapping function \( \bar{\nu} : \mathcal{X} \to \mathcal{D} \), where \( x_n^D \in \mathcal{D} \). We can divide \( x_n^D \) into multiple regions, and calculate the mean value (or max value, or standard deviation, etc.) of each region to get example \( x_n^D \). In the process of calculating \( \Omega(\theta) \), the example \( x_n^D \) will be fuzzified into \( \epsilon \) fuzzy sets. For instance, for the strength of energy we can define fuzzy sets low, medium and high.

where \( m_i \) denote the parameters of membership function. The fuzzy decision tree will be trained to fit the example \( \{x_n^P, y_n^S\} \). We will prune the trained decision tree by Algorithm 2 to improve the performance and reduce the size of the fuzzy decision tree. In the process of making a prediction for an input example \( x_n^P \), the average number of decision nodes that must be passed is the Average Decision Path Length (APL). The APL as the output of \( \Omega(\theta) \) will be used to measure the complexity of this fuzzy decision tree. Then, the number of nodes will be passed when making a prediction for \( x_n^P \) will be calculated. We define the function for calculating APL as \( \text{apl}(.) \).

**Algorithm 2** Post-pruning for fuzzy decision tree.

Require: decision tree \( T \), validation data \( \{x_n, y_n\}_{n=1}^{N} \).
1: Let the squared error on validation data as \( \text{mse}(T) = \sum_{n=1}^{N} (T(x_n) - y_n)^2 \)
2: procedure **PRUNETREE(T)**
3: \( \text{error} \leftarrow \text{mse}(T) \)
4: \( \text{nodes} \leftarrow \text{Sort nodes from leaf to root} \)
5: for each node in nodes do
6: \( T' \leftarrow \text{Remove node from } T \)
7: \( \text{error}' \leftarrow \text{mse}(T') \)
8: if error' < error then
9: \( T \leftarrow T' \), \( \text{error} \leftarrow \text{error}' \)
10: end if
11: end for
12: return \( T \)
end procedure

The loss function of the student model with the tree regular term can be defined as Eq. 12. However, there is a challenge of tree regularization, \( \Omega(\theta) \) is not differentiable. If this regular term is added directly to the loss function of student model, the loss function will not be optimized by the gradient descent method. Therefore, a method is to train a surrogate model to approximate \( \Omega(\theta) \) [21].

\[
\min_{\theta} L(\theta) + \eta \Omega(\theta) \tag{12}
\]

where \( \eta \) is a constant.

We use the MLP as the surrogate model, its prediction \( \hat{\Omega}(\theta) \) is used to evaluate \( \Omega(\theta) \). In other words, the MLP is required to fit the dataset \( \{(\theta_k, \Omega(\theta_k))\}_{k=1}^{K} \) of \( K \) parameter values. The elements in the training dataset \( \{(\theta_k, \Omega(\theta_k))\}_{k=1}^{K} \) of MLP will be appended in the process of training the student model. Training the surrogate MLP by minimizing the sum of squared errors:

\[
\min_{\varphi} \sum_{k=1}^{K} (\Omega(\theta_k) - \hat{\Omega}(\theta_k, \varphi))^2 + \varepsilon \| \varphi \|^2 \tag{13}
\]

where \( \varphi \) denotes the parameters of this MLP, \( \varepsilon > 0 \) is a regularization strength.

\[
\min_{\theta} L(\theta) + \eta \hat{\Omega}(\theta) \tag{14}
\]

In practice, \( \hat{\Omega}(\theta) \) will be substituted into Eq. 12 as this tree regularization term. Wu et al. [21] provide a detailed demonstration for tree-regularized MLP. In here, we directly give
solution in Eq. 14. The ensemble knowledge distillation of the learning process is summarized in Algorithm 3.

**Algorithm 3 Summary of our approach.**

| Input: data \( \{x_T^n, y_T^n\}_{n=1}^{N_T} \) for teacher \( T \), data \( \{x_S^n, y_S^n\}_{n=1}^{N_S} \) for student model \( S \) |
| Output: The student model \( S \) and fuzzy decision tree \( T \) |
| 1: Initialize teachers parameters, student parameters \( \theta \), MLP parameters \( \varphi \), set \( l = 0 \), |
| 2: \( T_l \leftarrow \text{Construct teacher parameters, Eq. 3.} \) |
| 3: \( q_j^{(i)} \leftarrow \text{Calculate soft prediction by } T_j \text{ with Eq. 4.} \) |
| 4: \( \pi_j^{(i)} \leftarrow \text{Transform } q_j^{(i)} \text{ into soft label with Eq. 5, Eq. 6.} \) |
| 5: for \( l = l + 1 \) until \( l < \text{MAX-ITERATE} \) do |
| 6: \( \mathcal{L}(\theta) \leftarrow \text{Get loss of ensemble distillation with Eq. 8.} \) |
| 7: Construct the student model \( S \) and get prediction function \( f(.; \theta) \) with Eq. 14. |
| 8: \( \hat{y}_S^n \leftarrow \text{Get the prediction of student by } f(x_S^n, \theta). \) |
| 9: \( x_n \leftarrow \text{Get the data of } T \text{ by mapping function } \nabla \text{ and membership function Eq. 9, Eq. 10, and Eq. 11.} \) |
| 10: \( T \leftarrow \text{Prune the trained decision tree by Algorithm 2.} \) |
| 11: \( \Omega(\theta) = \frac{1}{N} \sum_n \text{ap}(T, x_n^n). \) |
| 12: Construct the surrogate MLP and approximate \( \Omega(\theta) \) by \( \hat{\Omega}(\theta) \) with Eq. 13. |
| 14: end for |

V. EXPERIMENTS

**A. Dataset**

The dataset we use is the Respiratory Sound Database created by two research teams in Portugal and Greece [41]. The database consists of a total of 5.5 h of recordings containing 4,698 respiratory cycles, of which 1,864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes from eight types of subjects. The respiratory sound audio were recorded from 126 entities, including healthy and seven respiratory diseases: Asthma, Bronchiectasis, Bronchiolitis, COPD, Lower Respiratory Tract Infection, Upper Respiratory Tract Infection, and Pneumonia.

**B. Network Architecture**

The network architecture of the teacher model and the student model are shown in Table I. The network contains Convolutional (Conv) layers, Rectified Linear Units (ReLU), Max Pooling (MP) layers, Fully Connected (FC), and Softmax layer. The \( C \) indicates the number of model classification categories. The MLP has three hidden layers with 256, 64, and 64 hidden units respectively.

**C. Experimental Setup**

We designed two experiments to validate our method. In experiment 1, we select three categories of respiratory cycles, namely, crackles, wheezes, and cycles that do not contain crackles and wheezes (normal cycles) from the Respiratory Sound Database. In experiment 2, we classify between three categories, namely, COPD, Healthy, and Pneumonia. During data preprocessing, first the lung sound recordings will be sampled by target sampling rate of 8000 Hz and transform into mono audio data.

In all experiments, the audio data will be extracted as the time-frequency results Mel spectrograms feature [42] by 2048 length of the FFT window, 512 samples between successive frames, 128 Mel bands and Hann window. Then, the time-frequency results will be split into \( 128 \times 79 \) for model input. We use learning rate of 0.001, batch size of 128, 100 epochs, \( \lambda = 0.5 \), \( T = 2.0 \), and \( \eta = 1000 \). For this experiment, we divided the training set and the test set using the 5-fold cross-validation method. The \( C \) of teacher model is used for the classification of \( M \) categories and student model are both set to 3.

In all experiments, the training set will be augmented by changing dynamic range and adding distribution noise.

For each time-frequency result, we will divide it into 16 parts in the frequency domain, each part represents a frequency band as shown in Table II, and then calculate the mean value of each part to get 16 features \( (B_0, \ldots, B_{15}) \) as input of the decision tree. Then, each feature will be fuzzified into four fuzzy sets, indicating low, medium, high, and very high energy. In addition, we apply self-organizing map algorithm [43] to determine four clustering centers as the parameters of the membership function. The clustering centers is at \(-78.9, -71.9, -63.1 \) and \(-53.9 \) for experiment 1 and at \(-78.9, -58.7, -46.3 \) and \(-29.2 \) for experiment 2. Thus, the fuzzy membership function is shown in Fig. 2.

![Image](image-url) (a) Respiratory sounds classification. (b) Lung diseases classification.

**Fig. 2. Fuzzy membership function.**

In our experiment, we compared three models [12]–[14] on our preprocessed data. Similar to our processing, these methods use the CNN to analyze the lung sounds through the spectrograms.
TABLE II

FREQUENCY BANDS AND SYMBOLS.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Frequency(Hz)</th>
<th>Symbols</th>
<th>Frequency(Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_0$</td>
<td>0, 146</td>
<td>$B_6$</td>
<td>1194, 1389</td>
</tr>
<tr>
<td>$B_1$</td>
<td>[146, 293]</td>
<td>$B_7$</td>
<td>1200, 1398</td>
</tr>
<tr>
<td>$B_2$</td>
<td>[293, 439]</td>
<td>$B_{11}$</td>
<td>1879, 2185</td>
</tr>
<tr>
<td>$B_3$</td>
<td>[439, 586]</td>
<td>$B_{12}$</td>
<td>2185, 2542</td>
</tr>
<tr>
<td>$B_4$</td>
<td>[586, 732]</td>
<td>$B_{13}$</td>
<td>2542, 2956</td>
</tr>
<tr>
<td>$B_5$</td>
<td>[732, 879]</td>
<td>$B_{14}$</td>
<td>2956, 3439</td>
</tr>
<tr>
<td>$B_6$</td>
<td>[879, 1026]</td>
<td>$B_{15}$</td>
<td>3439, 4000</td>
</tr>
</tbody>
</table>

D. Evaluation Metrics

The evaluation using the metrics of Sensitivity (SE), Specificity (SP), and average score (AS) as the authors did in [15], [44]. In experiment 1, for confusion matrix as shown in Table III, $N_{(j)}^{(i)}$, $i, j \in \{c, n, w\}$ denote the numbers of the classification results for three categories, where the subscript $j$ represents the ground truth, and the superscript $i$ represents the prediction of the model. $SE$ and $SP$ can be calculated in Eq. 15.

$$ SE = \frac{(N_c^{(c)} + N_w^{(w)})}{\sum_{i \in \{c, n, w\}} (N_c^{(i)} + N_w^{(i)})} $$

$$ SP = \frac{N_n^{(n)}}{\sum_{i \in \{c, n, w\}} N_n^{(i)}} $$

In experiment 2, for confusion matrix as shown in Table III, $co, h, p$ denote COPD, Healthy, and Pneumonia three categories respectively. $SE$ and $SP$ can be calculated in Eq. 16.

$$ SE = \frac{(N_{co}^{(co)} + N_{p}^{(p)})}{\sum_{i \in \{co, h, p\}} (N_{co}^{(i)} + N_{p}^{(i)})} $$

$$ SP = \frac{N_{h}^{(h)}}{\sum_{i \in \{co, h, p\}} N_{h}^{(i)}} $$

$$ AS = (SE + SP)/2. $$

AS can be calculated as Eq. 17. In addition, we will also indicate if the method requires the data to be assimilated before training.

E. Experiment 1: Respiratory Sounds Classification

In experiment 1, since 97% of the respiratory cycles are within five second, we extract the sound clips for the fixed duration five seconds. For the short cycles, We will add enough 0 values to ensure that the duration is not less than the duration we want.

TABLE IV

COMPARISON RESULTS OF CLASSIFYING CYCLES.

<table>
<thead>
<tr>
<th>Method</th>
<th>SP</th>
<th>SE</th>
<th>AS</th>
<th>Assimilated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aykanet [12]</td>
<td>0.82 ± 0.03</td>
<td>0.52 ± 0.02</td>
<td>0.67 ± 0.02</td>
<td>Y</td>
</tr>
<tr>
<td>Bardou [14]</td>
<td>0.90 ± 0.02</td>
<td>0.56 ± 0.04</td>
<td>0.68 ± 0.02</td>
<td>Y</td>
</tr>
<tr>
<td>Tariq [13]</td>
<td>0.90 ± 0.04</td>
<td>0.83 ± 0.03</td>
<td>0.85 ± 0.01</td>
<td>Y</td>
</tr>
<tr>
<td>Teacher</td>
<td>0.82 ± 0.02</td>
<td>0.61 ± 0.01</td>
<td>0.72 ± 0.01</td>
<td>Y</td>
</tr>
<tr>
<td>No Distill</td>
<td>0.78 ± 0.01</td>
<td>0.56 ± 0.01</td>
<td>0.67 ± 0.01</td>
<td>Y</td>
</tr>
<tr>
<td>Distill</td>
<td>0.80 ± 0.02</td>
<td>0.58 ± 0.04</td>
<td>0.69 ± 0.02</td>
<td>N</td>
</tr>
<tr>
<td>Student</td>
<td>0.80 ± 0.02</td>
<td>0.60 ± 0.01</td>
<td>0.70 ± 0.01</td>
<td>N</td>
</tr>
</tbody>
</table>

Table IV evaluates the performance of difference methods for adventitious sounds. In Table IV, model ‘Aykanet’, ‘Bardou’, and ‘Tariq’ denote our three comparison networks respectively. ‘Teacher’ denotes the teacher model fit student dataset for three categories, ‘No Distill’ denotes a model with student network architecture without knowledge distillation and tree regularization, ‘Distill’ denotes a model with student network architecture and knowledge distillation without tree regularization, and ‘Student’ is our proposed framework. Compared with other models, the teacher model has good performance for three metrics $SP$, $SE$, and $AS$. We can find that distilling knowledge to the student model through the teacher model can improve the performance of the student model. Our student model has higher performance than ‘Distill’ in terms of $AS$. The tree regularization can prevent the model from overfitting compared with ‘No Distill’. Our student model with knowledge distillation and tree regularization makes it possible to balance various evaluation metrics comparing ‘No Distill’ and ‘Distill’. Comparing the other three networks, our student model outperforms model ‘Aykanet’, ‘Bardou’, and ‘Tariq’ by 3%, 2%, 5% in $AS$. Besides, the student model is only 0.01 lower in $SE$ than the teacher model and higher than other models. The performance of the student model on this metric and the characteristics of learning from multiple sources data are more suitable for the medical care. Although the student model is lower in the first three metrics than the teacher model, the student model has fewer parameters and higher prediction accuracy.

The decision tree generated by the model is shown in Fig. 3. The leaf nodes of the decision tree are the one class with the highest prediction probability and the prediction probability of this class. We use the probability that the predictions of the decision tree on the test set agree with the predictions of the student model as material for the confidence level of the decision tree. In this task, the confidence level of the fuzzy decision tree is 68%. The decision tree is an interpretable model whose decision rules can be extracted directly. The decision tree branches with a probability of less than 70% are pruned in order to pay more attention to high-probability rules. To show the decision rules more visually, we list the decision rules with probability greater than 90% as shown in Fig. 4.

The three types of test samples of crackles, normal, and wheezes are shown in Fig. 7 respectively. When the student model predicts a sample, we can obtain the Mel Spectrogram,
networks on various evaluation materials. Since our student

Table V reports the performance of our model and comparison

Fig. 3. Fuzzy decision tree of respiratory sounds classification. For branches with a probability greater than 90%, we added a branch number at the end of them to more intuitively associate the branch with the relevant rule.

1. wheezes(1.00) ← B15(low) ∧ B11(medium) ∧ B7(low) ∧ B3(medium) ∧ B5(high) ∧ B10(low)
2. normal(1.00) ← B15(low) ∧ B11(low) ∧ B3(high) ∧ B9(low)
3. wheezes(0.96) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(high) ∧ B5(high) ∧ B9(low)
4. crackles(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B14(low) ∧ B1(low) ∧ B3(medium)
5. crackles(1.00) ← B35(low) ∧ B11(low) ∧ B4(very high) ∧ B19(medium) ∧ B4(very high) ∧ B4(medium)
6. crackles(1.00) ← B35(low) ∧ B11(medium) ∧ B10(very high)
7. crackles(1.00) ← B35(low) ∧ B11(low) ∧ B7(high) ∧ B1(medium) ∧ B4(very high) ∧ B12(low) ∧ B10(high)
8. normal(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(medium) ∧ B4(very high) ∧ B6(medium) ∧ B3(very high)
9. normal(1.00) ← B15(low) ∧ B11(low) ∧ B10(medium) ∧ B12(medium) ∧ B3(high) ∧ B7(very high)
10. wheezes(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(high) ∧ B5(very high) ∧ B4(high) ∧ B12(low) ∧ B5(low)
11. crackles(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(high) ∧ B12(medium) ∧ B6(high) ∧ B10(low) ∧ B7(medium) ∧ B3(medium)
12. crackles(1.00) ← B15(medium) ∧ B11(low) ∧ B35(medium) ∧ B12(medium) ∧ B5(high) ∧ B7(high) ∧ B12(very high) ∧ B14(medium)
13. crackles(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(high) ∧ B14(low) ∧ B6(very high) ∧ B7(very high) ∧ B5(high) ∧ B3(high) ∧ B5(high) ∧ B6(high) ∧ B9(medium)
14. normal(1.00) ← B15(low) ∧ B11(low) ∧ B7(high) ∧ B1(high) ∧ B14(low) ∧ B6(very high) ∧ B7(very high) ∧ B5(high) ∧ B5(high) ∧ B6(high) ∧ B9(medium)

Fig. 4. The decision rule for respiratory sounds classification. The rule number here corresponds to the branch number in Fig. 3.

fuzzy energy level, and decision path of the sample. Through such a decision rule, the prediction of the model can be simply simulated by the doctor.

F. Experiment 2: Lung Diseases Classification

In this experiment, since the duration of most respiratory cycles is about 20s, we will the sound clips for the fixed duration five seconds to get the same length of sound clips. Table V reports the performance of our model and comparison networks on various evaluation materials. Since our student
model uses the knowledge distillation and tree regularization, it has higher performance than model ‘Aykanat’ and ‘Tariq’ in terms of AS. More importantly, the performance of student model is only 0.01 and 0.02 lower than model ‘Bardou’ and teacher model respectively. Similarly, the student model achieves better performance on metric SP than ‘No distill’ and ‘Distill’ due to the combined effect of knowledge distillation and tree regularization. The student model does not require the data to be assimilated to learn from multiple data sources.

The decision tree generated by the model is shown in Fig. 5. The confidence level of the fuzzy decision tree is 84%, which indicates that the decisions can be trusted for most of the samples. The pruning process of the decision tree is the same as that in Experiment 1. As above, we list the decision paths
that have a prediction probability greater than 90% on the decision tree as shown in Fig. 6. The three types of test samples of COPD, Healthy, and Pneumonia are shown in Fig. 8 respectively. The fuzzy decision trees allow us not to care about the specific values of the decision boundary, which is more in line with our behavior when making decisions. The decision path is directly translated into a decision rule that is easy for physicians to understand.

VI. CONCLUSION

This work proposes a novel explainable CNN framework based on fuzzy decision tree regularization for respiratory sounds analysis, which can learn distributed data from multiple hospitals and provide decision rules that can be simulated by
physicians. In this framework, teacher models can learn non-fixed categories of data and transfer their knowledge to student models. More importantly, fuzzy decision trees are used to deal with the uncertainty in the decision process, thus improving the interpretability of the model. Decision rules generated by fuzzy decision trees in this form are more easily accepted by physicians. This framework is evaluated on Respiratory Sound Database. The experimental results show that our framework can learn from distributed data, simplify the parameters of the model with less performance loss, and provide easily acceptable and interpretable fuzzy decision trees, compared to other methods. In future work, we aim to apply this framework to hospital auscultation systems.

REFERENCES


Jianqiang Li (Member, IEEE) received the B.S. and Ph.D. degrees from the South China University of Technology, Guangzhou, China, in 2003 and 2008, respectively. He is a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He led five projects of the National Natural Science Foundation, and four projects of the Natural Science Foundation of Guangdong Province, China. His current research interests include robotics, embedded systems and Internet of Things.

Cheng Wang received the B.S. degree from the University of South China, Hengyang, China, in 2019. He is currently pursuing the master’s degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His current research interests include intelligent healthcare system.

Jie Chen joined Shenzhen University as an assistant professor since 2021. He received his B.Sc. degree in EE from Taizhou University (China) in 2006, M.Sc. degree in CS from Zhejiang University (China) in 2008, and Ph.D. degree in CS from National University of Singapore (NUS) in 2013. He received Dean’s Graduate Research Excellence Award from NUS in 2015. He was a postdoctoral associate at Singapore-MIT Alliance for Research and Technology (SMART) from 2013 to 2018. He worked at Shenzhen University as a research-track associate professor from 2018 to 2021. He was a program committee member in IJCAI, AAAI, and ICRA. His research interests include machine learning, multi-agent system, robotics, intelligent transportation system and intelligent healthcare system.

Heng Zhang received the master’s degree from the Capital Medical University, Beijing, China, in 2012. She is an Attending Physician with the Department of Respiratory and Critical Care Medicine, Shenzhen People’s Hospital, Shenzhen, China. She is engaged in the frontline clinical work of respiratory and critical medicine. She is with Department of Respiratory and Critical Care Medicine, Shenzhen People’s Hospital, First Affiliated Hospital of Southern University of Science and Technology, Second Clinical Medical College of Jinan University, Shenzhen Institute of Respiratory Diseases, Shenzhen, China. Her current research interests include infectious diseases of the respiratory system.

Yuyan Dai received the B.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020. She is currently pursuing the master’s degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her current research interests include smart medicine and blind source signal separation.

Lingwei Wang is the Deputy Director of Shenzhen Institute of Respiratory Diseases and the Director of the Diagnostics Department at the School of Medicine, Southern University of Science and Technology. She is with Department of Respiratory and Critical Care Medicine, Shenzhen People’s Hospital, First Affiliated Hospital of Southern University of Science and Technology, Second Clinical Medical College of Jinan University, Shenzhen Institute of Respiratory Diseases, Shenzhen Key Laboratory of Respiratory Diseases, Shenzhen, China.

Li Wang received his B.Sc. degree in BME from Zhejiang University (China) in 2008, and Ph.D. degree in BME Zhejiang University (China) in 2017. He was a senior engineer at ZICT Technology Co., Ltd from 2017 to 2019. He is a postdoctoral associate at Shenzhen University since 2019. His research interests include medical informatics, regional healthcare system, machine learning, and artificial intelligence.

Asoke K. Nandi (Fellow, IEEE) received the degree of Ph.D. in Physics from the University of Cambridge (Trinity College), Cambridge (UK). He held academic positions in several universities, including Oxford (UK), Imperial College London (UK), Strathclyde (UK), and Liverpool (UK) as well as Finland Distinguished Professorship in Jyvaskyla (Finland). In 2013, he moved to Brunel University London (UK), to become the Chair and Head of Electronic and Computer Engineering. Professor Nandi is a Distinguished Visiting Professor at Shenzhen University (China) and an Adjunct Professor at University of Calgary (Canada). Professor Nandi is a Fellow of the Royal Academy of Engineering (UK) as well as a Fellow of seven other institutions, including the IEEE and the IET. He has published more than 600 research articles, with a Google Scholar H-index of 80.