# Deep auscultation with demographic data: detecting respiratory anomalies using convolutional neural networks and autoencoders

Mohan Xu[1] and Lena Wiese[1,2]

[1] Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany,
mohan.xu@item.fraunhofer.de,
lena.wiese@item.fraunhofer.de,
[2] Institute of Computer Science, Goethe University Frankfurt, Frankfurt a. M., Germany

**Abstract.** Digital auscultation, integrating digital signal processing and machine learning algorithms, has garnered significant attention in the field of automated disease diagnosis due to its simplicity, speed, and non-invasiveness. In our research, we introduce a multi-modal hybrid model that merges a wavegram logmel CNN, initially trained on the extensive audio dataset, with demographic data processed through an autoencoder. Additionally, we analyze the impact of two encoding methods on demographic information and integrating different numbers of snapshot models in the snapshot ensemble method on the model scores. The experimental results demonstrate that both the wavegram logmel CNN and multi-modal hybrid model with two encoding methods exhibit improvements of 2.1%, 3.3% and 2.8%, respectively, compared to a single model when utilizing four snapshot models. Through a 10-fold cross validation on the ICBHI dataset, the multi-modal hybrid model with one-hot encoding achieves a remarkable model score of 82.5% in the four-classification task, surpassing previous research outcomes and achieving a 1.2% enhancement over the wavegram logmel CNN model, which solely employs respiratory cycles as input.

**Keywords:** ICBHI dataset, neural networks, respiratory sounds classification, multi-modal

## 1 Introduction

The World Health Organization (WHO) [27] reports that over 8 million deaths in 2019 were attributable to respiratory diseases. The distinction between normal and abnormal respiratory sounds can provide valuable information for the diagnosis of respiratory diseases [22]. Among the abnormal sounds, crackles and wheezes are notably prevalent. Crackles, characterized as explosive, short-lived discontinuous sounds, are often linked with conditions like Interstitial lung fibrosis, Pneumonia, and Congestive heart failure. Wheezes, which are high-pitched continuous sounds, occur in airflow-restricted airways and are associated with diseases including Asthma and COPD [26].

This article proposes a multi-modal hybrid model aimed at classifying respiratory sounds as normal, crackle, wheeze, and both on the ICBHI dataset. The model is based on wavegram logmel CNN [13] with an additional autoencoder structure to extract features from the demographic information of the subjects. Based on the dataset's missing demographic information, we implemented one-hot and dummy encoding methods into our model design. The findings demonstrated that the one-hot encoding method yielded superior results. Utilizing the snapshot ensemble method to evaluate the integrated models' efficacy, we observe performance improvements in both the wavegram logmel CNN and multi-modal hybrid models to varying degrees, particularly when using 4 snapshot models. Our results surpass previous work under the same dataset partitioning condition, with the multi-modal hybrid model achieving a classification accuracy of up to 82.5%.

## 2 Related work

### 2.1 Multi-modal data in deep learning

The rapid acceleration in data collection in recent years has introduced multi-modal data in structured, semi-structured and unstructured formats become a new topic in the field of deep learning. Multi-modal data, as the name suggests, is data consisting of multiple modalities, each covering a partial description of the same thing of interest. Information fusion of multi-modal data can help to better understand the thing of interest in the presence of information deficit in any of the modalities [7].

The multi-modal autoencoder proposed by Ngiam et al. [18] learns the correlation between modalities, and some of the data that have only a single modality but complement the rest by adding zero values are likewise fed into the model to learn features. [15] features an image CNN for image representation, a matching CNN for encoding images and words into a joint representation, and multi-layer perceptron (MLP) for scoring image-sentence matches, enabling image retrieval and sorting via natural language queries. [25] applies multi-modal data measured by wearable devices to recurrent neural networks to help improve people's sleep characteristics.

In addition to the work mentioned above, several studies have used demographic information of subjects collected in the clinic to participate in the construction of multi-modal neural network structures. In classifying skin lesions, [28] not only trained the datasets acquired from two different imaging modalities separately in Resnet50, but also added the demographic data of the corresponding patients to the fully connected layer and fused it with the features learned from the other two modalities. [3] trained and extracted features from the four modality data in different network structures or settings, respectively, and predicted single cancer and pancancer overall survival by calculating the similarity loss for each modality pair. The features extracted from the Chest X-ray images by pre-trained CNN and demographic data were concat and continued to fed into the hidden layer in [9], and the output was used to determine the presence of TB in chest radiographs. For incomplete multi-modal data, [4] introduced modality dropout to randomly discard the modality being trained and use mean vector method to average only the available modality data during feature fusion.

### 2.2 Related studies using the ICBHI dataset

The ICBHI 2017 challenge dataset [23] is extensively employed in the field of digital auscultation as a public respiratory sound dataset for the comparison of deep learning algorithms. Based on the four symptoms included in the respiratory cycle (normal, wheeze, crackle and both (wheeze and crackle)), current deep learning classification models being applied to the data set are convolutional neural networks (CNN), recurrent neural networks (RNN) and hybrid models.
The spectrograms and scalograms obtained by [17] after preprocessing the respiratory sounds were used as input to CNN, respectively, and the features extracted by the network training were concatenated and input into the fully connected layer. [14] defined the CNN and added an attention mechanism to it so that the network training is mainly focused on some key features. A series of new techniques such as fine-tuning on devices, augmentation through concatenation and blank region clipping were proposed in [6] to preprocess the audio. For the class imbalance problem in the dataset, [11] used conditional GAN to increase the diversity of the dataset. The multi-stage CNN-RNN [1] model extracts abstract feature representations from mel spectrogram by CNN and learns temporal relations in RNN.
Despite the extensive usage of the ICBHI dataset, there is a lack of research involving the inclusion of demographic information as one of the modalities in the neural network workflow for classifying respiratory sounds.

## 3 Materials and Methods

### 3.1 ICBHI 2017 database

The ICBHI Scientific Challenge database (as briefly described in Section 2.2) is a large database of labelled respiratory sounds [23]; it has a total duration of 5.5 hours and contains 920 audio recordings from 126 subjects. The length of each recording ranges between 10s and 90s, and they can be segmented into various respiratory cycles based on the official annotation file that is provided. Considering the symptom categories contained in the respiratory cycles, 6898 respiratory cycles in the ICBHI dataset were labeled by the respiratory experts into four categories: normal, wheeze, crackle and both (wheeze and crackle). Each category contains 3642, 886, 1864 and 506 respiratory cycles respectively.

### 3.2 Demographic data

In addition to the 920 audio recordings and corresponding annotated files, the ICBHI Scientific Challenge database also provides demographic information for each subject in five areas: age, gender, BMI, child weight and child height. There are three cases of missing demographic information: (i) the subject does not provide the information; (ii) the subject is a child and therefore the corresponding BMI information is not available; (iii) the subject is an adult and therefore child weight and child height are not available. To leverage the demographic details of each respiratory cycle in the neural network's workflow effectively, we analyzed the data distribution of respiratory cycles on four categories: normal, crackle, wheeze and both, in terms of age, gender and BMI.

Following the grouping criteria outlined in Table 1, we organized subjects' respiratory cycles into age, gender, or BMI groups. "Min" and "max" in Table 1 represent the minimum and maximum values for age and BMI among the subjects. Our study involved the exploration of two encoding methods. One-hot encoding employs an N-bit state register to encode N states. Notably, this method accounts for missing demographic information by treating it as a distinct state, which is also included in the encoding. For instance, subjects' ages are encoded as [1 0 0 0], [0 1 0 0], [0 0 1 0], [0 0 1 0], and [0 0 0 1] for the age intervals [min, 18), [18, 60), [60, max], and missing values, respectively. Likewise, missing information for gender and BMI is encoded as [0 0 1] and [0 0 0 0 1]. Dummy encoding shares a similar concept to one-hot encoding. However, it excludes the states with missing demographic information from the encoding. Based on information from the ICBHI database, subjects have the option to provide demographic information anonymously to avoid privacy implication.

**Table 1.** For the four categories of normal, crackle, wheeze and both, the data distribution of respiratory cycles on different age groups, gender groups and BMI groups.

| Demographic variables | | Age | | | Gender | | BMI | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Grouping criteria | | [min,18) | [18,60) | [60,max] | Female | Male | [min,18.5) | [18.5,25) | [25,30) | [30,max] |
| One-hot encoding | | [1 0 0 0] | [0 1 0 0] | [0 0 1 0] | [1 0 0] | [0 1 0] | [1 0 0 0 0] | [0 1 0 0 0] | [0 0 1 0 0] | [0 0 0 1 0] |
| Dummy encoding | | [1 0 0] | [0 1 0] | [0 0 1] | [1 0] | [0 1] | [1 0 0 0] | [0 1 0 0] | [0 0 1 0] | [0 0 0 1] |
| Normal(3642) | Mean±std | 3.55±3.79 | 55.32±6.86 | 70.91±6.31 | - | | 16.90±0.30 | 22.75±1.70 | 27.85±1.39 | 33.84±5.51 |
| | Number | 642 | 500 | 2471 | 1239 | 2374 | 135 | 767 | 1393 | 652 |
| Crackle(1864) | Mean±std | 4.55±4.29 | 54.70±4.39 | 73.53±7.80 | - | | 17.02±0.20 | 22.12±1.66 | 27.44±1.26 | 32.79±5.17 |
| | Number | 65 | 365 | 1415 | 760 | 1085 | 356 | 450 | 515 | 446 |
| Wheeze(886) | Mean±std | 1.98±2.14 | 57.35±0.91 | 72.23±6.62 | - | | 16.81±0.30 | 21.46±1.99 | 27.75±1.23 | 32.88±2.63 |
| | Number | 75 | 54 | 744 | 197 | 676 | 117 | 136 | 437 | 100 |
| Both(506) | Mean±std | 3.06±3.52 | 56.90±0.82 | 75.31±7.98 | - | | 16.78±0.30 | 22.88±0.51 | 27.99±1.16 | 32.81±1.49 |
| | Number | 6 | 29 | 471 | 156 | 350 | 123 | 142 | 191 | 44 |

### 3.3 Proposed method

We now describe our multi-modal hybrid model workflow (see Figure 1). The ICBHI database was divided into 6898 respiratory cycles according to the official annotation file (Table 3). Next, the respiratory cycles and the corresponding demographic data were converted into data structures suitable for neural network inputs by different preprocessing methods. Different data categories were fed into autoencoder and wavegram logmel cnn for training. The extracted features are mapped to the corresponding categories by a concat operation.

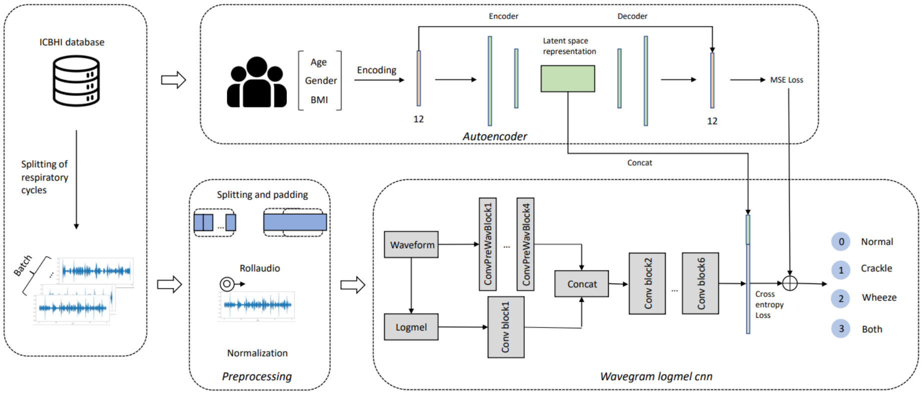
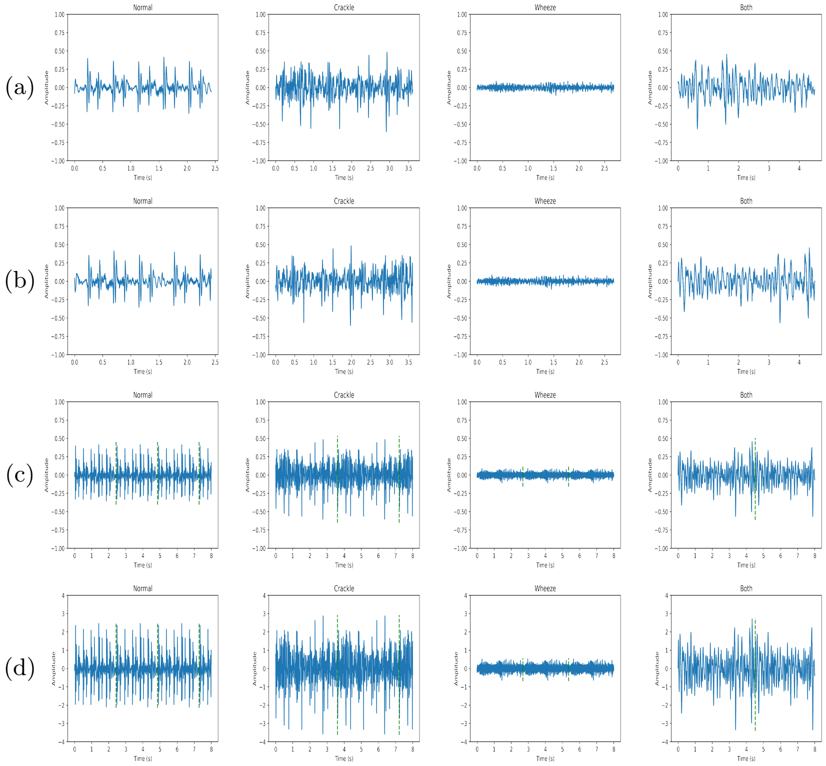


**Fig. 1.** Proposed multi-modal hybrid model workflow

**Preprocessing** To facilitate the pre-processing of the continuous audio signal in the workflow of Figure 1, the librosa library [16] reads the 6898 respiratory cycles divided according to Table 3 and generates discrete audio signals in JSON format at a sampling rate of 16 kHz. The newly generated data contain information on the corresponding labels as well as subject number, recording index, chest position, recording mode, recording device, and cycle number. From the discrete audio signal to the data structure suitable for wavegram logmel cnn input, three pre-processing methods were used: splitting and padding, data augmentation and normalization. Table 2 presents the results from four representative respiratory cycles (normal, crackle, wheeze and both) following various preprocessing steps.

*Splitting and Padding* Previous research [13, 6] has investigated the impact of varying respiratory cycle lengths on neural network performance. It has been observed that the neural network achieves optimal results when the respiratory cycle duration is set to 8s. However, the ICBHI dataset contains respiratory cycles with diverse durations, ranging from a minimum of 0.2s to a maximum of 16.1s. Figure 1 illustrates the preprocessing module, which outlines the process of standardizing respiratory cycles to a consistent length. Two scenarios are considered: for the length of the original respiratory cycle longer than 8s, the librosa library [16] is applied to randomly retain either the initial 8s or the final 8s; for the length of the original respiratory cycle shorter than 8s, they

**Table 2.** Waveforms of original respiratory cycles are shown in (a), followed by preprocessing steps: data augmentation (b), splitting and padding (c), normalization (d).



are duplicated until reaching a length greater than 8s, followed by preprocessing based on the first scenario.

*Data augmentation* The rollaudio method implements data augmentation by randomly changing the starting position of data reading. It achieves this by generating a random index within the range of the respiratory cycle length. Subsequently, the respiratory cycle is read starting from the position corresponding to the generated index, continuing until just before reaching the initial starting position.

*Normalization* Due to the diversity of data sources, data containing different scales for features must be put on a common scale to prevent any adverse impacts on the model's performance. Z-score normalization, as a method of data normalization, converts the data to the same scale. The normalized dataset has a mean of 0 and a variance of 1. To achieve this, mean ($\mu$) and standard deviation ($\delta$) values are computed individually for each respiratory cycle and subsequently

**Table 3.** Annotation file containing start and end time of some respiratory cycles and corresponding labels

| Cycle | Start (s) | End (s) | Crackle | Wheeze | Label | Symptom |
|---|---|---|---|---|---|---|
| incomplete | 0 | 0.662 | - | - | - | - |
| 1 | 0.662 | 3.483 | 0 | 0 | 0 | normal |
| 2 | 3.483 | 5.218 | 1 | 0 | 1 | crackle |
| 3 | 5.218 | 9.072 | 0 | 1 | 2 | wheeze |
| 4 | 9.072 | 12.647 | 1 | 1 | 3 | both |
| incomplete | 12.647 | 13.251 | - | - | - | - |

averaged across ICBHI dataset to determine the global $\mu$ and $\delta$. For any respiratory cycle $x$, $x'$ is the new respiratory cycle after normalization: $x' = \frac{x-\mu}{\delta}$.

**Autoencoder** Multi-modal data has been widely used in the field of deep learning to help models perform better prediction tasks. In the use case of this work, we use two modalities, respiratory cycle and demographic data from each subject, to predict the category of respiratory cycle. An Autoencoder consists of an encoder, which learns how to interpret the input and map it to a low-dimensional representation called "latent space", and a decoder, which learns how to reconstruct the original input data from the latent space representation. To extract the information in the demographic data, the encoded demographic data is used as the input to autoencoder. By minimizing the reconstruction loss, autoencoder learns to capture the most important features of the input data in the latent space. The latent space representation is concatenated as the extracted information into the fully connected layer of wavegram logmel cnn. Table 4 shows the architecture of the autoencoder used to extract the demographic information.

**Table 4.** The proposed autoencoder architecture

| Architecture | Layers | Output (one-hot/dummy) |
|---|---|---|
| Encoder | Input layer | (b,1,12)/(b,1,9) |
| | Conv1D+BN+LeakyReLU | (b,64,10)/(b,64,7) |
| | Conv1D+BN+LeakyReLU | (b,128,8)/(b,128,5) |
| Decoder | Input layer | (b,128,8)/(b,128,5) |
| | Conv1D+BN+LeakyReLU | (b,64,10)/(b,64,7) |
| | Conv1D+Sigmoid | (b,1,12)/(b,1,9) |

**Wavegram logmel CNN** Transfer learning allows a model to transfer what it learns in one task to another related or different task, thus reducing the training time of the model in the new task and improving the generalization ability of the model. Transfer learning is particularly useful for smaller data sets. The Wavegram logmel CNN [13] in Figure 1 is loaded with pre-trained weights trained on the *Audioset* [8] and fine-tuned at fully connected layers of the model so that it can fuse features with the latent space representation from our autoencoder. The prediction results of the respiratory cycle on the four-classification task (normal, crackle, wheeze and both) were obtained after the activation function log softmax. The raw input waveforms are separately trained on two branches; the combined results are subsequently passed through five dropout layers and five blocks comprising *conv2d*, *batchNorm2d*, *relu* and *avg_pool2d*.

**Branch1:** The 1D-data, after undergoing processing by the convolutional layer and batch normalization, is then passed through three identical blocks. Each of these blocks comprises convolutional layer, batch normalization, activation function, and pooling layer, resulting in the output for next step.
**Branch2:** The input raw waveforms undergo Short-Time Fourier Transform (STFT) to generate Mel spectrograms. The branch's outcomes are derived by progressing through the convolution layer, batch normalization, activation function, and pooling layer, all of which are applied to process 2D data.

**Loss function** The loss function of the multi-modal hybrid model proposed in this paper consists of two terms: the mean squared error (MSE) loss function, which enables the output of autoencoder to maximize the reconstruction of the original input; and the cross entropy loss function, which is used in the classification task to measure the dissimilarity between the true label and the predicted probability. The overall loss becomes:

$$L_{multi} = \frac{1}{N} \sum_{i=1}^{N} [(x_i - x_i')^2 + \sum_{c=1}^{K} y_{ic} log(y_{ic}')] \tag{1}$$

$K$ is the number of categories contained in the dataset and $N$ is the total number of respiratory cycles. For a respiratory cycle $i$, $x_i$, $x_i'$ correspond to its demographic data (in one-hot encoding format) and its predicted value on autoencoder, respectively; $y_{ic}$, $y_{ic}'$ correspond to its true category and wavegram logmel CNN's predicted probability on category $c$.

### 3.4 Evaluation metrics

Our work adopted the evaluation metrics officially provided in [24] for categorizing respiratory cycles into four groups. To simplify the expression, the four categories, namely normal, crackle, wheeze, and both, are abbreviated as N, C, W, and B, respectively. Consequently, the formulas for Sensitivity (Se), Specificity (Sp), and Score (Sc) are defined as follows:

$$Se = \frac{C_{correct} + W_{correct} + B_{correct}}{C_{total} + W_{total} + B_{total}} \text{ , } Sp = \frac{N_{correct}}{N_{total}} \text{ , } Sc = \frac{Se + Sp}{2} \tag{2}$$

$i_{correct}$ represents the count of accurately classified respiratory cycles, and $i_{total}$ corresponds to the total number of respiratory cycles within the respective class when $i \in \{N, C, B, W\}$.

# 4 Experiments and Results

This section provides details of the experimental setup and subsequently compares the score differences between the wavegram logmel CNN model and the multi-modal hybrid model with 2 encoding methods using the snapshot ensemble approach [10]. We compare the results of the model after 10-fold cross validation with other works, based on the best configuration from the first experiment.

## 4.1 Experimental Setup

In deep learning tasks, optuna [2], an open-source automated hyperparameter optimization library, efficiently automates the search and optimization of hyperparameters for model training. We configured three hyperparameters: batch size, optimizer, and learning rate, each with specific search spaces in Optuna: [16, 32, 64], [SGD, Adam], and (0.001, 0.1), respectively. The objective function was defined as the model score, and we conducted a total of 100 trials. To expedite optimization, Optuna implemented early termination for trials with poor objective function performance. The optimized hyperparameter configuration for maximizing the objective function consisted of a batch size of 64, the Adam optimizer, and a learning rate of 0.001. These configurations were adopted for our experimental setup. Regarding dataset division, we split the dataset, consisting of 6898 respiratory cycles, into training and test sets with an 8:2 ratio. Moreover, within the first experiment, the training set was further divided, allocating 80% for model training and 20% for validation.

## 4.2 Performance comparison of wavegram logmel CNN model and multi-modal hybrid model

The snapshot ensemble method saves multiple snapshots of the model at regular intervals during the training of the neural network, enabling model integration without incurring any additional training costs. This method divides the entire training process into multiple cycles, each consisting of an equal number of epochs. The learning rate converges in each cycle using cyclic cosine annealing, which can be mathematically formulated as follows, where $\alpha_0$ denotes the initial learning rate at the restart of each cycle, $\alpha(t)$ denotes the learning rate when epoch is $t$, $T$ and $M$ are the total number of epochs and cycles, respectively

$$\alpha(t) = \frac{\alpha_0}{2} \cdot (cos(\frac{\pi \operatorname{mod} \lceil T/M \rceil}{\lceil T/M \rceil}) + 1) \tag{3}$$

Whenever a new cycle begins, the model initiates the exploration of the local optima and captures a snapshot at that specific location. At the end of each cycle, the model snapshot obtained during that cycle serves as the initialization for the subsequent cycle. During the model testing phase, the outputs of multiple snapshot models are averaged and compared against the corresponding actual labels, following the methodology outlined in Section 3.4.

**Table 5.** Performance of wavegram logmel and multi-modal hybrid models at different cycle numbers M

| | Unimodal wavegram logmel CNN | | | Multi-modal hybrid model (one-hot) | | | Multi-modal hybrid model (dummy) | | |
|---|---|---|---|---|---|---|---|---|---|
| M | Se(%) | Sp(%) | Sc(%) | Se(%) | Sp(%) | Sc(%) | Se(%) | Sp(%) | Sc(%) |
| 1 | 69.9 | 88.5 | 79.2 | 70.6 | 89.3 | 79.9 | 72.8 | 87.8 | 80.3 |
| 2 | 72.0 | 89.0 | 80.5 | 72.3 | 90.5 | 81.4 | 74.7 | 86.8 | 80.7 |
| 3 | 70.1 | 91.8 | 80.9 | 75.2 | 89.9 | 82.6 | 75.5 | 89.4 | 82.5 |
| 4 | **70.7** | **91.9** | **81.3** | **75.4** | **91.1** | **83.2** | **75.8** | **90.3** | **83.1** |
| 5 | 70.4 | 91.7 | 81.0 | 74.6 | 90.8 | 82.7 | 75.2 | 89.9 | 82.6 |

Each cycle comprises 30 epochs; Table 5 shows scores of the wavegram logmel CNN and multi-modal hybrid models with 2 encoding methods, ranging from single model scores to five-model integrations. Among these, the multi-modal hybrid model incorporating subject age, gender, and BMI in one-hot encoding outperforms the wavegram logmel CNN, which utilized only the respiratory cycle as input, across all integrated models. When subjected to four-model integrations, models achieve optimal performance of 81.3%, 83.2% and 83.1%, respectively.

### 4.3 Comparison to other works

Table 6 presents a comparison of the proposed method with existing approaches using the ICBHI dataset, as evaluated according to the methods described in section 3.4. A score of 64.8% [12] was attained by employing noise masking to mask abnormal sounds and utilizing RNN for respiratory cycle classification. A comprehensive investigation in [6] was conducted, encompassing various preprocessing techniques, resulted in an impressive score of 68.5% using a basic CNN. [1] added a Bi-LSTM layer after the CNN to capture temporal relations,

**Table 6.** ICBHI Challenge Comparison (four classes: normal, crackle, wheeze, both)

| Method | Se(%) | Sp(%) | Sc(%) |
|---|---|---|---|
| NMRNN [12] | 56.0 | 73.6 | 64.8 |
| RespireNet [6] | 53.7 | 83.3 | 68.5 |
| Hybrid CNN-RNN [1] | 56.9 | 86.7 | 71.8 |
| LSTM [20] | 62.0 | 85.0 | 74.0 |
| MBTCNSE [29] | 65.3 | 86.1 | 75.7 |
| CNN+Snapshot Ensemble [19] | 69.4 | 87.3 | 78.4 |
| CNN-MoE [21] | 68.0 | 90.0 | 79.0 |
| Wavegram logmel CNN (this paper) | 72.0 | 90.6 | 81.3 |
| Multi-modal hybrid model with one-hot (this paper) | **73.8** | **91.1** | **82.5** |
| Multi-modal hybrid model with dummy (this paper) | 73.7 | 90.3 | 82.0 |

leading to a score of 71.8% at the output of the fully connected layer. The RNN-based learning framework proposed by [20] achieved a model score of 74.0%. By integrating different types of neural networks, the model in [29] achieved a score of 75.7%. In [19], the snapshot ensemble method was applied to a custom CNN model, resulting in a score of 78.4%. [21] replaced the dense block in the C-DNN with a mixture-of-experts (MoE) block, and the Softmax Gate decided which expert to apply to which input region, achieving a score of 79.0%. Utilizing the experimental results presented in Section 4.2, we made the decision to incorporate four-model snapshots for computing the prediction scores. To ensure robustness, we conducted a 10-fold cross-validation on both the wavegram logmel CNN model and the multi-modal hybrid model. Subsequently, we conducted a comparative analysis with other works in the same data partitioning scenario. Notably, the classification outcomes obtained from the pre-trained wavegram logmel CNN outperformed other research endeavors. Moreover, the utilization of one-hot encoded demographic information during training in the multi-modal hybrid model yielded the most advanced results.

## 5 Conclusion and future work

Our work presents a multi-modal hybrid model incorporating respiratory cycles and demographic information (age, gender, and BMI) to accurately predict four classes (normal, crackle, wheeze and both) of respiratory cycles on the ICBHI dataset. To address the problem of missing demographic information, we conducted a comparative analysis of the effects of one-hot encoding and dummy encoding on model performance. Notably, the multimodal hybrid model utilizing one-hot encoding achieved an impressive score of 82.5%, surpassing previous works. We also examined the influence of the number of snapshot models on the model performance and determined the optimal number for our ensemble.

The promising results obtained from our multi-modal model indicate its potential for future research. To enhance the performance of multi-modal models, we intend to explore the impact of more encoding methods in our future work. Moreover, we plan to extend our investigations to encompass diverse datasets that contain various modalities. These efforts will contribute to a deeper understanding of multi-modal model and its applications in audio analysis.

## Acknowledgements

This work was partially funded by the BMBF under project number 13GW0554B (*Digitalung*). Figures were created with icons from [5].

## References

1. Acharya, J., Basu, A.: Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. IEEE transactions on biomedical circuits and systems 14(3), 535–544 (2020)

2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
3. Cheerla, A., Gevaert, O.: Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 35(14), i446–i454 (2019)
4. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y.: Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 626–635. Springer (2022)
5. Flaticon: Access 10.4m+ vector icons & stickers. https://www.flaticon.com/ (2023)
6. Gairola, S., Tom, F., Kwatra, N., Jain, M.: Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 527–530. IEEE (2021)
7. Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. Neural Computation 32(5), 829–864 (2020)
8. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
9. Heo, S.J., Kim, Y., Yun, S., Lim, S.S., Kim, J., Nam, C.M., Park, E.C., Jung, I., Yoon, J.H.: Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. International journal of environmental research and public health 16(2), 250 (2019)
10. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109 (2017)
11. Kochetov, K., Filchenkov, A.: Generative adversarial networks for respiratory sound augmentation. In: Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System. pp. 106–111 (2020)
12. Kochetov, K., Putin, E., Balashov, M., Filchenkov, A., Shalyto, A.: Noise masking recurrent neural network for respiratory sound classification. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. pp. 208–217. Springer (2018)
13. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, 2880–2894 (2020)
14. Li, C., Du, H., Zhu, B.: Classification of lung sounds using cnn-attention. EasyChair Preprint (4356) (2020)
15. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE international conference on computer vision. pp. 2623–2631 (2015)
16. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. vol. 8, pp. 18–25 (2015)

17. Minami, K., Lu, H., Kim, H., Mabu, S., Hirano, Y., Kido, S.: Automatic classification of large-scale respiratory sound dataset based on convolutional neural network. In: 2019 19th International Conference on Control, Automation and Systems (ICCAS). pp. 804–807. IEEE (2019)
18. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696 (2011)
19. Nguyen, T., Pernkopf, F.: Lung sound classification using snapshot ensemble of convolutional neural networks. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 760–763. IEEE (2020)
20. Perna, D., Tagarelli, A.: Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). pp. 50–55. IEEE (2019)
21. Pham, L., Phan, H., Palaniappan, R., Mertins, A., McLoughlin, I.: Cnn-moe based framework for classification of respiratory anomalies and lung disease detection. IEEE journal of biomedical and health informatics 25(8), 2938–2947 (2021)
22. Reichert, S., Gass, R., Brandt, C., Andrès, E.: Analysis of respiratory sounds: state of the art. Clinical medicine. Circulatory, respiratory and pulmonary medicine 2, CCRPM–S530 (2008)
23. Rocha, B., Filos, D., Mendes, L., Vogiatzis, I., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., et al.: A respiratory sound database for the development of automated classification. In: Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017. pp. 33–37. Springer (2018)
24. Rocha, B.M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y.P., Jakovljevic, N., Turukalo, T.L., Vogiatzis, I.M., Perantoni, E., et al.: An open access database for the evaluation of respiratory sound classification algorithms. Physiological measurement 40(3), 035001 (2019)
25. Sano, A., Chen, W., Lopez-Martinez, D., Taylor, S., Picard, R.W.: Multimodal ambulatory sleep detection using lstm recurrent neural networks. IEEE journal of biomedical and health informatics 23(4), 1607–1617 (2018)
26. Vyshedskiy, A., Alhashem, R.M., Paciej, R., Ebril, M., Rudman, I., Fredberg, J.J., Murphy, R.: Mechanism of inspiratory and expiratory crackles. Chest 135(1), 156–164 (2009)
27. World Health Organization: The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (2020)
28. Yap, J., Yolland, W., Tschandl, P.: Multimodal skin lesion classification using deep learning. Experimental dermatology 27(11), 1261–1267 (2018)
29. Zhao, Z., Gong, Z., Niu, M., Ma, J., Wang, H., Zhang, Z., Li, Y.: Automatic respiratory sound classification via multi-branch temporal convolutional network. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 9102–9106. IEEE (2022)