BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Deep learning-based cough classification using application-recorded sounds: a transfer learning approach with VGGish

Sanghoon Han[1], Yu-rim Lee[1], Ji-ho Lee[2], JinHee Jeon[3], Choongki Min[1], Kyungnam Kim[1], Donghoon Kim[4], Myung Pyo Kim[5], Young Mi Park[6], Uiri An[7] and Kyoung Min Moon[4,8,9*]

## Abstract

**Background** Coughing sounds contain various bio-metric information with regards to respiratory diseases that can help in the assessment of respiratory diseases. While clinicians find coughs insightful, non-experts struggle to identify abnormalities in cough sounds. Furthermore, respiratory diseases has characterized by widespread health complications and elevated mortality rates, the development of early diagnostic systems is imperative for ensuring timely intervention and improving outcomes for both clinicians and patients. Accordingly, we propose a deep learning–based model for early diagnosis. To enhance the reliability of the training data, we utilized annotations provided by multiple medical specialists. Additionally, we examined how clinical expertise and diagnostic input influence the model's generalization performance.

**Methods** This study introduces a deep learning framework utilizing VGGish as a transfer learning model, enhanced with additional detection and classification networks. The detection model identifies cough events within recorded audio, and then the classification model determines whether a detected cough is normal or abnormal. Both models were trained on raw cough sound data collected via smartphones and labeled by medical experts through a rigorous inspection process.

**Results** Experimental evaluations demonstrated that the cough detection model achieved an average accuracy of 0.9883, while the cough classification model attained accuracies of 0.8417, 0.8629, and 0.8662 among dataset1, 2, and 3. To enhance interpretability, we applied Grad-CAM to visualize the features that influenced the model's decision-making. Model performance was further evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC).

**Conclusions** Our proposed cough classification model has the potential to assist individuals with limited access to healthcare as well as medical professionals with limited experience in diagnosing cough-related conditions. By leveraging deep learning and smartphone-recorded cough sounds, this approach aims to enhance early detection and management of respiratory diseases.

*Correspondence:
Kyoung Min Moon
ml.pulmogicu@gmail.com

Full list of author information is available at the end of the article

**Keywords** Respiratory health, Cough classification, Cough detection, Deep learning, Medical diagnosis, Smartphone-based screening, VGGish model

## Background

Over the past decade, artificial intelligence (AI) has attracted remarkable interest across various industries, including smart healthcare, driven by advancements in computing power and storage capacity [1–6]. Additionally, several machine learning (ML)-based frameworks are being used for the general diagnosis of severe diseases affecting various systems, including the nervous, cardiovascular, digestive, and pulmonary systems [7–10].

In case of respiratory diseases, they represent a significant global health burden, characterized by widespread health complications and elevated mortality rates. As these conditions increasingly affect global health, the importance of early diagnosis and effective monitoring strategies has become increasingly critical for improving patient outcomes and reducing healthcare costs. Owing to the significance of early diagnosis and effective monitoring, voice-based diagnosis has emerged as a promising area of interest. Although AI algorithms have already shown commendable performance in image-based diagnosis [11, 12], sounds can carry the signature of many diseases [13–15]. In particular, AI speech analysis creates new opportunities in healthcare, such as the remote monitoring of various clinical outcomes and symptoms using vocal biomarkers for diagnosis, risk prediction, and overall health assessment [16]. Various studies use deep learning to detect and diagnose respiratory abnormalities. In the case of cough diagnosis, research has been conducted to develop models for classifying cough sounds by collecting data on asthma and normal coughs from children under the age of 16 [17], research has studied a classification framework based on cough sounds for identifying bronchitis and pneumonia in children [18], research used Audio Spectrogram Transformer model to classify cough status [19], research used a Bi-LSTM model to learn coughs of diseased patients and coughs of normal people [20], and researches [21, 22] used Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection and compared the performance of Deep Neural Decision Tree (DNDT), Deep Neural Decision Forest (DNDF), and various other machine learning models. Other case of cough detection, studies have explored cough detection methods as efficient alternatives for monitoring patients with wearable devices [23] and developing algorithms to assess recovery from pulmonary tuberculosis in areas with limited facilities around the world [24]. A study on classifying of lung sound using the VGGish model [25] employs a learning model for symptom classification. Additionally, this research explores the application of several convolutional neural network (CNN) models—AlexNet [26], VGG [27], Inception [28], and ResNet [29]—hich are highly effective in the image domain, to the audio domain as part of an experimental approach [30]. In contrast to previous studies [21, 22] that developed Decision Tree, Random Forest, and various machine learning models from scratch, we have applied transfer learning method to our experiment, because transfer learning is particularly useful when dealing with limited data and when the model's performance needs to be enhanced. Some systems use customized learning models [31–33], while others are built using pretrained models with transfer learning [34, 35]. We trained a learning model to identify valid cough samples from recordings labeled by professional medical staff and classify these cough sounds to diagnose abnormalities. Our model is designed to assist individuals with limited access to hospitals as well as assist doctors with limited medical resources, as it relies on cough sounds collected without any medical devices. Whereas studies [21, 22] depended on public COVID-19 datasets, we gathered and employed a dataset tailored for cough disease classification. To develop the learning model, we collected cough sounds from patients with asthma, COPD, and pneumonia at the hospital and studied the correlation between lung health and cough sounds.

Our cough data reflect the medical opinions of several healthcare professionals, making it a valuable resource for investigating how the inclusion of medical expertise affects model performance. In this study, we categorized the datasets based on the number of medical experts involved, trained models on each dataset, and compared their performance. Unlike previous studies, our research enhances the reliability of the data by incorporating the medical opinions of multiple healthcare professionals.

We utilized transfer learning with the pretrained VGGish model [25] and incorporated additional learning network blocks. We fine-tuned our learning model using labeled data from medical experts and through data inspection. After fine-tuning, we extracted Grad- CAM [36] from the model to analyze the features of true positive (TP) and true negative (TN) cough sounds. We then used the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) metrics to compare the model's performance across different datasets.

## Methods

### Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee

at Gangneung Asan Hospital (approval number: GNAH 2021-08-002-001). Informed consent was obtained from all participants, who signed a form authorizing the anonymous use of their clinical data for research, as approved by the Ethics Committee.

We collected cough sounds from the patients, and the process of collecting these sounds did not affect their treatment. The cough sounds were gathered as unique patient information in accordance with IRB approval. Furthermore, individual patients cannot be identified solely based on the cough sounds.

### Data Availability

The cough audio dataset analyzed in this study was ethically collected with approval from the Ethics Committee at Gangneung Asan Hospital (approval number: GNAH 2021-08-002-001). Given that individual cough recordings could potentially identify participants, full public dissemination of the dataset and original source code is restricted by ethical and privacy considerations mandated by the IRB. However, de-identified subsets of the data and the custom-developed code utilized for model training, validation, and evaluation can be made available upon reasonable request from qualified researchers. Access requests should be directed to the corresponding author and are subjected to review and approval by the relevant institutional ethics committees.

### Data collection

As shown in Figs. 1 and 2, medical staff directly collected patients' cough sounds using an app. The cough label data provided by the medical staff includes information on cough abnormality and the duration of cough. The cough data, collected in a tertiary care hospital over six months, involved a diverse cohort of patients. Our data are labelled two subject groups; one is healthy group, and the other is respiratory disease group. The respiratory diseased group includes conditions associated with coughing, such as asthma, pneumonia, and chronic obstructive pulmonary disease (COPD). As illustrated in Fig. 1, we collect patients' cough sounds using a smartphone app or a web page. With user consent, these platforms facilitate easy collection of cough sounds. Figure 2 displays a web page where specialists review and examine the cough recordings. The lower section of Fig. 2 is the medical examination data section, and the upper part is related to the cough signal. In the top section of Fig. 2, the x-axis represents time, while the y-axis represents the cough waveform. This interface allows medical experts to listen to individual cough sounds and modify annotations related to abnormality and cough duration. The labeling of the data was based on the clinical judgment of medical professionals for disease diagnosis. To enhance objectivity, the diagnoses from seven experts were aggregated and utilized.

This process enabled us to gather cough data from a total of 739 patients. Finally, we used data from a total of

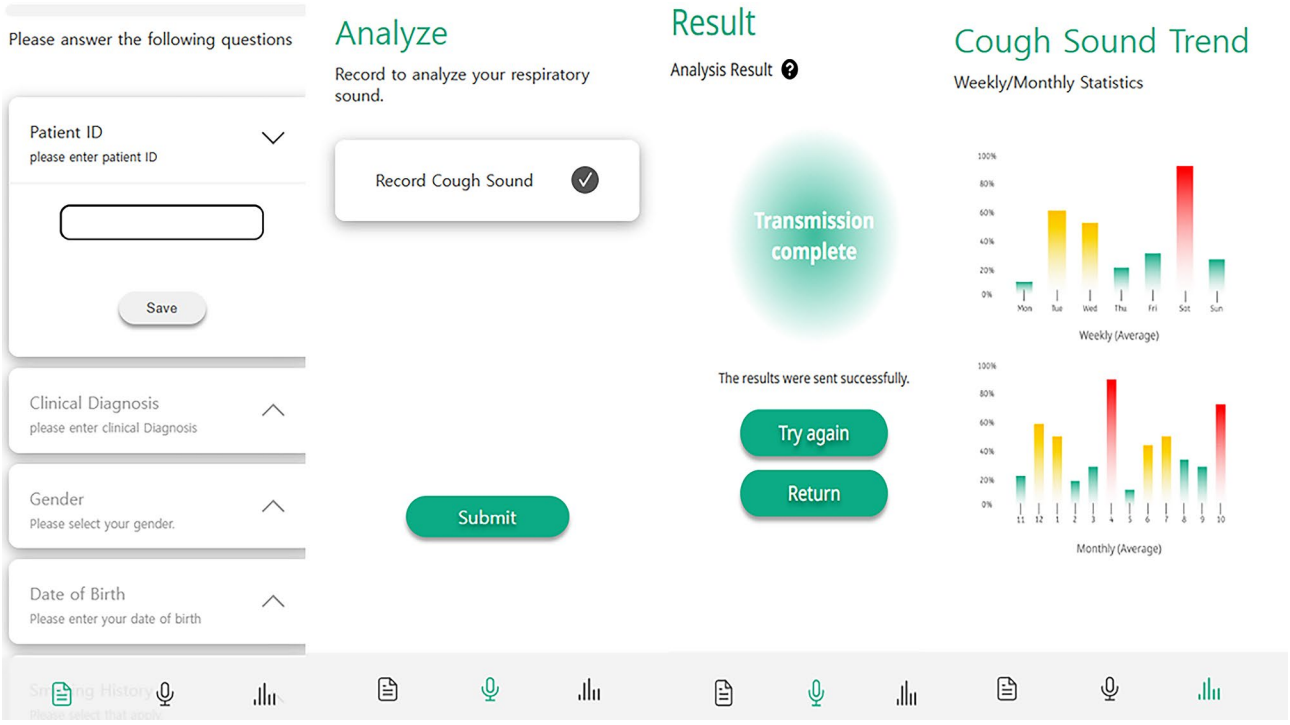


**Fig. 1** App interface for collecting patient cough sounds

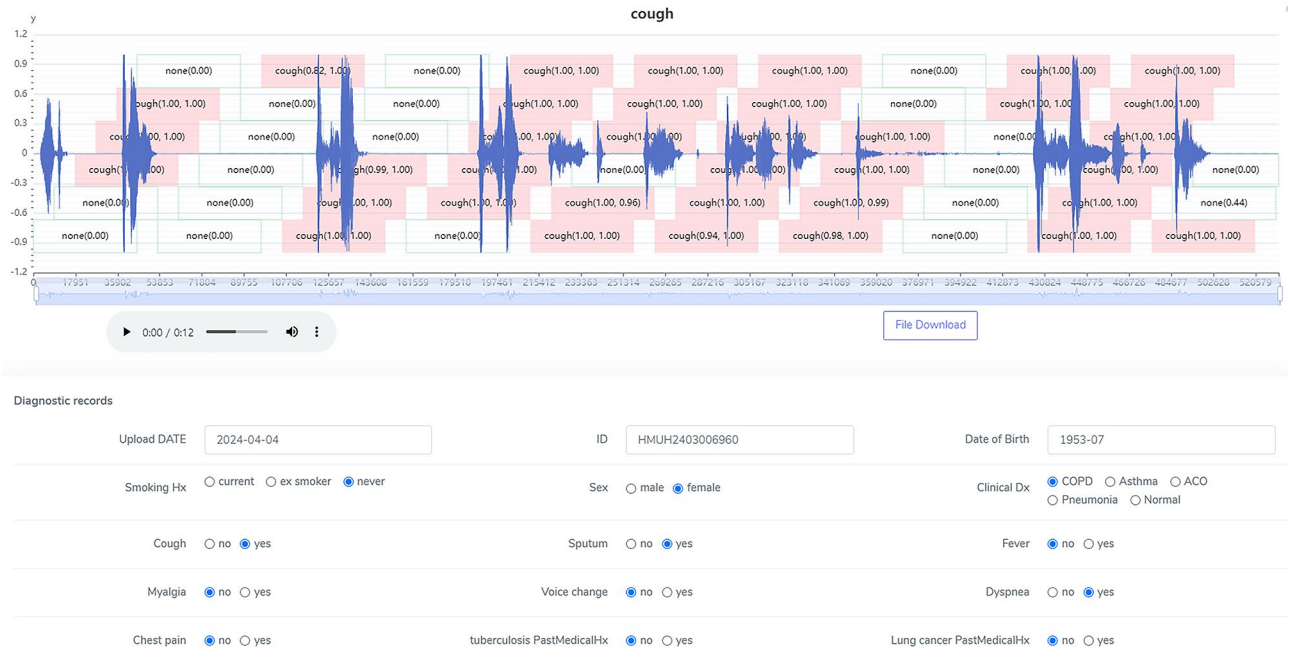


**Fig. 2** Web interface for inspecting patient cough sounds

476 patients after filtering out 33 patients whose lacked recording files, 63 patients whose data did not match their personal information and medical records, and 157 patients whose data did not reach a consensus among inspectors (Supplementary Tables 1, and Fig. 3).

After data collection, seven medical specialists listen to each patient's cough and assess its normality and duration in conjunction with clinical information. In this process, medical specialists assess the cough sounds, medical examination data, and apply their clinical expertise, as shown in Fig. 2.

Coughs identified as abnormal were categorized to have been collected from patients suffering from chronic obstructive pulmonary disease (COPD), asthma, and pneumonia. In contrast, normal coughs were extracted from healthy individuals. When medical experts identify an abnormal cough, they use the definitions of respiratory conditions (COPD, asthma, and pneumonia) described below to determine the cough's normality across all datasets. COPD is a heterogeneous lung condition characterized by symptoms such as dyspnea, cough, sputum production, and exacerbations [37]. Asthma is a heterogeneous disease marked by chronic airway inflammation and symptoms such as wheezing, shortness of breath, chest tightness, and cough [38]. Pneumonia is an acute infection of the lung parenchyma caused by various pathogens, distinct from bronchiolitis [39].

## Data pre-processing

We pre-processed the cough sounds to extract 1-second samples for model training, each containing temporal information and disease labels. Diagnosing respiratory diseases generally requires the consideration of multiple symptoms and patient examination results, which makes it difficult to accurately diagnose cough abnormalities based solely on cough sounds. Nevertheless, previous studies [40–42] have shown that cough sounds can serve as valuable biomarkers for cough classification. Therefore, to improve the diagnostic value of our data for cough classification, we conducted multiple rounds of data review with medical experts.

Asthma is a respiratory disease characterized by symptoms such as shortness of breath and chest tightness. In adults, asthma can lead to a distinct cough, often accompanied by wheezing and a high-pitched whistling sound in the upper frequency spectrum (above 2 kHz) during exhalation due to constricted airways [43]. This wheezing falls within a distinct high-frequency range. Asthma coughs are typically dry and have a sharp, abrupt quality, resulting in higher-frequency sound elements [44]. In contrast, COPD presents with a heterogeneous lung condition and chronic respiratory symptoms. COPD coughs are characterized by coarse crackling sounds, a series of short, explosive noises, and extended, low-pitched (typically below 1 kHz) tones [43]. These sounds may indicate of the presence of fluid in the patient's air sacs, a common symptom of COPD. Pneumonia, an acute respiratory infection commonly caused by viruses or bacteria,

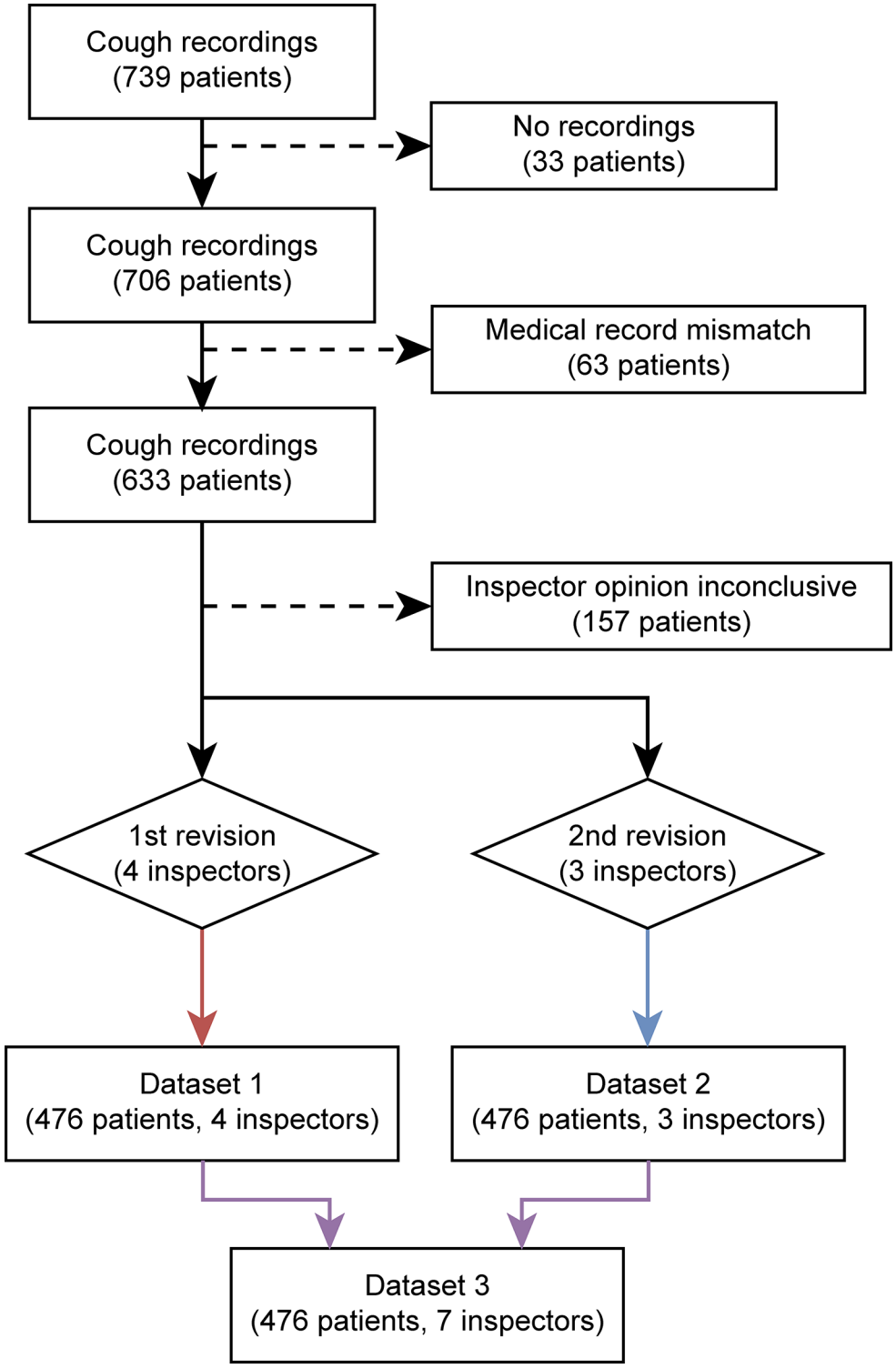


**Fig. 3** Flowchart of dataset division

typically presents with a deeper or louder cough compared to other coughs. A typical cough, which is commonly caused by a cold or a respiratory illness, affects the acoustic properties and may add lower-frequency components or a wide range of frequency characteristics. When labeling coughs, medical experts classified each one as normal or abnormal based on these symptom-related criteria.

During the inspection, we gathered the assessments from all medical experts and decided on the label (normal

vs. abnormal) for each cough based on the majority opinion. Once all medical opinions were collected, the majority consensus was used to assign the final labels. Following this data inspection process, we created three versions of the dataset (Dataset 1, Dataset 2, and Dataset 3) and augmented the cough sample based on the final labels.

As shown in Fig. 3, we divided the datasets based on the number of inspecting medical experts to evaluate model performance and suitability relative to the number of inspectors. We created three datasets by integrating the opinions of four, three, and seven reviewers, respectively. Dataset 1 includes labels from four medical experts (K. Moon, J. Lee, Y. Park, and M. Kim). The cough samples in Dataset 1 were labeled by these experts. Dataset 2 comprises the opinions of three medical experts (D. Kim, J. Jun, and U. Ahn) who independently labeled the cough samples, separate from Dataset 1. After labeling, we combined the experts' labels and removed samples where there was no majority consensus. Dataset 3 includes the combined opinions of all seven medical experts included in both Datasets 1 and 2, all with over a decade of clinical experience (four pulmonologists: K. Moon, J. Lee, M. Kim, and U. Ahn; one cardiologist: D. Kim; one pediatrician: Y. Park; one general practitioner: J. Jun). During the first inspection (Dataset 1), we filtered out patient data where there was no majority opinion, such as cases with an equal number of normal and abnormal classifications. As a result, we compiled a dataset set using data from 476 patients. For the second data inspection (Dataset 2), we used the same patient data as in the first data inspection. In total 476 patients, the dataset consists of a total of 12,970 cough sound samples. As shown in Table 1, Dataset 1 consists of 4,390 normal and 8,580 abnormal samples; Dataset 2 consists of 6,905 normal and 6,065 abnormal samples; and the final Dataset 3 consists of 5,780 normal and 7,190 abnormal samples. The variation in the number of normal and abnormal samples across the datasets is attributed to the fact that each dataset was independently annotated by different medical experts based on their diagnostic assessments. When recruiting medical experts to validate our data, we encountered challenges related to the workload of medical professionals and difficulties in finding suitable experts. As a result, we initially created Dataset 1 with four medical professionals, and later, we added three more professionals to form Dataset 2.

#### *Training data*

We labeled each 1-second sample of the recorded sound as either "cough" or "not cough". As shown in Fig. 4a, we generated samples by sliding a window from 0 s to T (the end of the recording) in 0.2-second increments. The gray windows represent noise and silence, while the orange windows denote cough segments. This sampling method was used to create the dataset for the cough detection model. As shown in Fig. 4b, we extracted samples from each cough by sliding a window of 0.1 s around the median cough time, spanning from − 0.2 s to 0.2 s. Using sliding windows, we capture a range of cough variations from different individuals by generating samples with various time positions within each window. After applying the sliding window technique, we label each window sample based on its overlap with the labeled cough period. If the overlap with the cough period exceeds 50% (0.5 s), the sample is labeled as abnormal if medical experts diagnose the cough as related to COPD, asthma, or pneumonia; otherwise, it is labeled as normal.

### Model architecture

Recently, the use of AI models in the field of medical data research, particularly for analyzing and detecting cough sounds, has been growing rapidly. Deep learning networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have proven effective in various domains and are increasingly being adapted for medical data analysis. However, these networks often suffer from overfitting when data are limited, leading to suboptimal performance. To address this issue, transfer learning can enhance the performance of deep learning models, especially in scenarios with insufficient data. Herein, we address the challenge of data scarcity by employing the VGGish model with transfer learning to improve performance.

#### *Proposed model*

We propose two models built upon the VGGish architecture, augmented with two additional learning networks.

**Table 1** Training and testing data samples for each fold in classification

| | | 1-Fold | | 2-Fold | | 3-Fold | | 4-Fold | | 5-Fold | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Normal | Abnormal | Normal | Abnormal | Normal | Abnormal | Abnormal | Normal | Normal | Abnormal |
| **Dataset 1** | Train | 3,510 | 6,860 | 3,510 | 6,865 | 3,510 | 6,865 | 3,515 | 6,865 | 3,515 | 6,865 |
| | Test | 880 | 1,720 | 880 | 1,715 | 880 | 1,715 | 875 | 1,715 | 875 | 1,715 |
| **Dataset 2** | Train | 5,520 | 4,850 | 5,525 | 4,850 | 5,525 | 4,850 | 5,525 | 4,855 | 5,525 | 4,855 |
| | Test | 1,385 | 1,215 | 1,380 | 1,215 | 1,380 | 1,215 | 1,380 | 1,210 | 1,380 | 1,210 |
| **Dataset 3** | Train | 4,620 | 5,750 | 4,625 | 5,750 | 4,625 | 5,750 | 4,625 | 5,755 | 4,625 | 5,755 |
| | Test | 1,160 | 1,440 | 1,155 | 1,440 | 1,155 | 1,440 | 1,155 | 1,435 | 1,155 | 1,435 |

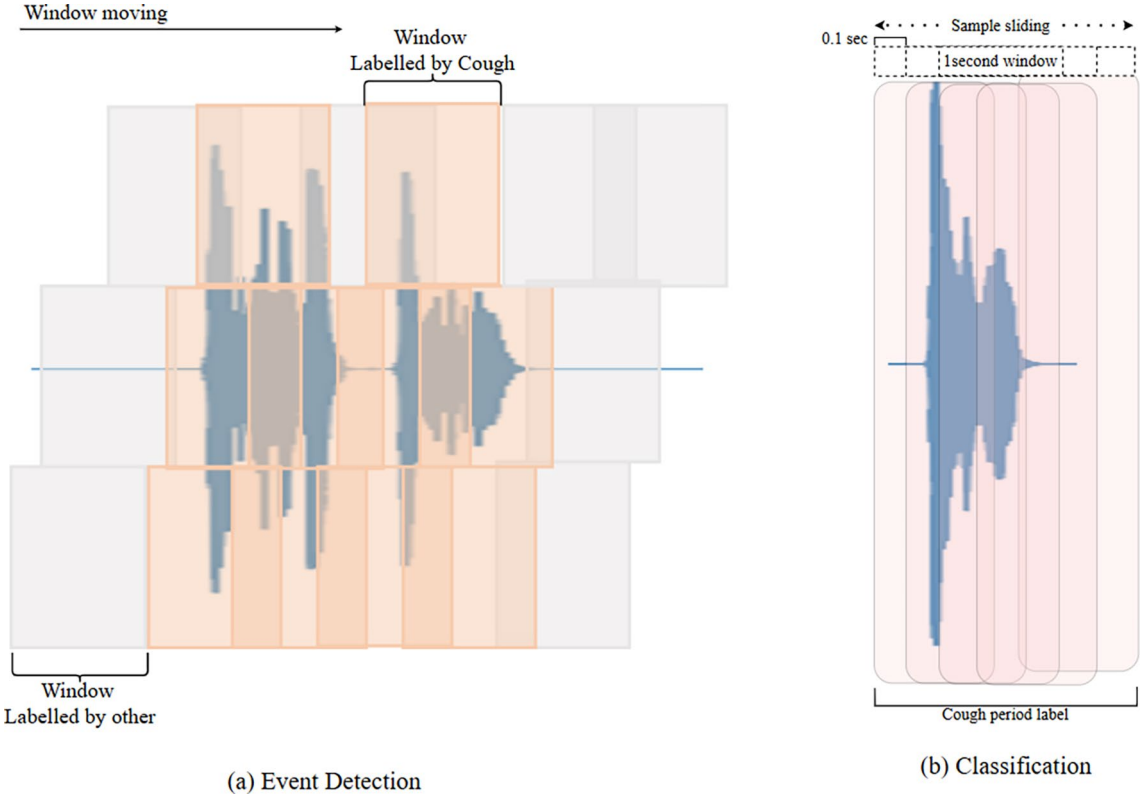


**Fig. 4** Visualized image of (**a**) event detection and (**b**) classification sampling. Event detection sampling (left) slides the window from the start to the end of the total cough sounds, while classification sampling (right) involves augmentation from a single cough sound

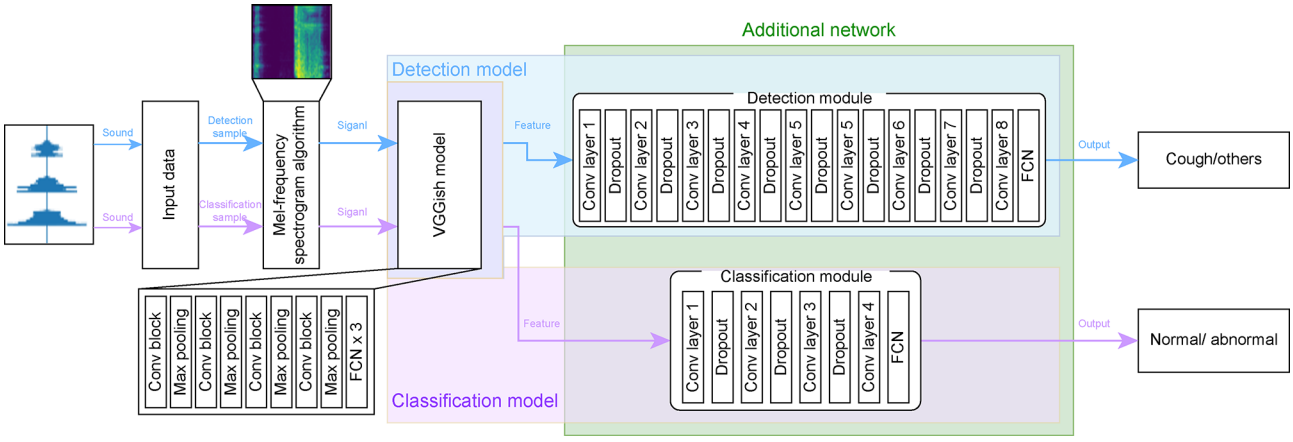


**Fig. 5** Overall architecture of the model

As shown in Fig. 5, the VGGish model is composed of 4 convolution blocks, max pooling layers, and 3 fully connected layers. To enhance the model's capabilities, we integrate two additional networks at the end of the VGGish model: a cough detection network and a cough classification network that analyze the cough samples and determines whether they are normal or abnormal. Specifically, the cough detection network and the cough classification network each feature different configurations to address their respective tasks. The cough classification network includes 8 convolutional layers with channel sizes of [256, 128, 64, 32], while the cough detection network comprises 4 convolutional layers with channel sizes of [256, 512, 1024, 512, 256, 128, 64, 32]. Both networks incorporate dropout layers to mitigate overfitting. Each convolution layer in the additional networks uses a kernel size of 2 and a stride of 1, with ReLU as the active function. We use the dropout layer because the transferred parameters of VGGish have already been sufficiently learned and reflect the characteristics of the sound signal,

since the general dataset of Google Audio Set includes a large number of samples and various categories.

#### *Pretrained model*

The VGGish model is a pre-trained model with Google Audio Set [45]. The Google Audio Set is a large labeling dataset in which people directly label about 10 s of audio extracted from YouTube videos. A total of 2,084,320 data consists of 632 classes, 527 classes of which are used for learning, whereas 105 classes are typically excluded due to ambiguity [45]. In the complete VGGish model, the transfer learning layer spans from the start of the VGGish architecture up to, but not including the final fully connected layer. Both the pre-trained and VGGish models used for transfer learning share the same structure and parameters.

As depicted in Fig. 5, we extracted features for input into additional networks from the VGGish model using our collected data. For feature extraction, which produces a 128-dimensional vector, we utilized the parameters from the pretrained VGGish model.

#### *Fine-tuning*

We created two learning models: cough detection and cough classification models. We set the parameters of the VGGish model learned with Google Audio Set to the initial state of our model before fine-tuning. After setting the parameters, we fine-tuned each model to improve model performance using our collected data. As shown in Fig. 5, each additional network uses the input feature that was extracted from the VGGish model. While fine-tuning, we updated all parameters of the VGGish model and the two additional networks (Cough Detection and Cough Classification) using our collected data. After this process, we obtained two fine-tuned models: one for cough detection and one for cough classification.

#### Analysis: AUROC, AUPRC graph

In this study, we investigated the AUROC and AUPRC to evaluate model performance. The AUROC is a valuable tool for assessing prediction accuracy and represents the model's discriminatory performance. In the case of imbalanced data, AUPRC is often a more appropriate evaluation metric than AUROC [46]. For this reason, we used both AUROC and AUPRC to comprehensively compare the performance of each dataset's model.

AUROC is plotted with the true positive rate (recall) on the y-axis and the false positive rate on the x-axis. A higher AUROC value, approaching 1, indicates better model performance, with the curve skewed towards the top left corner of the plot.

AUPRC is plotted with precision on the y-axis and recall on the x-axis. A higher AUPRC value, approaching 1, signifies better model performance, with the curve skewed towards the top right corner of the plot.

A detailed explanation of AUROC and AUPRC has been included in the supplementary material.

#### Analysis: Grad-Cam

We extracted the Grad-CAM from the cough classification model to analyze the VGGish features. By examining the Grad-CAM, we can identify which frequency bands were most influential in the learning model, given that the input data are a spectrogram with both time and frequency axes. The Grad-CAM was obtained from the fourth convolution block of the VGGish model, which is part of the transfer learning component of our cough classification model. This fourth convolution block consists of two layers, and we focused on extracting Grad-CAM from the second layer. In our analysis, we emphasized the magnitude of the Grad-CAM values rather than their sign [47].

## Experiment

### Experimental design

As shown in Fig. 5, we trained our two models using cough signal data extracted from the cough samples. We transformed our cough signal data (window samples) into a Log-Mel spectrogram using the Fourier transform and Mel-frequency spectrogram algorithm. The cough signal was converted to a Mel spectrum using the following equation for the Mel frequency $f$:

$$f = 1172\ln\left(1 + \frac{f}{700}\right)$$

After data transformation, the signal data is converted from the time domain into a frequency band. We used this transformed signal as input for the VGGish model instead of the raw cough sound. The parameters for the Log-Mel spectrogram are as follows: a sample rate of 16,000 Hz, 64 Mel bins, a minimum Mel frequency of 125 Hz, a maximum Mel frequency of 7,500 Hz, and a short-term-Fourier transform window length of 0.025 s with a hop length of 0.01 s.

### Data construction

As shown in Fig. 4, we sampled the cough window with a length of 1-second. Each sample was converted into a Log-Mel spectrogram (94 frames x 64 Mel bins) using the abovementioned data processing. The VGGish model then extracts a 128-dimensional feature vector from the Log-Mel spectrogram. We applied 5-fold cross-validation to Datasets 1, 2, and 3, and the number of samples in each dataset is shown in Tables 1 and 2. We applied the 5-fold cross-validation method, creating a training set and a test set for each fold. The data was divided into

**Table 2** Training and testing data samples for each fold in event detection

| | | 1-Fold | | 2-Fold | | 3-Fold | | 4-Fold | | 5-Fold | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cough | Other | Cough | Other | Cough | Other | Cough | Other | Cough | Other |
| Dataset | Train | 25,000 | 24,204 | 25,000 | 24,039 | 25,000 | 23,989 | 25,000 | 23,661 | 25,000 | 24,204 |
| | Test | 7,000 | 5,907 | 7,000 | 6,010 | 7,000 | 6,027 | 7,000 | 6,312 | 7,000 | 5,721 |

training and test sets with an 8:2 ratio. We did not create a separate validation set, as the 5-fold cross-validation process includes validation within each fold.

In the cough detection experiment, we applied a downsampling method to address data imbalance. In the cough classification experiment, we compared the performance of the classification models using Datasets 1, 2, and 3. We designated normal cough samples as negative and abnormal cough samples as positive.

This paper investigates how variations in the number of data inspectors, who are medical experts, impact the classification model's performance. All datasets consist of 12,970 cough samples, and each is divided into approximately 10,370 samples for training and 2,600 samples for testing. Since each dataset has a different number of medical experts and combines the experts' opinions, the number of normal and abnormal samples is different in dataset 1, 2 and 3.

### Cough detection

In the cough detection part, the input data consists of both cough and other samples (Table 1). As shown in Fig. 4a, "other" samples include all sounds except cough samples (indicated in orange). These other samples encompass silence, patients' speech, and background noise, while cough samples specifically represent patients' coughs. The total sampling data spans from 0 to T seconds (end of the signal), and each window sample was labeled based on time label data. We trained our cough detection model, as shown in Fig. 5, using an optimizer with a learning rate of 0.001 and a dropout rate of 0.3 and applied a binary-cross entropy loss function.

### Cough classification

In the cough classification part, the input data consists of normal and abnormal cough samples. As shown in Fig. 4b, we sampled all patients' coughs from Datasets 1, 2, and 3. The number of medical experts inspecting each dataset varies, resulting in different train data ratios. Dataset 1 was labeled based on the diagnoses of four medical experts, while Dataset 2 was labeled by three medical experts. Dataset 3 was constructed by integrating the diagnostic results from both Datasets 1 and 2. The number of medical experts involved in the diagnostic process varied for each dataset, resulting in different ratios of normal to abnormal samples across the datasets. This deliberate variation allowed us to systematically investigate how the number of medical experts contributing to the labeling process affects the performance of cough classification models. We trained our cough classification model, as shown in Fig. 5, with an optimizer learning rate of 0.0003 and a dropout rate of 0.7 and applied a binary-cross entropy loss function. We set the

**Table 3** Performance scores for cough detection and classification models

| | | AUROC | Accuracy | Precision | Recall | AUPRC | Specificity | F1 score |
|---|---|---|---|---|---|---|---|---|
| Detection | | NA | 0.9883 ± 0.0027 | 0.9966 ± 0.0022 | 0.9816 ± 0.0033 | NA | 0.9960 ± 0.0026 | 0.9890 ± 0.0025 |
| Classification | Dset 1 | 0.9345 ± 0.0075 | 0.8417 ± 0.0181 | **0.8845 ± 0.0269** | **0.8771 ± 0.0365** | 0.9298 ± 0.0099 | 0.7725 ± 0.0703 | **0.8798 ± 0.0143** |
| | Dset 2 | 0.9028 ± 0.0094 | 0.8629 ± 0.0085 | 0.8658 ± 0.0262 | 0.8384 ± 0.0300 | 0.9415 ± 0.0071 | **0.8845 ± 0.0278** | 0.8511 ± 0.0098 |
| | Dset 3 | **0.9348 ± 0.0046** | **0.8662 ± 0.0076** | 0.8818 ± 0.0126 | 0.8765 ± 0.0153 | **0.9443 ± 0.0032** | 0.8535 ± 0.0195 | 0.8790 ± 0.007 |

Dset, Dataset; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve

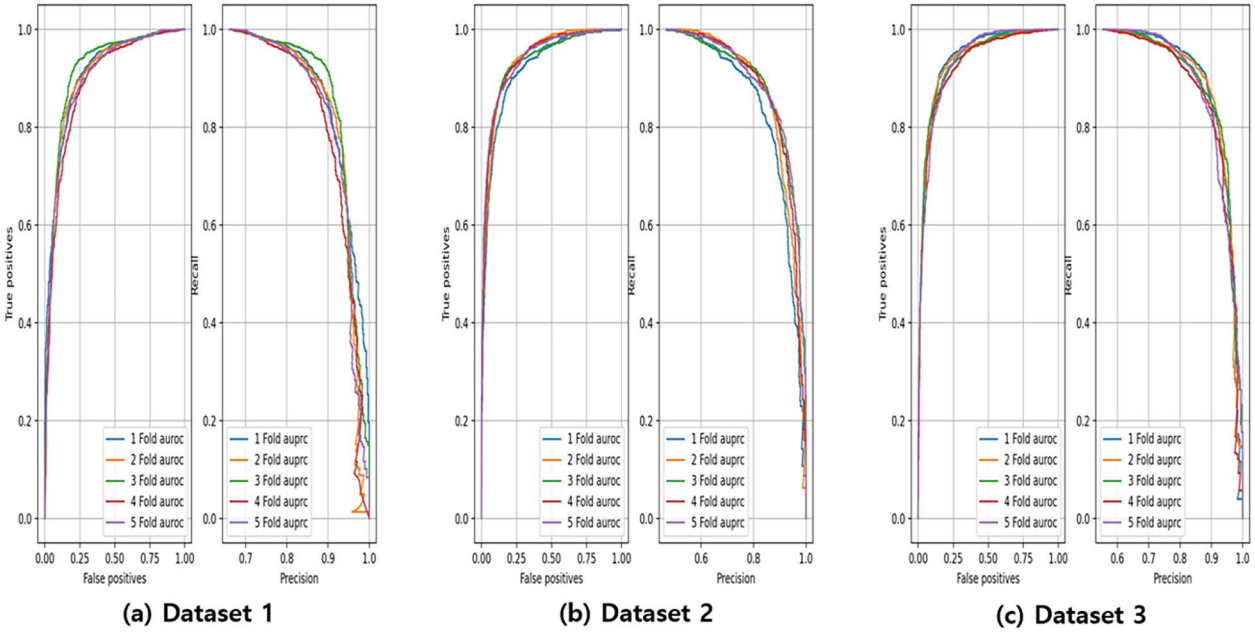


**Fig. 6** AUPRC and AUROC for cough classification in Datasets 1, 2, and 3. AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve

dropout rate to 0.7 to mitigate overfitting, given that the VGGish model was pretrained.

## Results

We summarized the results of the detection and classification experiments in Table 3, presenting the performance of each fold with mean values and standard deviations. The evaluation metrics used to compare the performance of each learning model included accuracy, precision, recall, specificity, and F1 score. We also calculated the AUROC and AUPRC to assess model performance across different thresholds. As illustrated in Fig. 6, we have extracted the AUROC and AUPRC for each dataset to evaluate the performance comprehensively.

### Detection model result

As shown in Table 3, the cough detection model achieved the following performance metrics: Accuracy: 0.9883 ± 0.0027, Precision: 0.9966 ± 0.0022, Recall: 0.9816 ± 0.0033, Specificity: 0.9960 ± 0.0026, and F1 Score: 0.9890 ± 0.0025. The model effectively differentiates between coughs and other sounds, a distinction that is also clear to medical experts. Specifically, the model shows a marked difference in classifying silence samples versus cough samples, as well as distinguishing between other samples and cough samples. Silence samples produce minimal signal in the Log-Mel spectrogram, while other sounds are more discernible to the human ear. These factors contribute to the model's high performance.

### Classification model result

For the classification models, the performance differences across the datasets are demonstrated. In Dataset 1, the classification model achieved the following metrics: Accuracy: 0.8417 ± 0.0181, Precision: 0.8845 ± 0.0269, Recall: 0.8771 ± 0.0365, Specificity: 0.7725 ± 0.0703, and F1 Score: 0.8798 ± 0.0143. For Dataset 2, the classification model yielded: Accuracy: 0.8629 ± 0.0085, Precision: 0.8658 ± 0.0262, Recall: 0.8384 ± 0.0300,

Specificity: 0.8845 ± 0.0278, and F1: 0.8511 ± 0.0098. For Dataset 3, the classification model showed: Accuracy: 0.8662 ± 0.076, Precision: 0.8818 ± 0.0126, Recall: 0.8765 ± 0.0153, Specificity: 0.8535 ± 0.0195, and F1: 0.8790 ± 0.007.

Based on the results, the model trained on Dataset 1 achieved the highest precision, recall, and F1 score, but demonstrated lower performance in specificity. The model trained on Dataset 2, on the other hand, achieved the highest specificity but underperformed in the other metrics. In contrast, the model trained on Dataset 3 showed the best overall performance in terms of AUROC, AUPRC, and accuracy, indicating a more balanced and consistently high performance across all evaluation metrics. We consider the model trained on Dataset 3 to be the most optimal for cough classification, as it demonstrates more balanced performance across all evaluation metrics compared to the models trained on the other datasets.

Additionally, we trained the comparative models using Dataset 3 and evaluated their performance alongside our proposed model. The results of this comparison, including models such as VGG + LSTM, VGG + ConvLSTM, and ResNet50, are summarized in Table 4. As shown in Table 4, the models that extend VGG with additional modules exhibit overall improved performance compared to the ResNet-based model. Among them, the proposed VGG + CNN model demonstrates superior performance relative to the other comparative models in the experiment. Thus, we believe the Dataset 3 model is the most effective for our needs, particularly for use by individuals with limited access to medical services and doctors with constrained resources.

We attribute the observed differences in performance to the varying number of medical experts involved in inspecting the cough samples. As shown in Table 2, there is a remarkable disparity in the ratio of normal-to-abnormal samples. Specifically, Dataset 1 contains approximately 2,000 more abnormal samples than normal samples. This imbalance leads to the Dataset 1 model being biased towards detecting abnormal samples, resulting in higher precision and recall scores but lower specificity compared to the Dataset 3 model. Conversely, Dataset 3 features a more balanced ratio of normal and abnormal samples, yielding a more balanced performance overall. This balance is reflected in Fig. 6, where the AUPRC for Dataset 3 (Fig. 6c, right) shows denser lines compared to that of Datasets 1 (Fig. 6a, right) and 2 (Fig. 6b, right). Additionally, the AUROC for Dataset 3 (Fig. 6c, left) exhibits lines that are closer together than those in Datasets 1 (Fig. 6a, left) and 2 (Fig. 6b, left). Figures 7 showed the confusion matrixes of the model trained with Dataset 1 (a), Dataset 2 (b), and Dataset 3 (c). As discussed, we believe that the ratio of normal-to-abnormal samples is an important factor influencing model performance. Additionally, improving the quality of data labeling, which is enhanced by involving more medical experts, will extensively impact the effectiveness of model training.

### Grad-CAM analysis result

In this paper, we compared Grad-CAM images for true positives (TPs) and true negatives (TNs). A true positive occurs when an abnormal cough is correctly identified as abnormal, and a true negative occurs when a normal cough is correctly identified as normal. As shown in Fig. 8, we generated the mean Grad-CAM for TN and TP to compare across each dataset and this figure visualizes which aspects the classify model focuses on when classifying data structured based on medical expertise. For TPs, the model is predominantly influenced by the mid-frequency bands, whereas for TNs, the model is influenced by more dispersed frequency bands.

## Discussion

One limitation of our implementation is that it does not account for the characteristics of time series data. To address this, we propose using the entire cough recording as input rather than a single sample. This approach allows for variable input lengths and enables the model to learn cough intervals more effectively. Additionally, it can improve efficiency by eliminating the need for a separate cough detection step.

Another limitation is that our data were exclusively collected from Korean hospitals, which may present challenges when the model is used by individuals in other countries. To overcome this, we plan to expand our data collection to include cough sounds from hospitals in other countries.

**Table 4** Performance scores for comparing cough classification models

| | | AUROC | Accuracy | Precision | Recall | AUPRC | Specificity | F1 score |
|---|---|---|---|---|---|---|---|---|
| Classification (Dset 3) | VGG + CNN | **0.9348 ± 0.0046** | **0.8662 ± 0.0076** | **0.8818 ± 0.0126** | **0.8765 ± 0.0153** | **0.9443 ± 0.0032** | **0.8535 ± 0.0195** | **0.8790 ± 0.007** |
| | VGG + LSTM | 0.9127 ± 0.0041 | 0.8312 ± 0.0142 | 0.8338 ± 0.0319 | 0.8725 ± 0.0366 | 0.9258 ± 0.0049 | 0.7799 ± 0.0628 | 0.8515 ± 0.0104 |
| | VGG+ CONVLSTM | 0.9126 ± 0.0090 | 0.8331 ± 0.0056 | 0.8585 ± 0.0168 | 0.8381 ± 0.0315 | 0.9261 ± 0.0071 | 0.8268 ± 0.0314 | 0.8475 ± 0.0083 |
| | ResNet50 | 0.8780 ± 0.0193 | 0.7977 ± 0.0231 | 0.7946 ± 0.0394 | 0.8611 ± 0.0163 | 0.8945 ± 0.0149 | 0.7188 ± 0.0701 | 0.8256 ± 0.0146 |

Dset, Dataset; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve

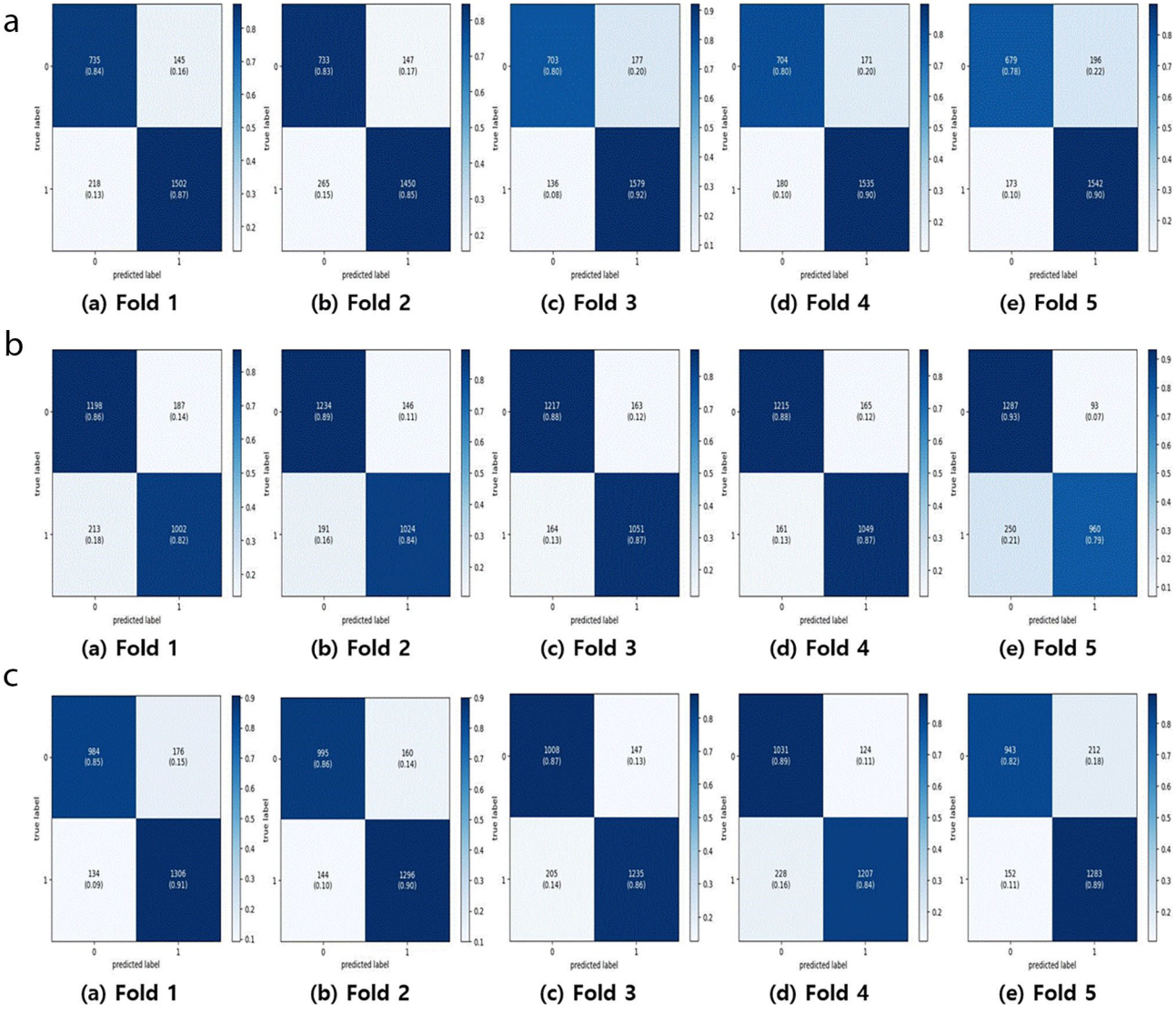


**Fig. 7** Confusion matrix of the model trained with (**a**) Dataset 1, (**b**) Dataset 2, and (**c**) Dataset 3

Finally, the cough sounds with background noise have a negative effect on the classification model. To solve this problem, we plan to study how to remove noise in coughing sounds.

For future work, we plan to expand our research in several directions to further improve model performance and explore additional clinical insights.

First, we aim to develop a symptom-specific classification model by utilizing the available symptom annotations and patient severity levels for each cough sample. This will allow us to train models to classify various symptoms and compare their performance across symptom categories.

Second, we intend to investigate the effect of patient sex on model performance. To do this, we will divide the dataset by sex, train separate models for each group, and analyze differences in classification accuracy.

Third, we plan to examine how increasing the dataset size influences model performance. By collecting more diverse cough data, we aim to evaluate the scalability and robustness of the model.

Additionally, we will explore the impact of medical experts' specialties on model accuracy. Specifically, we will compare models trained on labels provided by respiratory specialists versus those provided by non-respiratory clinicians. This comparison will help determine whether specialist knowledge contributes to more reliable model outputs.

To support this study, we intend to expand our dataset by incorporating diagnostic annotations from a greater number of medical experts.

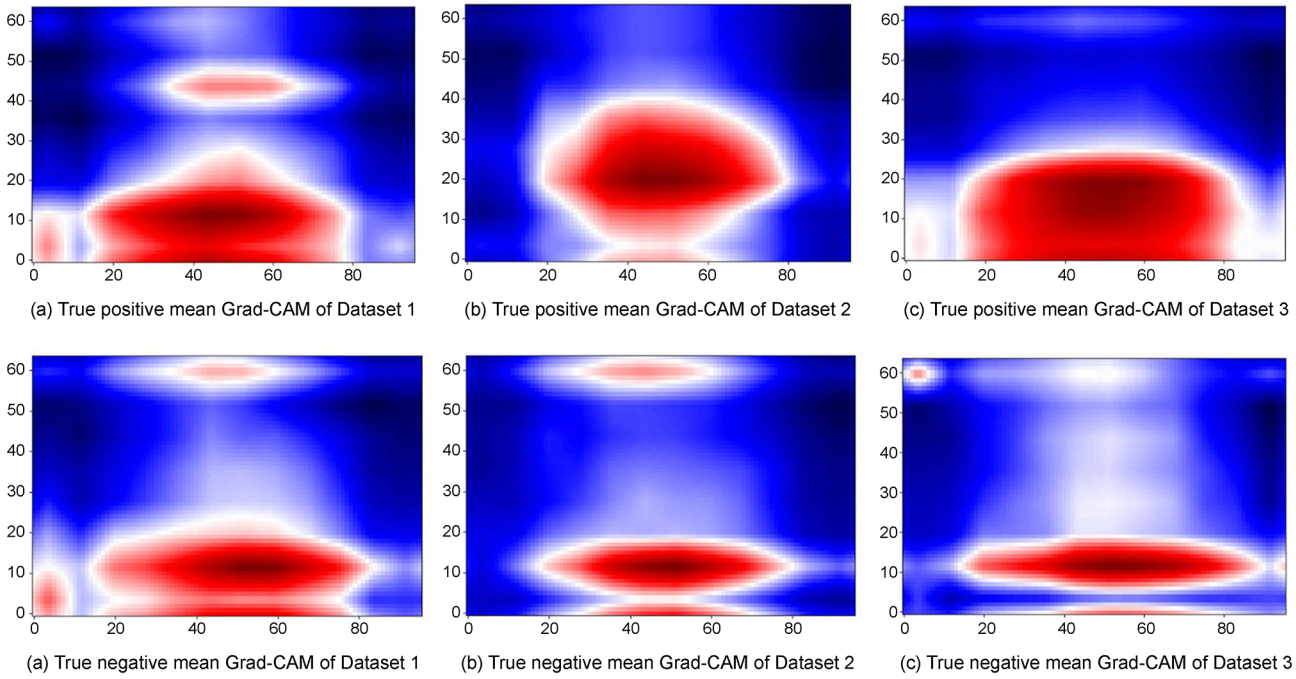


**Fig. 8** Visualized mean Grad-CAM of Datasets 1, 2, and 3

## Conclusions

In this study, we investigated the influence of the number of medical experts involved in data annotation on the performance of cough classification models. As highlighted in Sect. 4.2, both data acquisition and labeling quality are critical to achieving reliable and effective model performance.

Our experimental setup focused on classifying cough sounds collected via smartphones into either normal or abnormal categories, without relying on additional clinical equipment or metadata. Among the models evaluated, the one trained on Dataset 3, which included annotations from seven medical specialists, demonstrated the most balanced and robust performance. This model not only outperformed others across key evaluation metrics but also proved to be the most generalizable, making it well-suited for real-world applications. Therefore, as our model is trained using diagnostic data annotated by medical professionals, it may be suitable for self-screening purposes and could offer advantages in the early detection and management of respiratory diseases. Nonetheless, further clinical validation is necessary to confirm its practical applicability.

As illustrated in Fig. 9, the proposed model holds strong potential for deployment as a practical healthcare support tool. It may assist individuals in remote or underserved areas with limited access to medical resources and help alleviate the diagnostic burden on healthcare professionals, particularly in environments facing labor shortages.

Moving forward, we aim to further enhance the model by expanding the dataset, incorporating symptom-specific and demographic information, and exploring specialist-driven annotation strategies. These efforts will contribute to the development of more accurate, scalable, and accessible AI-based diagnostic tools.

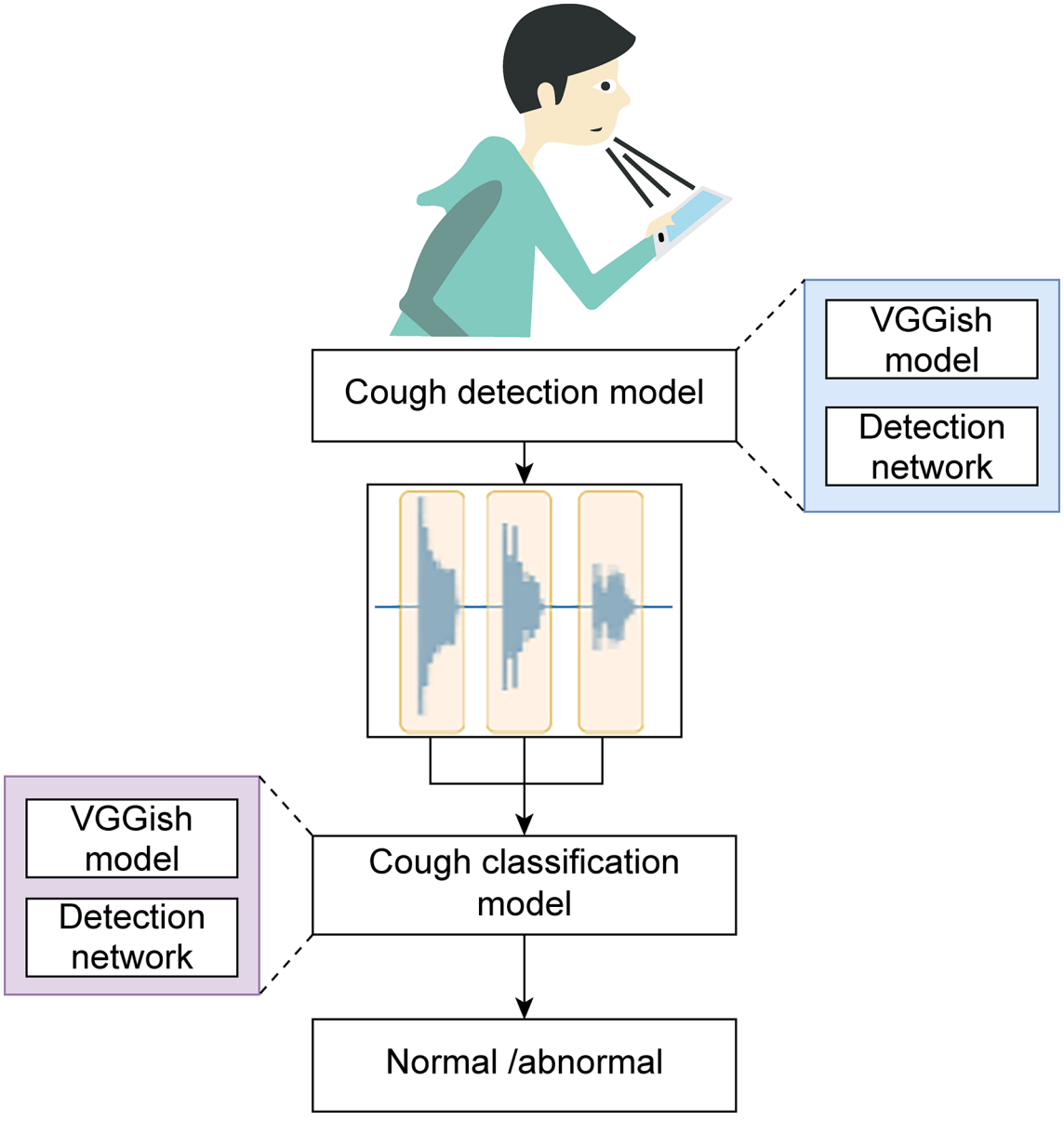


**Fig. 9** Overall service flow

**Abbreviations**

| | |
|---|---|
| AI | Artificial Intelligence |
| AUPRC | Area Under the Precision-Recall Curve |
| AUROC | Area Under the Receiver Operating Characteristic Curve |
| CNN | Convolutional Neural Network |
| COPD | Chronic Obstructive Pulmonary Disease |
| FN | False Negatives |
| FP | False Positives |
| Hz | Hertz |
| ML | Machine Learning |
| RNN | Recurrent Neural Network |
| TN | True Negatives |
| TP | True Positives |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03065-w.

Supplementary Material 1

Supplementary Material 2

**Acknowledgements**

We would like to thank Editage (www.editage.co.kr) for English language editing.

**Author contributions**

Sanghoon Han: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft; Yu-rim Lee: Methodology, Software; Ji-ho Lee: Validation, Investigation; JinHee Jeon: Resources; Choongki Min: Software, Formal analysis; Kyungnam Kim: Software; Donghoon Kim: Supervision; Myung Pyo Kim: Investigation; Young Mi Park: Validation, Investigation; Uiri An: Validation, Investigation, Resources; Kyoung Min Moon: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Validation, Investigation, Resources.

**Funding**
This study was supported by research grant from Biomedical Research Institute, Chung-Ang University Hospital (2024).

**Data availability**
The cough audio dataset analyzed in this study was ethically collected with approval from the Ethics Committee at Gangneung Asan Hospital (approval number: GNAH 2021-08-002-001). Given that individual cough recordings could potentially identify participants, full public dissemination of the dataset and original source code is restricted by ethical and privacy considerations mandated by the IRB. However, de-identified subsets of the data and the custom-developed code utilized for model training, validation, and evaluation can be made available upon reasonable request from qualified researchers. Access requests should be directed to the corresponding author and are subjected to review and approval by the relevant institutional ethics committees.

## Declarations

**Ethics approval and consent to participate**
This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee at Gangneung Asan Hospital (approval number: GNAH 2021-08-002-001). Informed consent was obtained from all participants, who signed a form authorizing the anonymous use of their clinical data for research, as approved by the Ethics Committee.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Waycen Inc, Seoul 06167, Republic of Korea
[2]Department of Internal Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
[3]Abler Private Clinic, Ulsan, Republic of Korea
[4]Division of Pulmonary and Allergy Medicine, Department of Internal Medicine, Chung-Ang University Hospital, Chung-Ang University College of Medicine, Seoul, Republic of Korea
[5]Department of Pulmonology, Hyundae Hospital, Namyangju, Korea
[6]Department of Pediatrics, Gangnam Severance Hospital, Yonsei University College of Medicine, Eonjuro Gangnamgu, Seoul, Korea
[7]Department of Internal Medicine, Armed Forces Capital Hospital, Seongnam, Korea
[8]Biomedical Research Institute, Chung-Ang University Hospital, Seoul, Korea
[9]Division of Pulmonary and Allergy Medicine, Department of Internal Medicine, Biomedical Research Institute, Chung-Ang University Hospital, Chung-Ang University College of Medicine, 102 Heukseok-ro, Dongjak-gu, Seoul 06973, Republic of Korea

Received: 26 September 2024 / Accepted: 9 June 2025
Published online: 01 July 2025

**References**

1. Chui KT, Alhalabi W, Pang SSH, de Pablos PO, Liu RW, Zhao M. Disease diagnosis in smart healthcare: innovation, technologies and applications. Sustainability. 2017;9:2309. https://doi.org/10.3390/su9122309.
2. Islam MM, Rahaman A, Islam MR. Development of smart healthcare monitoring system in Iot environment. SN Comput Sci. 2020;1:185. https://doi.org/10.1007/s42979-020-00195-y.
3. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230–43. https://doi.org/10.1136/svn-2017-000101.
4. Kumar A, Gupta PK, Srivastava A. A review of modern technologies for tackling COVID-19 pandemic. Diabetes Metab Syndr. 2020;14:569–73. https://doi.org/10.1016/j.dsx.2020.05.008.
5. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and Covid-19. Nat Med. 2020;26:459–61. https://doi.org/10.1038/s41591-020-0824-5.
6. Zang Y, Zhang F, Di C-a, Zhu D. Advances of flexible pressure sensors toward artificial intelligence and health care applications. Mater Horiz. 2015;2:140–56. https://doi.org/10.1039/C4MH00147H.
7. Raghavendra U, Acharya UR, Adeli H. Artificial intelligence techniques for automated diagnosis of neurological disorders. Eur Neurol. 2019;82(1–3):41–64. https://doi.org/10.1159/000504292.
8. Mori Y, Kudo S-E, Mohmed HEN, Misawa M, Ogata N, Itoh H, et al. Artificial intelligence and upper Gastrointestinal endoscopy: current status and future perspective. Dig Endosc. 2019;31(4):378–88. https://doi.org/10.1111/den.13317.
9. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69:2657–64. https://doi.org/10.1016/j.jacc.2017.03.571.
10. Catherwood PA, Rafferty J, McLaughlin J. Artificial intelligence for long-term respiratory disease management. Electronic Workshops in Computing. Electronic Workshops in Computing British HCI Conference. 2018. https://doi.org/10.14236/ewic/HCI2018.65
11. Mart´ın-Isla C, Asadi-Aghbolaghi M, Gkontra P, Campello VM, Escalera S, Lekadir K. Stacked BCDU-net with semantic CMR synthesis: application to myocardial pathology segmentation challenge. In: Myocardial pathology segmentation combining multi-sequence cardiac magnetic resonance images: first challenge, MyoPS 2020, held in conjunction with MICCAI 2020, Proceedings. Springer; 2020. pp. 1–16.
12. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of Covid-19. PLoS ONE. 2020;15:e0235187. https://doi.org/10.1371/journal.pone.0235187.
13. Palaniappan R, Sundaraj K, Ahamed NU. Machine learning in lung sound analysis: a systematic review. Biocybern Biomed Eng. 2013;33:129–35. https://doi.org/10.1016/j.bbe.2013.07.001.
14. Chambres G, Hanna P, Desainte-Catherine M. Automatic detection of patient with respiratory diseases using lung sound analysis. In: International Conference on Content-Based Multimedia Indexing (CBMI). IEEE; 2018. pp. 1–6. https://doi.org/10.1109/CBMI.2018.8516489
15. Palaniappan R, Sundaraj K, Sundaraj S. Artificial intelligence techniques used in respiratory sound analysis–a systematic review. Biomed Tech (Berl). 2014;59:7–18. https://doi.org/10.1515/bmt-2013-0074.
16. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. Digit Biomark. 2021;5:78–88. https://doi.org/10.1159/000515346.
17. Hee HI, Balamurali B, Karunakaran A, Herremans D, Teoh OH, Lee KP, et al. Development of machine learning for asthmatic and healthy voluntary cough sounds: a proof of concept study. Appl Sci. 2019;9:2833. https://doi.org/10.3390/app9142833.
18. Liao S, Song C, Wang X, Wang Y. A classification framework for identifying bronchitis and pneumonia in children based on a small-scale cough sounds dataset. PLoS ONE. 2022;17(10):e0275479.
19. Habashy K, Vald´es J, Cohen-McFarlane M, Xi P, Wallace B, Goubran R, Knoefel F. Cough classification using audio spectrogram transformer, in 2022 IEEE Sensors Applications Symposium (SAS), pp. 1–6, IEEE, 2022.
20. Balamurali B, Hee HI, Kapoor S, Teoh OH, Teng SS, Lee KP, Herremans D, Chen JM. Deep neural network-based respiratory pathology classification using cough sounds. Sensors. 2021;21(16):5555.
21. Islam R, Chowdhury NK, Kabir MA. Robust covid-19 detection from cough sounds using deep neural decision tree and forest: A comprehensive cross-datasets evaluation. ArXiv Preprint arXiv:2501.01117, 2025.
22. Chowdhury NK, Kabir MA, Rahman MM, Islam SMS. Machine learning for detecting covid-19 from cough sounds: an ensemble based Mcdm method. Comput Biol Med. 2022;145:105405.
23. Kadambi P, Mohanty A, Ren H, Smith J, McGuinnes K, Holt K et al. Towards a wearable cough detector based on neural networks. In: IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. pp. 2161-65. https://doi.org/10.1109/ICASSP.2018.8461394
24. Larson S, Comina G, Gilman RH, Tracey BH, Bravard M, López JW. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. PLoS ONE. 2012;7:e46229. https://doi.org/10.1371/journal.pone.0046229.
25. Shi L, Du K, Zhang C, Ma H, Yan W. Lung sound recognition algorithm based on VGGish-BiGRU. IEEE Access., Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90. https://doi.org/10.1145/3065386

26. Simonyan K, Zisserman A. 'Very deep convolutional networks for large-scale image recognition,' arXiv preprint arXiv:1409.1556; 2014.
27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE; 2016. pp. 2818-26. https://doi.org/10.1109/CVPR.2016.308
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE; 2016. pp. 770-8. https://doi.org/10.1109/CVPR.2016.90
29. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC et al. CNN architectures for large-scale audio classification. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP). IEEE. 2017:131-5. https://doi.org/10.1109/ICASSP.2017.7952132
30. Oh Y, Park S, Ye JC. Deep learning Covid-19 features on Cxr using limited training data sets. IEEE Trans Med Imaging. 2020;39:2688–700. https://doi.org/10.1109/TMI.2020.2993291.
31. Fan D-P, Zhou T, Ji G-P, Zhou Y, Chen G, Fu H, et al. Inf-net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans Med Imaging. 2020;39:2626–37. https://doi.org/10.1109/TMI.2020.2996645.
32. Pereira RM, Bertolini D, Teixeira LO, Silla CN Jr, Costa YMG. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed. 2020;194:105532. https://doi.org/10.1016/j.cmpb.2020.105532.
33. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Singh V. Application of deep learning for fast detection of Covid-19 in x-rays using Ncovnet. Chaos Solitons Fractals. 2020;138:109944. https://doi.org/10.1016/j.chaos.2020.109944.
34. El Asnaoui K, Chawki Y. Using x-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn. 2021;39:3615–26. https://doi.org/10.1080/07391102.2020.1767212.
35. Alqudaihi KS, Aslam N, Khan IU, Almuhaideb AM, Alsunaidi SJ, Ibrahim NMAR, et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. IEEE Access. 2021;9:102327–44. https://doi.org/10.1109/ACCESS.2021.3097559.
36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. IEEE; 2017. pp. 618 – 26. https://doi.org/10.1109/ICCV.2017.74
37. 2023 gold report, global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease [2023 report]. Global Initiative for Asthma; 2023.
38. 2023. gina report, global strategy for asthma management and prevention. Global Initiative for Asthma; 2023.
39. Mackenzie G. The definition and classification of pneumonia. Pneumonia (Nathan). 2016;8:14. https://doi.org/10.1186/s41479-016-0012-z.
40. Zimmer AJ, Ugarte-Gil C, Pathri R, Dewan P, Jaganath D, Cattamanchi A, et al. Making cough count in tuberculosis care. Commun Med (Lond). 2022;2:83. https://doi.org/10.1038/s43856-022-00149-w.
41. Ni X, Ouyang W, Jeong H, Kim J-T, Tzaveils A, Mirzazadeh A, et al. Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for Covid-19 patients. Proc Natl Acad Sci U S A. 2021;118:e2026610118. https://doi.org/10.1073/pnas.2026610118.
42. Laguarta J, Hueto F, Subirana B. Covid-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol. 2020;1:275–81. https://doi.org/10.1109/OJEMB.2020.3026928.
43. Korpáš J, Sadloňová J, Vrabec M. Analysis of the cough sound: an overview. Pulm Pharmacol. 1996;9:261–8. https://doi.org/10.1006/pulp.1996.0034.
44. Global strategy for the. diagnosis, management, and prevention of chronic obstructive pulmonary disease. Global Initiative for Chronic Obstructive Lung Disease; 2023.
45. Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC et al. Audio Set: an ontology and human-labeled dataset for audio events; In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2017. pp. 776 – 80. https://doi.org/10.1109/ICASSP.2017.7952261
46. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y. A relationship between the incremental values of area under the Roc curve and of area under the precision-recall curve. Diagn Prognostic Res. 2021;5:1–15.
47. Srinivas S, Fleuret F. Full-gradient representation for neural network visualization. In: Wallach H, Larochelle H, Beygelzimer A, dAlché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems. Curran Associates, Inc.; 2019.


## Publisher's note