Explainable CNN With Fuzzy Tree Regularization for Respiratory Sound Analysis

Jianqiang Li, *Member, IEEE*, Cheng Wang, Jie Chen, Heng Zhang, Yuyan Dai, Lingwei Wang, Li Wang, and Asoke K. Nandi, *Fellow, IEEE*

***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 article 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 the Respiratory Sound database and compared with other existing approaches.**

***Index Terms*—Convolutional neural network, fuzzy decision tree, interpretable, knowledge distillation, respiratory sounds.**

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 2021, the outbreak of coronavirus disease 2019 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 in highly infectious environments.

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 signal 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 is very limited and needs to be further improved.

To achieve better predictive accuracy, deep-learning-based methodologies have been proposed for respiratory sound 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 (CNNs) or recurrent neural networks (RNNs) [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 feeding them into the CNN models for training and

Manuscript received August 4, 2021; revised December 11, 2021; accepted January 6, 2022. Date of publication January 21, 2022; date of current version June 2, 2022. This work was supported in part by the National Key R&D Program of China under Grant 2020YFA0908700, in part by the National Natural Science Foundation of China under Grant 62073225, Grant 62072315, Grant 61836005, and Grant 62006157, in part by the Natural Science Foundation of Guangdong Province Outstanding Youth Program under Grant 2019B151502018, in part by the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, in part by the Shenzhen Science and Technology Program under Grant JCYJ20210324093808021, and in part by the Shenzhen Science and Technology Innovation Commission under Grant R2020A045. *(Corresponding author: Jie Chen.)*

Jianqiang Li, Cheng Wang, Jie Chen, Yuyan Dai, and Li Wang are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: lijq@szu.edu.cn; wangcheng2019@email.szu.edu.cn; chenjie@szu.edu.cn; daiyuyan2020@email.szu.edu.cn; wangli100@szu.edu.cn).

Heng Zhang and Lingwei Wang are with the Department of Respiratory and Critical Care Medicine, Shenzhen People's Hospital, Shenzhen 518020, China (e-mail: zhanghabcde@163.com; limey12@vip.sina.com).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH London, U.K. (e-mail: asoke.nandi@brunel.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TFUZZ.2022.3144448.

Digital Object Identifier 10.1109/TFUZZ.2022.3144448

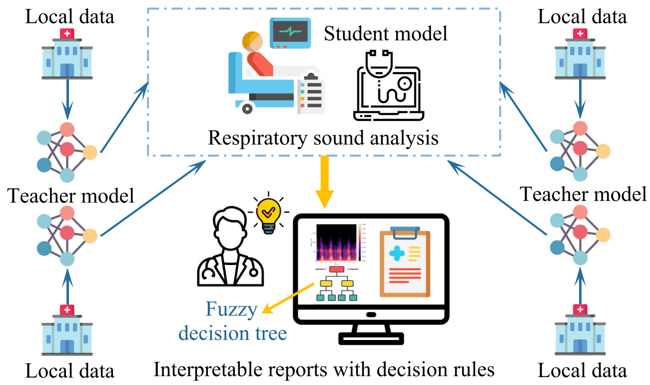


Fig. 1. Framework overview.

prediction. In addition, the RNN is designed to discover temporal patterns and can be used for lung sound analysis. As an advanced extension of the RNN, a long short-term memory network has been developed for the detection of respiratory abnormal sounds and chronic/nonchronic 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 degenerates 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. First, 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]. Second, these models are still difficult to be understood by the healthcare practitioners.

In this article, we propose a novel framework of ensemble knowledge distillation with fuzzy logic that can achieve both interpretability and predictive accuracy in the 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 are 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 article are as follows.

1) To learn an interpretable model from medical data that is distributed among multiple hospitals, this article proposes the ensemble knowledge distillation to learn from multiple sources. The proposed model can overcome the poor generalization due to training from a small single-source dataset. The distilled student model with a more compacted network structure can also be more efficient.
2) To introduce interpretability of learning model in respiratory sound analysis, the proposed method incorporates a fuzzy decision tree regularization with a CNN model such that each lung sound prediction can be explained by a sequence of fuzzy logic.
3) 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 sound [26]. An automated lung sound analysis process needs to be patient-friendly. For this reason, a noninvasive electronic stethoscope system based on support vector machines and the CNN machine learning algorithm was constructed [12]. In order to overcome the challenge of labeling large amount of data, a semisupervised algorithm has been proposed to train two support vector machine classifiers to identify wheezes and crackles [22]. In addition, some research works 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 the teacher–student framework as an effective model compression method tries to achieve a tradeoff 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 pretrained 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 embedding 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, Du *et al.* [32] formulated ensemble knowledge distillation as a multiobjective optimization problem and assigned dynamic weights to each teacher model. You *et al.* [33] proposed to use a voting mechanism to unify multiple relative dissimilarity information, which can be transferred into the student network. These existing methods use multiple-teacher models to learn category-fixed training samples. In this article, we use multiple-teacher models to learn local data distributed across multiple hospitals with flexible categories.

C. Explainable Machine Learning

There have been 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 toward selecting a few visual concepts rather than all. To overcome the typical bias-interpreting problem, the researchers distilled knowledge from a pretrained 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 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]. Wu *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 article, the fuzzy decision tree will be constructed by the fuzzy ID3 algorithm [25].

For dataset $\mathcal{D}$, it has l attributes $\mathcal{A} = \{A_1, \ldots, A_l\}$, n classes, and e fuzzy sets $F_{i1}, F_{i2}, \ldots, F_{ie}$ for the attribute A_i. We can calculate the information entropy as

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

where r_k denotes the proportion of the kth class of samples in the current dataset $\mathcal{D}$. The fuzzy ID3 algorithm splits the attributes based on fuzzy information gain by (2), and the pseudocode is shown in Algorithm 1

$$G(A_i, \mathcal{D}) = I(\mathcal{D}) - \sum_{j=1}^{e} \frac{|\mathcal{D}_{F_{ij}}|}{\sum_{j=1}^{e} |\mathcal{D}_{F_{ij}}|} I(\mathcal{D}_{F_{ij}}) \tag{2}$$

where $\sum_{j=1}^{e} |\mathcal{D}_{F_{ij}}|$ 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 multisource

Algorithm 1: Fuzzy ID3 Algorithm to Build Decision Tree.

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

$root \leftarrow$ Create a root node with $\mathcal{D}$ and $\mathcal{A}$.
$tree \leftarrow$ TREEGENERATE($root$, $\mathcal{D}$, $\mathcal{A}$).
procedure TreeGenerate ($node$, $\mathcal{D}$, $\mathcal{A}$)
  if $node$ satisfies the leaf node condition **then**
    $node \leftarrow$ mark $node$ as a leaf.
  else
    for each A_i in $\mathcal{A}$ **do**
      $G(A_i, \mathcal{D}) \leftarrow$ Calculate information gains by Eq. 2.
    end for
    $\mathcal{A}_* \leftarrow$ Select the attribute that maximizes information gains.
    for each A_*^j in $\mathcal{A}_*$ **do**
      $D_j \leftarrow$ split the sample subset by A_*^j.
      if D_j is null **then**
        return $node$.
      end if
      $child_j \leftarrow$ create new child node by D_j and A_*^j.
      $node \leftarrow$ Connect TREEGENERATE($child_j$, D_j, $\mathcal{A} \backslash \mathcal{A}_*$) as branch node.
    end for
  end if
  return $node$
end procedure

data and fuzzy decision tree regularization to provide model interpretation.

A. Problem Formulation

Assuming that the well-annotated lung sound is distributed across M different sources. For the ith source, its dataset is defined as $(\mathcal{X}^{\mathrm{T}_i}, \mathcal{Y}^{\mathrm{T}_i})$, where the element $x_n^{\mathrm{T}_i}$ 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 the ith source is $\mathcal{C}_i \subseteq \{1, \ldots, M\}$. For a hospital with dataset $(\mathcal{X}, \mathcal{Y})$, we want to learn a function $f_S : \mathbb{R}^{fd \times td} \mapsto \{0, 1\}$ 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 a 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 $\{\mathrm{T}_i | i \in \{1, \ldots, M\}\}$ and a student model S. In the learning task, teacher T_i will learn dataset $(\mathcal{X}^{\mathrm{T}_i}, \mathcal{Y}^{\mathrm{T}_i})$; the label of $x_n^{\mathrm{T}_i}$ is $y_n^{\mathrm{T}_i} \in \mathcal{Y}^{\mathrm{T}_i}$. Let x_n^{S} denote an example in the dataset $\mathcal{X}$ of the student model S, where y_n^{S} is the label of x_n^{S}. The number of elements in the dataset $\mathcal{X}$ is N.

To merge the knowledge of multiple-teacher models and construct an explainable strategy requires the following four steps: 1) we need to train teacher model and calculate the soft prediction with a pretrained teacher model and expand the dimension of the soft prediction calculated by multiple-teacher models; 2) train the student model S with soft label and fuzzy tree regularization term; 3) update the tree regularization term by the surrogate model; and 4) we need to iterate the training process. The details of our framework will be described in the following subsections.

B. Ensemble Knowledge Distillation

1) Teacher Model: The teacher model is a classification network, and each teacher model T_i corresponds to the lung sound category $\mathcal{C}_i$. We can train the teacher model via the following loss minimization objective:

$$\mathcal{L}_{\mathrm{T}_i} = -\frac{1}{N^{\mathrm{T}_i}} \sum_{n=1}^{N^{\mathrm{T}_i}} y_n^{\mathrm{T}_i} \log f^{\mathrm{T}_i}\left(x_n^{\mathrm{T}_i}\right) \tag{3}$$

where N^{T_i} is the number of elements in $\mathcal{X}^{\mathrm{T}_i}$ and $f^{\mathrm{T}_i}(.)$ is the predictive function of T_i.

The different datasets from multiple sources cannot be shared directly; the joint training of the teacher model and the 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_n^{S}, the calculation of the soft prediction $q_k^{(i)}$ at a temperature of T is given as

$$q_k^{(i)}\left(x_n^{\mathrm{S}}\right) = \frac{\exp\left(g_k^{(i)}\left(x_n^{\mathrm{S}}\right)/T\right)}{\sum_k \exp\left(g_k^{(i)}\left(x_n^{\mathrm{S}}\right)/T\right)} \tag{4}$$

where $g_k^{(i)}(.)$ is the logit layer output of the teacher model T_i that corresponds to the kth ($k \in \mathcal{C}_i$) category. The larger the value of T, the smoother the soft prediction distribution [28].

2) Transform Soft Label for the Student Model: Since the categories of teacher models are not fixed, the dimension of their soft predictions needs to be transformed into the same dimension as the student model before knowledge distillation. First, the soft predictions will be extended through

$$q_j^{(T_i)}\left(x_n^{\mathrm{S}}\right) = \begin{cases} q_k^{(i)}\left(x_n^{\mathrm{S}}\right), & j = k \\ 0, & j \neq k \end{cases} \tag{5}$$

where $j \in \{1, \ldots, M\}$ and $q_j^{(T_i)}(x_n^{\mathrm{S}})$ denotes the transformed soft predictions. After extending the dimension of the soft predictions, these soft predictions will be grouped together as ensemble soft labels $\pi_j(x_n^{\mathrm{S}})$ for the student model with the following formula:

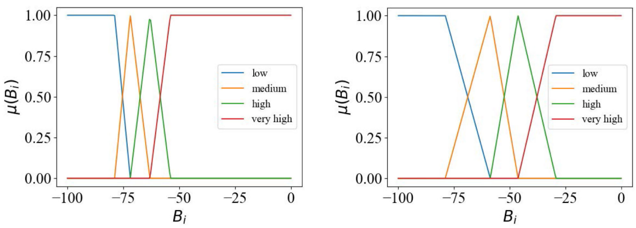
$$\pi_j(x_n^{\mathrm{S}}) = \sum_{i=1}^{M} w_i q_j^{(T_i)}\left(x_n^{\mathrm{S}}\right) \tag{6}$$

where $w_i \in [0, 1]$ denotes weight for $q_j^{(T_i)}(x_n^{\mathrm{S}})$ and here $\sum_{i=1}^{M} w_i = 1$.

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

$$p_j(x_n^{\mathrm{S}}) = \frac{\exp\left(h_j\left(x_n^{\mathrm{S}}\right)/T\right)}{\sum_{j=1}^{M} \exp\left(h_j\left(x_n^{\mathrm{S}}\right)/T\right)} \tag{7}$$

where $h_j(.)$ is the logit layer output of the student model to the jth category.



(a) (b)

Fig. 2. Fuzzy membership function. (a) Respiratory sound classification. (b) Lung disease classification.

On the one hand, we want to reduce the difference of information between the teacher and the student by knowledge distillation. On the other hand, we do not want the student model to be biased toward 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:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\Bigg\{-\lambda\, y_n^{\mathrm{S}} \log f\left(x_n^{\mathrm{S}}, \boldsymbol{\theta}\right) + (1-\lambda)\, T^2 \sum_{j=1}^{M} KL\left(\pi_j(x_n^{\mathrm{S}}), p_j(x_n^{\mathrm{S}})\right)\Bigg\} \tag{8}$$

where N is the number of elements in the dataset for S. $f(.,\theta)$ is the predictive function of S, and $\boldsymbol{\theta}$ is the parameter 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 Kullback–Leibler divergence.

C. Fuzzy Tree Regularization for the 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 simulated. We are inspired by [21], distilling the knowledge from the student model to the decision tree via tree regularization term $\Omega(\boldsymbol{\theta})$. 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 $\hat{y}_n^{\mathrm{S}}$, which 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 a low-dimensional feature by using a mapping function $\nabla : \mathcal{X} \rightarrow \mathcal{D}$, where $x_n^{\mathcal{D}} \in \mathcal{D}$. We can divide x_n^{S} into multiple regions and calculate the mean value (or max value, or standard deviation, etc.) of each region to get example $x_n^{\mathcal{D}}$. In the process of calculating $\Omega(\boldsymbol{\theta})$, the example $x_n^{\mathcal{D}}$ will be fuzzified into e fuzzy sets. For instance, for the strength of energy, we can define fuzzy sets low, medium, and high. For the membership degree, values of the ith ($i \in \{1,\ldots,e\}$) fuzzy sets can be calculated by triangular

Algorithm 2: Postpruning for Fuzzy Decision Tree.

Require: decision tree $\mathcal{T}$, validation data $\{x_n, y_n\}_{n=1}^{N}$.
Let the squared error on validation data as $\mathrm{mse}(\mathcal{T}) = \sum_{n=1}^{N}(\mathcal{T}(x_n) - y_n)^2$
procedure PruneTree$\mathcal{T}$
  $error \leftarrow \mathrm{mse}(\mathcal{T})$.
  $nodes \leftarrow$ Sort nodes from leaf to root.
  for each $node$ in $nodes$ **do**
    $\mathcal{T}' \leftarrow$ Remove $node$ from $\mathcal{T}$.
    $error' \leftarrow \mathrm{mse}(\mathcal{T}')$.
    if $error' < error$ **then**
      $\mathcal{T} \leftarrow \mathcal{T}'$, $error \leftarrow error'$
    end if
  end for
  return $\mathcal{T}$
end procedure

membership function $\mu_i(.)$ as follows:

$$\mu_1\left(x_n^{\mathcal{D}}\right) = \begin{cases} 1, & x_n^{\mathcal{D}} \leq m_1, \\ \frac{m_2 - x_n^{\mathcal{D}}}{m_2 - m_1}, & m_1 < x_n^{\mathcal{D}} < m_2 \\ 0, & x_n^{\mathcal{D}} \geq m_2 \end{cases} \tag{9}$$

$$\mu_i\left(x_n^{\mathcal{D}}\right) = \begin{cases} 0, & x_n^{\mathcal{D}} \leq m_{i-1} \\ \frac{x_n^{\mathcal{D}} - m_{i-1}}{m_i - m_{i-1}}, & m_{i-1} \leq x_n^{\mathcal{D}} \leq m_i \\ \frac{m_{i+1} - x_n^{\mathcal{D}}}{m_{i+1} - m_i}, & m_i \leq x_n^{\mathcal{D}} \leq m_{i+1} \\ 0, & x_n^{\mathcal{D}} \geq m_{i+1} \end{cases} \tag{10}$$

$$\mu_e\left(x_n^{\mathcal{D}}\right) = \begin{cases} 0, & x_n^{\mathcal{D}} \leq m_{e-1} \\ \frac{x_n^{\mathcal{D}} - m_{e-1}}{m_e - m_{e-1}}, & m_{e-1} < x_n^{\mathcal{D}} < m_e \\ 1, & x_n^{\mathcal{D}} \geq m_e \end{cases} \tag{11}$$

where m_i denote the parameters of the membership function. The fuzzy decision tree will be trained to fit the example $\{x_n^{\mathcal{D}}, \hat{y}_n^{\mathrm{S}}\}$. 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 $x_n^{\mathcal{D}}$, the average number of decision nodes that must be passed is the average decision path length (APL). The APL as the output of $\Omega(\boldsymbol{\theta})$ 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 $x_n^{\mathcal{D}}$ will be calculated. We define the function for calculating APL as $\mathrm{apl}(.)$.

The loss function of the student model with the tree regular term can be defined as (12). However, there is a challenge of tree regularization: $\Omega(\boldsymbol{\theta})$ is not differentiable. If this regular term is added directly to the loss function of the 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 $\Omega(\boldsymbol{\theta})$ [21]

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \eta\Omega(\boldsymbol{\theta}) \tag{12}$$

where η is a constant.

We use the MLP as the surrogate model; its prediction $\hat{\Omega}(\boldsymbol{\theta})$ is used to evaluate $\Omega(\boldsymbol{\theta})$. In other words, the MLP is required

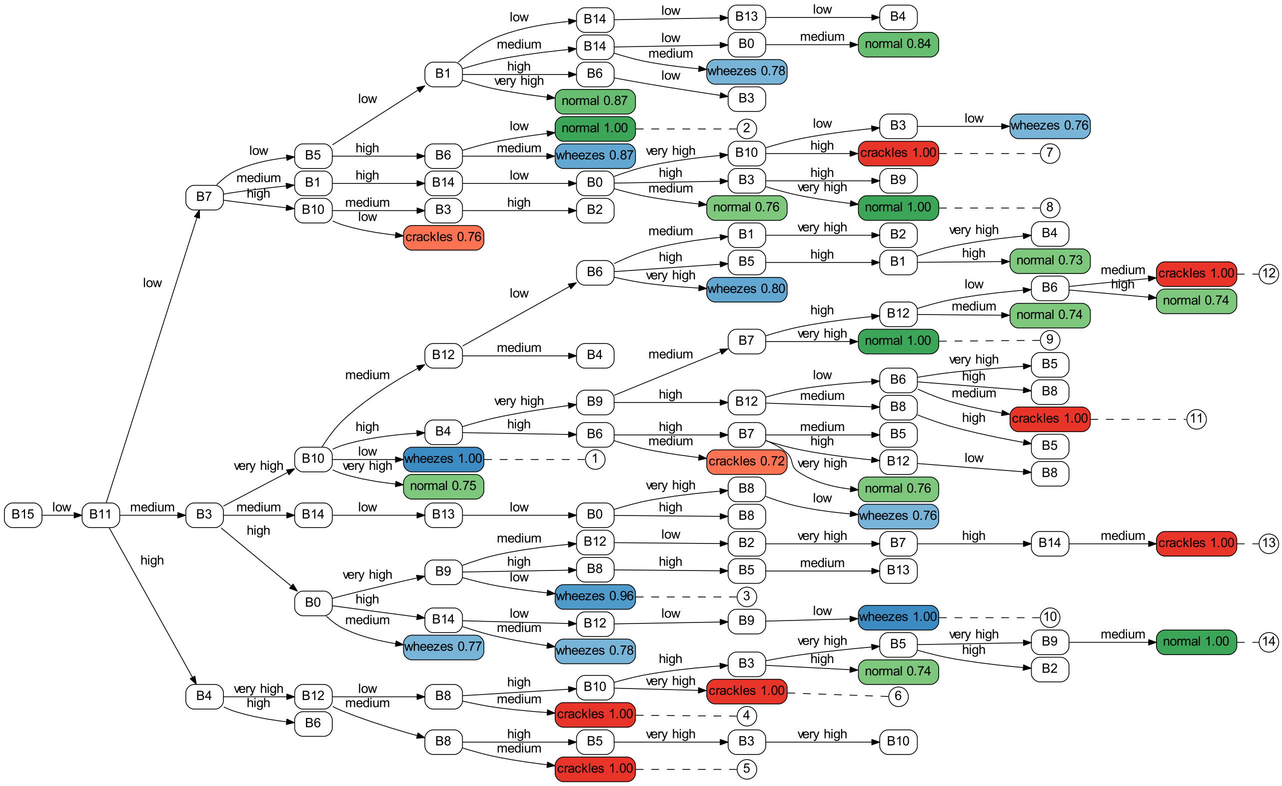


Fig. 3. Fuzzy decision tree of respiratory sound classification. For branches with a probability greater than 90%, we added a branch number at the end of them to more intuitively associate the branch with the relevant rule.

1 $\mathit{wheezes}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{medium}) \wedge B_7(\text{low}) \wedge B_3(\text{medium}) \wedge B_5(\text{high}) \wedge B_{10}(\text{low})$

2 $\mathit{normal}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_0(\text{high}) \wedge B_6(\text{low})$

3 $\mathit{wheezes}(0.96) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{high}) \wedge B_3(\text{high}) \wedge B_9(\text{low})$

4 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_{14}(\text{low}) \wedge B_1(\text{low}) \wedge B_8(\text{medium})$

5 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_4(\text{very high}) \wedge B_{10}(\text{medium}) \wedge B_0(\text{very high}) \wedge B_8(\text{medium})$

6 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_9(\text{medium}) \wedge B_{10}(\text{very high})$

7 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{medium}) \wedge B_4(\text{very high}) \wedge B_{12}(\text{low}) \wedge B_{10}(\text{high})$

8 $\mathit{normal}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{medium}) \wedge B_4(\text{very high}) \wedge B_6(\text{medium}) \wedge B_3(\text{very high})$

9 $\mathit{normal}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_{10}(\text{medium}) \wedge B_{12}(\text{medium}) \wedge B_9(\text{high}) \wedge B_7(\text{very high})$

10 $\mathit{wheezes}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{high}) \wedge B_6(\text{low}) \wedge B_6(\text{high}) \wedge B_{13}(\text{low}) \wedge B_9(\text{low})$

11 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{high}) \wedge B_{12}(\text{medium}) \wedge B_0(\text{high}) \wedge B_{10}(\text{low}) \wedge B_7(\text{medium}) \wedge B_6(\text{medium})$

12 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_{10}(\text{medium}) \wedge B_{12}(\text{medium}) \wedge B_9(\text{high}) \wedge B_5(\text{very high}) \wedge B_7(\text{high}) \wedge B_{12}(\text{low}) \wedge B_6(\text{medium})$

13 $\mathit{crackles}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_{10}(\text{medium}) \wedge B_{12}(\text{medium}) \wedge B_9(\text{high}) \wedge B_5(\text{very high}) \wedge B_7(\text{high}) \wedge B_{14}(\text{medium})$

14 $\mathit{normal}(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_1(\text{high}) \wedge B_{14}(\text{low}) \wedge B_0(\text{very high}) \wedge B_0(\text{high}) \wedge B_3(\text{very high}) \wedge B_6(\text{high}) \wedge B_5(\text{high}) \wedge B_9(\text{medium})$

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

to fit the dataset $\{\boldsymbol{\theta}_k, \Omega(\boldsymbol{\theta}_k)\}_{k=1}^{K}$ of K parameter values. The elements in the training dataset $\{\boldsymbol{\theta}_k, \Omega(\boldsymbol{\theta}_k)\}_{k=1}^{K}$ of the MLP will be appended in the process of training the student model. Training the surrogate MLP by minimizing the sum of squared errors

$$\min_{\boldsymbol{\varphi}} \sum_{k=1}^{K} \left(\Omega(\boldsymbol{\theta}_k) - \hat{\Omega}(\boldsymbol{\theta}_k, \boldsymbol{\varphi})\right)^2 + \varepsilon \left\|\boldsymbol{\varphi}\right\|_2^2 \qquad (13)$$

where $\boldsymbol{\varphi}$ denotes the parameters of this MLP and $\varepsilon > 0$ is a regularization strength, we have

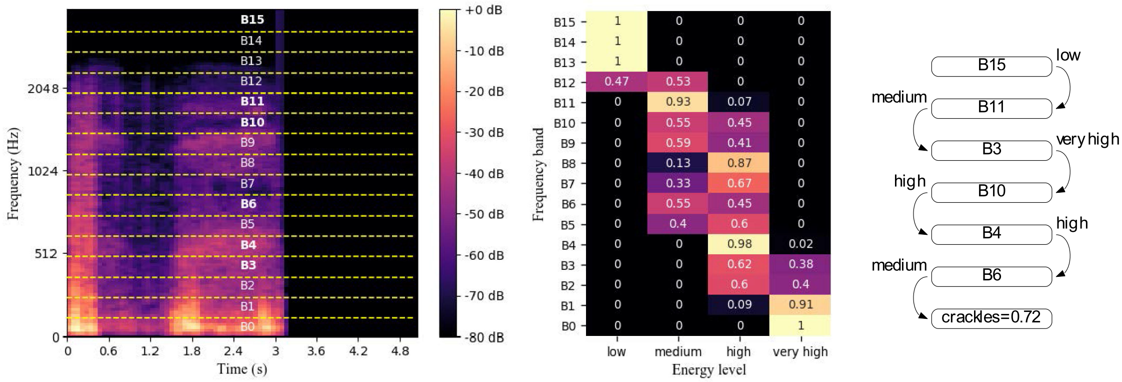
$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \eta\hat{\Omega}(\boldsymbol{\theta}) \qquad (14)$$

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

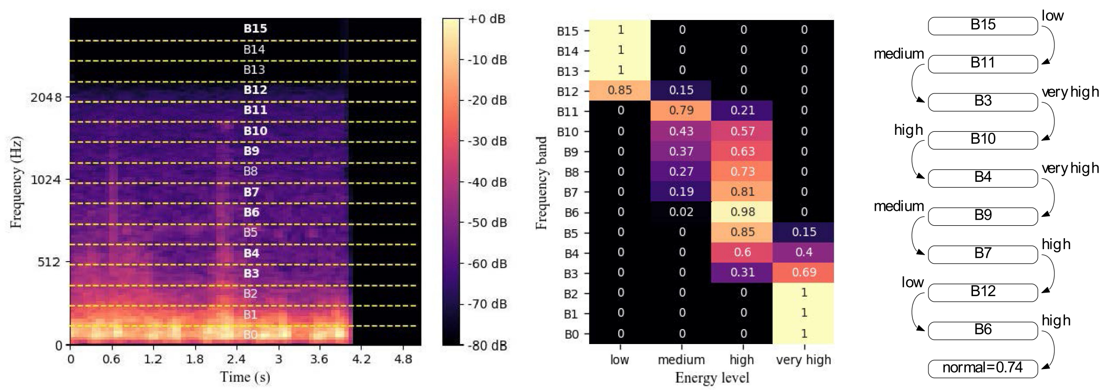
V. Experiments

A. Dataset

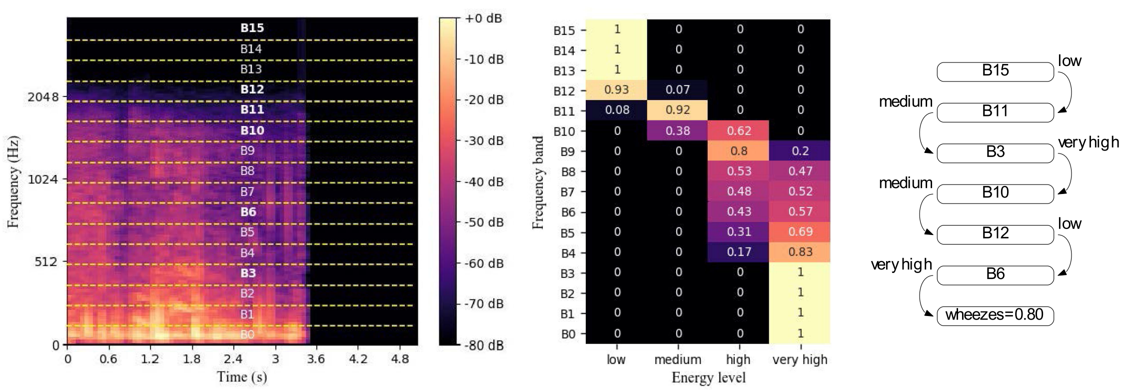
The dataset we use is the Respiratory Sound database created by two research teams in Portugal and Greece [41]. The database consists of a total of 5.5 h of recordings containing 6898



(a)



(b)



(c)

Fig. 5. (a) Model prediction as crackles. (b) Model prediction as normal. (c) Model prediction as wheezes. The three pictures in each row represent their Mel spectrogram, fuzzy energy level, and decision path, respectively. We mark the different feature values with rectangular boxes on the Mel spectrograms. The values of different frequency bands denote the fuzzy membership degree at low, medium, high, and very high energy, respectively.

respiratory cycles, of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes from eight types of subjects. The respiratory sound audio was 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 is shown in Table I. The network contains convolutional (Conv) layers, rectified linear units (ReLU), max pooling (MP) layers, fully connected (FC), and softmax layer. 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

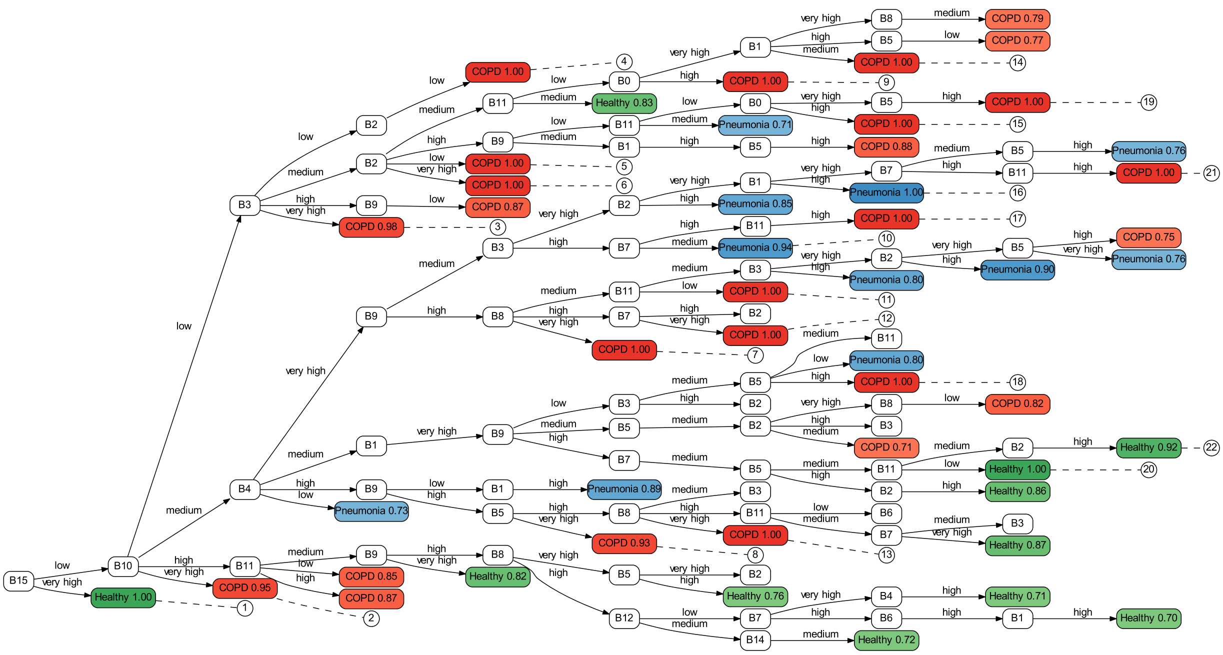


Fig. 6. Fuzzy decision tree of lung disease classification.

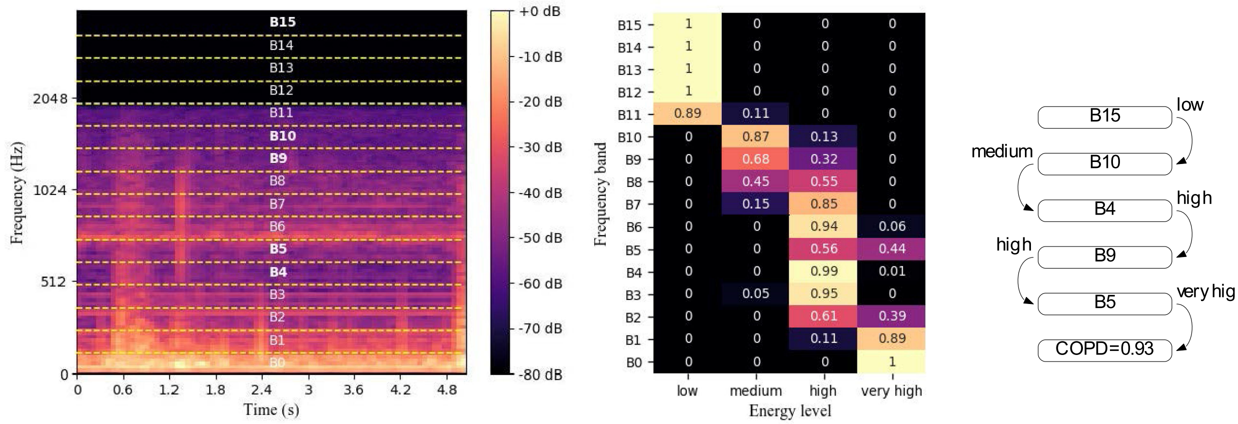
1 $Healthy(1.00) \leftarrow B_{15}(\text{very high})$
2 $COPD(0.95) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{very high})$
3 $COPD(0.98) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_3(\text{very high})$
4 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{medium}) \wedge B_3(\text{low}) \wedge B_4(\text{medium}) \wedge B_2(\text{low})$
5 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{low})$
6 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_{11}(\text{medium}) \wedge B_2(\text{very high})$
7 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{low}) \wedge B_8(\text{very high})$
8 $COPD(0.93) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_5(\text{very high})$
9 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{medium}) \wedge B_9(\text{medium}) \wedge B_9(\text{low}) \wedge B_0(\text{high})$
10 $Pneumonia(0.94) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{high}) \wedge B_9(\text{low}) \wedge B_7(\text{medium})$
11 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{medium}) \wedge B_9(\text{medium}) \wedge B_{11}(\text{low})$
12 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_7(\text{very high})$
13 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{medium}) \wedge B_9(\text{medium}) \wedge B_8(\text{very high})$
14 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{high}) \wedge B_9(\text{low}) \wedge B_8(\text{high}) \wedge B_7(\text{high}) \wedge B_7(\text{medium}) \wedge B_1(\text{medium})$
15 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{high}) \wedge B_9(\text{low}) \wedge B_0(\text{high})$
16 $Pneumonia(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{medium}) \wedge B_9(\text{high}) \wedge B_{11}(\text{low}) \wedge B_7(\text{high}) \wedge B_5(\text{very high}) \wedge B_1(\text{high})$
17 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_1(\text{high}) \wedge B_{11}(\text{high})$
18 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{high}) \wedge B_3(\text{medium}) \wedge B_4(\text{high}) \wedge B_9(\text{high}) \wedge B_9(\text{low}) \wedge B_5(\text{high})$
19 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_5(\text{high}) \wedge B_3(\text{high}) \wedge B_{12}(\text{medium}) \wedge B_3(\text{very high}) \wedge B_5(\text{high})$
20 $Healthy(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_5(\text{high}) \wedge B_3(\text{high}) \wedge B_{12}(\text{medium}) \wedge B_3(\text{very high}) \wedge B_{11}(\text{low})$
21 $COPD(1.00) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_5(\text{high}) \wedge B_3(\text{medium}) \wedge B_{12}(\text{low}) \wedge B_5(\text{high}) \wedge B_{11}(\text{high})$
22 $Healthy(0.92) \leftarrow B_{15}(\text{low}) \wedge B_{10}(\text{low}) \wedge B_2(\text{high}) \wedge B_1(\text{very high}) \wedge B_5(\text{high}) \wedge B_3(\text{high}) \wedge B_{12}(\text{medium}) \wedge B_3(\text{very high}) \wedge B_{11}(\text{medium}) \wedge B_2(\text{high})$

Fig. 7. Decision rule for lung disease classification. The rule number here corresponds to the branch number in Fig. 6.

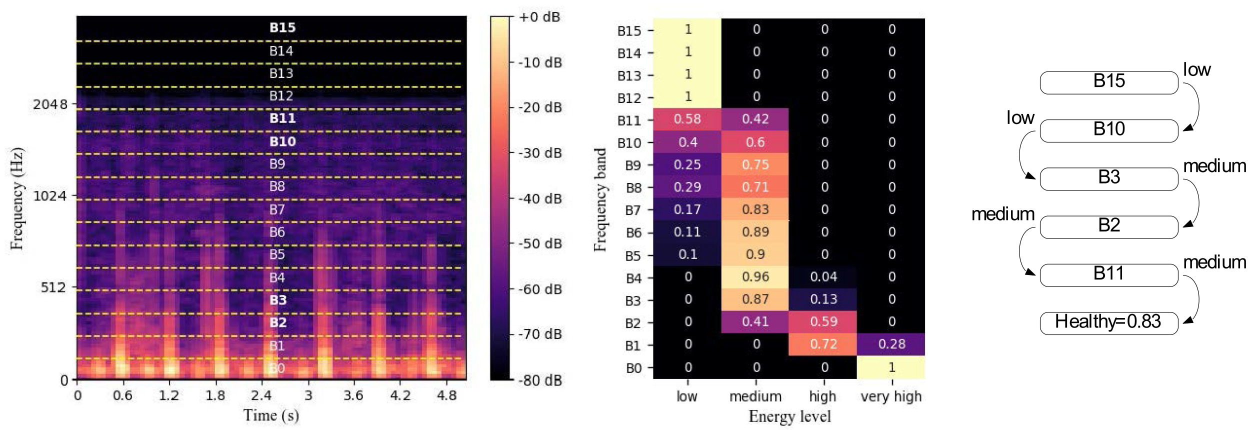
database. In experiment 2, we classify between three categories, namely, COPD, healthy, and pneumonia. During data preprocessing, first, the lung sound recordings will be sampled by the target sampling rate of 8000 Hz and transform into mono audio data.

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 128×79 for model input. We use the learning rate of 0.001, the batch size of 128, 100 epochs, λ of 0.5, T of 2.0, and η of 1000. For this experiment, we divided the training set and the test set using the fivefold cross-validation method. C of the teacher model that is used for the classification of M categories and C of the student model are both set to 3. In all experiments, the training set will be augmented by changing dynamic range and adding distribution noise.

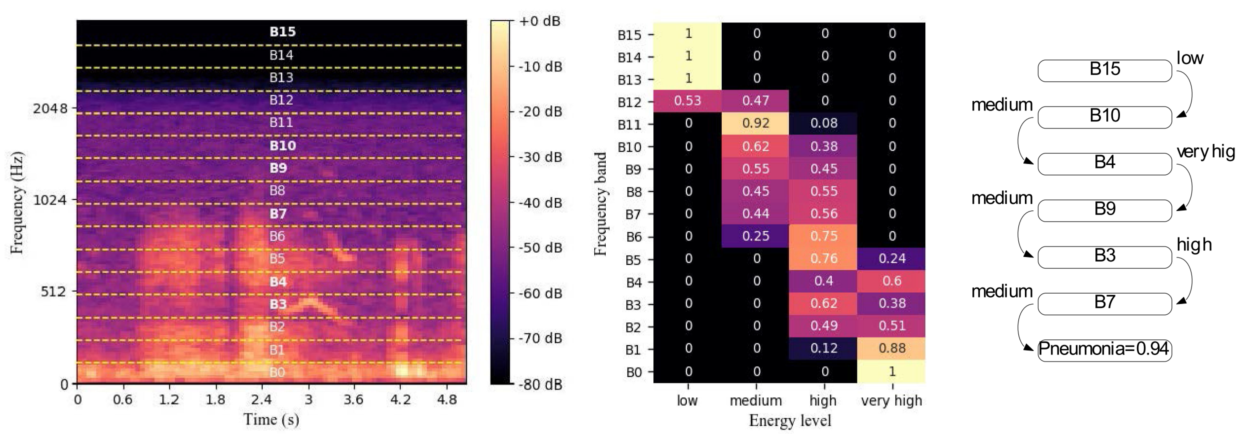
For each time–frequency result, we will divide it into 16 parts in the frequency domain; each part represents a frequency band, as shown in Table II, and then calculate the mean value of each part to get 16 features ($B_0, \ldots, B_{15}$) as input of the decision tree. Then, each feature will be fuzzified into four fuzzy sets,



(a)



(b)



(c)

Fig. 8. (a) Model prediction as COPD. (b) Model prediction as healthy. (c) Model prediction as pneumonia.

indicating low, medium, high, and very high energy. In addition, we apply the self-organizing map algorithm [43] to determine four clustering centers as the parameters of the membership function. The clustering center 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.

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.

D. Evaluation Metrics

The evaluation uses the metrics of sensitivity (SE), specificity (SP), and average score (AS) as the authors did in [15] and [44]. In experiment 1, for the confusion matrix as shown in Table III, $N_{(j)}^{(i)}, i, j \in \{c, n, w\}$, denote the number of classification results

Algorithm 3: Summary of Our Approach.

Input: data $\{x_n^{\mathrm{T}_i}, y_n^{\mathrm{T}_i}\}_{n=1}^{N^{\mathrm{T}_i}}$ for teacher T_i,
data $\{x_n^{\mathrm{S}}, y_n^{\mathrm{S}}\}_{n=1}^{N}$ for student model S
Output: The student model S and fuzzy decision tree $\mathcal{T}$

1: Initialize teachers parameters, student parameters $\boldsymbol{\theta}$, MLP parameters $\boldsymbol{\varphi}$, set $l = 0$.
2: $\mathrm{T}_i \leftarrow$ Construct teacher model with (3).
3: $q_j^{(i)} \leftarrow$ Calculate soft prediction by T_i with (4).
4: $\pi_j(.) \leftarrow$ Transform $q_j^{(i)}$ into soft label with (5) and (6).
5: **for** $l = l + 1$ **until** $l <$ MAX-ITERATE **do**
6: $\mathcal{L}(\boldsymbol{\theta}) \leftarrow$ Get loss of ensemble distillation with (8).
7: Construct the student model S and get prediction function $f(., \boldsymbol{\theta})$ with (14).
8: $\hat{y}_n^{\mathrm{S}} \leftarrow$ Get the prediction of student by $f(x_n^{\mathrm{S}}, \boldsymbol{\theta})$.
9: $x_n^D \leftarrow$ Get the data of $\mathcal{T}$ by mapping function ∇ and membership function (9)–(11).
10: Fit a decision tree $\mathcal{T}$ on $(\{x_n^{\mathcal{D}}, \hat{y}_n^{\mathrm{S}}\})$ by Algorithm 1.
11: $\mathcal{T} \leftarrow$ Prune the trained decision tree by Algorithm 2.
12: $\Omega(\boldsymbol{\theta}) = \frac{1}{N}\sum_n \mathrm{apl}(\mathcal{T}, x_n^{\mathcal{D}})$.
13: Construct the surrogate MLP and approximate $\Omega(\boldsymbol{\theta})$ by $\hat{\Omega}(\boldsymbol{\theta})$ with (13).
14: **end for**

TABLE I
MODEL NETWORK ARCHITECTURE

Input	Layer	Output
Teacher model		
$1 \times 128 \times 79$	Conv$(5*5)$-ReLU-MP$(2*2)$	$16 \times 63 \times 38$
$16 \times 63 \times 38$	Conv$(3*3)$-ReLU-MP$(2*2)$	$32 \times 31 \times 19$
$32 \times 31 \times 19$	Conv$(3*3)$-ReLU-MP$(2*2)$	$64 \times 15 \times 9$
$64 \times 15 \times 9$	FC-Softmax	C
Student model		
$1 \times 128 \times 79$	Conv$(5*5)$-ReLU-MP$(2*2)$	$16 \times 63 \times 38$
$16 \times 63 \times 38$	Conv$(3*3)$-ReLU-MP$(2*2)$	$16 \times 31 \times 19$
$16 \times 31 \times 19$	Conv$(3*3)$-ReLU-MP$(2*2)$	$32 \times 15 \times 9$
$32 \times 15 \times 9$	FC-Softmax	C

TABLE II
FREQUENCY BANDS AND SYMBOLS

Symbols	Frequency(Hz)	Symbols	Frequency(Hz)
B_0	$[0, 146]$	B_8	$[1194, 1389]$
B_1	$[146, 293]$	B_9	$[1389, 1615]$
B_2	$[293, 439]$	B_{10}	$[1615, 1879]$
B_3	$[439, 586]$	B_{11}	$[1879, 2185]$
B_4	$[586, 732]$	B_{12}	$[2185, 2542]$
B_5	$[732, 879]$	B_{13}	$[2542, 2956]$
B_6	$[879, 1026]$	B_{14}	$[2956, 3439]$
B_7	$[1026, 1194]$	B_{15}	$[3439, 4000]$

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 as

$$\mathrm{SE} = \left(N_{(c)}^{(c)} + N_{(w)}^{(w)}\right) / \sum_{i \in \{c,n,w\}} \left(N_{(c)}^{(i)} + N_{(w)}^{(i)}\right)$$

$$\mathrm{SP} = N_{(n)}^{(n)} / \sum_{i \in \{c,n,w\}} N_{(n)}^{(i)}. \tag{15}$$

TABLE III
CONFUSION MATRIX

	Crackles	Normal	Wheezes
Respiratory sounds classification			
Crackles	$N_{(c)}^{(c)}$	$N_{(c)}^{(n)}$	$N_{(c)}^{(w)}$
Normal	$N_{(n)}^{(c)}$	$N_{(n)}^{(n)}$	$N_{(n)}^{(w)}$
Wheezes	$N_{(w)}^{(c)}$	$N_{(w)}^{(n)}$	$N_{(w)}^{(w)}$
Lung diseases classification			
	COPD	Healthy	Pneumonia
COPD	$N_{(co)}^{(co)}$	$N_{(co)}^{(h)}$	$N_{(co)}^{(p)}$
Healthy	$N_{(h)}^{(co)}$	$N_{(h)}^{(h)}$	$N_{(h)}^{(p)}$
Pneumonia	$N_{(p)}^{(co)}$	$N_{(p)}^{(h)}$	$N_{(p)}^{(p)}$

TABLE IV
COMPARISON RESULTS OF CLASSIFYING CYCLES

Method	SP	SE	AS	Assimilated
Aykanat [12]	**0.82 ± 0.03**	0.52 ± 0.02	0.67 ± 0.02	Y
Bardou [14]	0.80 ± 0.02	0.56 ± 0.04	0.68 ± 0.02	Y
Tariq [13]	0.76 ± 0.04	0.53 ± 0.03	0.65 ± 0.01	Y
Teacher	**0.82 ± 0.02**	**0.61 ± 0.01**	**0.72 ± 0.01**	Y
No Distill	0.78 ± 0.01	0.56 ± 0.01	0.67 ± 0.01	Y
Distill	0.80 ± 0.02	0.58 ± 0.04	0.69 ± 0.02	**N**
Student	0.80 ± 0.02	0.60 ± 0.01	0.70 ± 0.01	**N**

The bold entities indicate the best achieved performance among the comparisons.

In experiment 2, for the confusion matrix as shown in Table III, co, h, and p denote COPD, healthy, and pneumonia, respectively. SE and SP can be calculated as

$$\mathrm{SE} = \left(N_{(co)}^{(co)} + N_{(p)}^{(p)}\right) / \sum_{i \in \{co,h,p\}} \left(N_{(co)}^{(i)} + N_{(p)}^{(i)}\right)$$

$$\mathrm{SP} = N_{(h)}^{(h)} / \sum_{i \in \{co,h,p\}} N_{(h)}^{(i)}. \tag{16}$$

AS can be calculated as

$$\mathrm{AS} = (\mathrm{SE} + \mathrm{SP})/2. \tag{17}$$

In addition, we will also indicate if the method requires the data to be *assimilated* before training.

E. Experiment 1: Respiratory Sound Classification

In experiment 1, since 97% of the respiratory cycles are within 5 s, we extract the sound clips for the fixed duration of 5 s. 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 evaluates the performance of difference methods for adventitious sounds. In Table IV, models "Aykanat," "Bardou," and "Tariq" denote our three comparison networks, "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

TABLE V
COMPARISON RESULTS OF CLASSIFYING DISEASES

Method	SP	SE	AS	Assimilated
Aykanat [12]	0.85 ± 0.04	0.96 ± 0.01	0.91 ± 0.02	Y
Bardou [14]	0.91 ± 0.04	0.97 ± 0.01	0.94 ± 0.02	Y
Tariq [13]	0.85 ± 0.01	0.97 ± 0.01	0.91 ± 0.00	Y
Teacher	**0.92 ± 0.02**	**0.98 ± 0.00**	**0.95 ± 0.01**	Y
No Distill	0.85 ± 0.01	0.97 ± 0.01	0.91 ± 0.01	Y
Distill	0.88 ± 0.03	0.97 ± 0.00	0.93 ± 0.02	**N**
Student	0.89 ± 0.03	0.97 ± 0.01	0.93 ± 0.01	**N**

The bold entities indicate the best achieved performance among the comparisons.

"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 "Aykanat," "Bardou," and "Tariq" by 3%, 2%, and 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 sample of crackles, normal, and wheezes are shown in Fig. 5. When the student model predicts a sample, we can obtain the Mel spectrogram, 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 Disease Classification

In this experiment, since the duration of most respiratory cycles is about 20 s, we will extract the sound clips for the fixed duration of 5 s 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 models "Aykanat" and "Tariq" in terms of AS. More importantly, the performance of the student model is only 0.01 and 0.02 lower than the model "Bardou" and the 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. 6. 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. 7. The three types of test sample of COPD, healthy, and pneumonia are shown in Fig. 8. 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 article proposed a novel explainable CNN framework based on fuzzy decision tree regularization for respiratory sound 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 nonfixed 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 the 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

[1] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "Evaluation of features for classification of wheezes and normal respiratory sounds," *PLoS One*, vol. 14, no. 3, pp. 1–21, 2019.

[2] Y. Cheng, Y. Lin, K. Chiang, and V. S. Tseng, "Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: A case study on chronic obstructive pulmonary disease," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 303–311, Mar. 2017.

[3] World Health Organization, "COVID-19 weekly epidemiological update," 69th ed., Dec. 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[4] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Ann. Thoracic Med.*, vol. 10, pp. 158–168, 2015.

[5] A. Bohadana, G. Izbicki, and S. Kraman, "Fundamentals of lung auscultation," *New England J. Med.*, vol. 370, no. 8, pp. 744–751, 2014.

[6] A. Rizal, R. Hidayat, and H. A. Nugroho, "Signal domain in respiratory sound analysis: Methods, application and future development," *J. Comput. Sci.*, vol. 11, no. 10, pp. 1005–1016, 2015.

[7] A. Kandaswamy, C. Kumar, R. Ramanathan, S. Jayaraman, and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," *Comput. Biol. Med.*, vol. 34, no. 6, pp. 523–537, 2004.

[8] S. A. Taplidou and L. J. Hadjileontiadis, "Wheeze detection based on time-frequency analysis of breath sounds," *Comput. Biol. Med.*, vol. 37, no. 8, pp. 1073–1083, 2007.

[9] S. Reichert, R. Gass, C. Brandt, and E. Andres, "Analysis of respiratory sounds: State of the art," *Clin. Med. Circulatory, Respir. Pulmonary Med.*, vol. 2, pp. 45–58, 2008.

[10] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: A systematic review," *Biocybern. Biomed. Eng.*, vol. 33, no. 3, pp. 129–135, 2013.

[11] R. X. A. Pramono, S. A. Bowyer, and E. Rodríguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PLoS One*, vol. 12, no. 5, 2017, Art. no. e0177926.

[12] M. Aykanat, Ö. Kiliç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP J. Image Video Process.*, vol. 2017, pp. 1–9, 2017.

[13] Z. Tariq, S. K. Shah, and Y. Lee, "Lung disease classification using deep convolutional neural network," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 732–735.

[14] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artif. Intell. Med.*, vol. 88, pp. 58–69, 2018.

[15] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst.*, 2019, pp. 50–55.

[16] J. Vaughan, A. Sudjianto, E. Brahimi, J. J. Chen, and V. N. Nair, "Explainable neural networks based on additive index models," 2018, *arXiv:1806.01933*.

[17] Q. Zhang, Y. N. Wu, and S. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 8827–8836.

[18] M. Yeganejou, S. Dick, and J. Miller, "Interpretable deep convolutional fuzzy classifier," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1407–1419, Jul. 2020.

[19] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 3504–3512.

[20] R. Chen, H. Chen, G. Huang, J. Ren, and Q. Zhang, "Explaining neural networks semantically and quantitatively," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 9186–9195.

[21] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1670–1678.

[22] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, "Application of semi-supervised deep learning to lung sound analysis," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 804–807.

[23] S. Das, D. Guha, and B. Dutta, "Medical diagnosis with the aid of using fuzzy logic and intuitionistic fuzzy logic," *Appl. Intell.*, vol. 45, no. 3, pp. 850–867, 2016.

[24] G. Capizzi, G. L. Sciuto, C. Napoli, D. Polap, and M. Wozniak, "Small lung nodules detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 6, pp. 1178–1189, Jun. 2020.

[25] M. Umanol *et al.*, "Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems," in *Proc. IEEE 3rd Int. Fuzzy Syst. Conf.*, 1994, vol. 3, pp. 2113–2118.

[26] N. Sengupta, M. Sahidullah, and G. Saha, "Lung sound classification using cepstral-based statistical features," *Comput. Biol. Med.*, vol. 75, pp. 118–129, 2016.

[27] S. Datta, A. D. Choudhury, P. Deshpande, S. Bhattacharya, and A. Pal, "Automated lung sound analysis for detecting pulmonary abnormalities," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 4594–4598.

[28] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[29] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, San Francisco, CA, USA, 2016, pp. 3439–3443.

[30] M. Wu, C. Chiu, and K. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 2202–2206.

[31] B. Shin, H. Yang, and J. D. Choi, "The pupil has become the master: Teacher-student model-based word embedding distillation with ensemble learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, 2019, pp. 3439–3445.

[32] S. Du *et al.*, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12345–12355.

[33] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 1285–1294.

[34] Q. Zhang, Y. Yang, Y. Liu, Y. N. Wu, and S. Zhu, "Unsupervised learning of neural networks to explain neural networks," 2018, *arXiv: 1805.07468*.

[35] Z. Che, S. Purushotham, R. G. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. Amer. Med. Inform. Assoc. Annu. Symp.*, Chicago, IL, USA, vol. 2016, 2016, pp. 371–380.

[36] N. Frosst and G. E. Hinton, "Distilling a neural network into a soft decision tree," 2017, *arXiv:1711.09784*.

[37] S. Tan, R. Caruana, G. Hooker, P. L. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," 2018, *arXiv:1801.08640*.

[38] Q. Zhang, Y. Yang, Y. N. Wu, and S.-C. Zhu, "Interpreting CNNs via decision trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2019, pp. 6254–6263.

[39] A. Segatori, F. Marcelloni, and W. Pedrycz, "On distributed fuzzy decision trees for Big Data," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 174–192, Feb. 2018.

[40] Z. Hu, X. Ma, Z. Liu, E. H. Hovy, and E. P. Xing, "Harnessing deep neural networks with logic rules," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 2410–2420.

[41] B. M. Rocha *et al.*, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health*. Singapore: Springer, 2018, pp. 33–37.

[42] B. McFee *et al.*, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–24.

[43] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[44] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, and Y. Kahya, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiol. Meas.*, vol. 40, 2019, Art. no. 035001.

Jianqiang Li (Member, IEEE) received the B.S. and Ph.D. degrees in automation from the South China University of Technology, Guangzhou, China, in 2003 and 2008, respectively.

He is currently 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 in information technology from the University of South China, Hengyang, China, in 2019. He is currently working toward the master's degree in computer science with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His current research interests include intelligent healthcare systems.

Jie Chen received the B.Sc. degree in electrical engineering from Taizhou University, Taizhou, China, in 2006, the M.Sc. degree in computer science from Zhejiang University, Hangzhou, China, in 2008, and the Ph.D. degree in computer science from the National University of Singapore (NUS), Singapore, in 2013.

He was a Postdoctoral Associate with Singapore–MIT Alliance for Research and Technology from 2013 to 2018. He was a Research-Track Associate Professor with Shenzhen University from 2018 to 2021. Since 2021, he has been an Assistant Professor with Shenzhen University, Shenzhen, China. His research interests include machine learning, multiagent systems, robotics, intelligent transportation systems, and intelligent healthcare systems.

Dr. Chen is the recipient of the Dean's Graduate Research Excellence Award from NUS in 2013. He was a Program Committee Member of International Joint Conference on Artificial Intelligence, AAAI Conference on Artificial Intelligence, and International Conference on Robotics and Automation.

Heng Zhang received the master's degree in the area of mechanical ventilation from the Capital Medical University, Beijing, China, in 2012.

She is currently an Attending Physician with the Department of Respiratory and Critical Care Medicine, Shenzhen People's Hospital, Shenzhen, China. She is involved in the frontline clinical work of respiratory and critical medicine. She is with the 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, and Shenzhen Key Laboratory of Respiratory Diseases, Shenzhen. Her current research interests include infectious diseases of the respiratory system.

Yuyan Dai received the B.S. degree in information technology from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020. She is currently working toward the master's degree in computer science 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 received the M.D. degree in airway remodeling in asthma from Jinan University, Guangzhou, China, in 2016.

She is the Deputy Director of the Shenzhen Institute of Respiratory Diseases and the Director of the Department of Diagnostics, School of Medicine, Southern University of Science and Technology, Shenzhen, China. She is currently with the 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, and Shenzhen Key Laboratory of Respiratory Diseases, Shenzhen.

Li Wang received the B.Sc. and Ph.D. degrees in biomedical engineering from Zhejiang University, Hangzhou, China, in 2008 and 2017, respectively.

He was a Senior Engineer with ZICT Technology Company, Ltd., Shenzhen, China, from 2017 to 2019. He has been a Postdoctoral Associate with Shenzhen University, Shenzhen, since 2019. His research interests include medical informatics, regional healthcare systems, machine learning, and artificial intelligence.

Asoke K. Nandi (Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College), Cambridge, U.K., in 1979.

He held academic positions in several universities, including Oxford University, Oxford, U.K.; Imperial College London, London, U.K.; University of Strathclyde, Glasgow, U.K.; and University of Liverpool, Liverpool, U.K., as well as Finland Distinguished Professorship with Jyväskylä University, Jyväskylä, Finland. In 2013, he joined Brunel University London, Uxbridge, U.K., as the Chair and Head of Electronic and Computer Engineering Department. He is a Distinguished Visiting Professor with Shenzhen University, Shenzhen, China, and an Adjunct Professor with the University of Calgary, Calgary, AB, Canada. He has authored or coauthored more than 600 research articles, with a Google Scholar H-index of 80.

Dr. Nandi is a Fellow of the Royal Academy of Engineering. He is also a Fellow of seven other institutions, including the Institution of Engineering and Technology.