EasyChair Preprint

№ 10651

CASCO: a Contactless Cough Screening System based on Audio Signal Processing

Xinxin Zhang, Hang Liu, Xinru Chen, Rui Qin, Yan Zhu, Wenfang Li, Menghan Hu and Jian Zhang

August 2, 2023

CGI2023 manuscript No.
(will be inserted by the editor)

CASCO: A Contactless Cough Screening System based on Audio Signal Processing

Xinxin Zhang[†] · **Hang Liu**[†] · **Xinru Chen** · **Rui Qin** · **Yan Zhu** · **Wenfang Li** · **Menghan Hu**[*] · **Jian Zhang**

Abstract Cough is a common symptom of respiratory disease, which produces a specific sound. Cough detection has great significance to prevent, assess, and control epidemics. This paper proposes CASCO (Cough Analysis System using Short-Time Fourier Transform (STFT) and Convolutional Neural Networks (CNN) in the WeChat mini Program), a cough detection system capable of quantifying the number of coughs through an audio division algorithm. This system combines STFT with CNN, achieving accuracy, precision, recall, and F1-score with 97.0%, 95.6%, 98.7%, and 0.97 respectively in cough detection. The model is embedded into the WeChat mini program to make it feasible to apply cough detection on smartphones and realize large-scale and contactless cough screening. Future research can combine audio and video signals to further improve the accuracy of large-scale cough screening.

Keywords Cough detection · Deep neural network · Audio Signal Processing

1 Introduction

Cough is a common symptom associated with various respiratory diseases such as bronchitis, and asthma. It serves as a powerful mechanism of the human body to expel foreign particles and clear secretions from the upper respiratory tract, resulting in a specific sound that plays a significant role in disease diagnosis [1].

Xinxin Zhang, Hang Liu, Xinru Chen, Rui Qin, Menghan Hu, Jian Zhang
Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, China.

Yan Zhu, Wenfang Li
Department of Emergency and Critical Care, Shanghai Changzheng Hospital, China.
Corresponding author: Menghan Hu (mhhu@ce.ecnu.edu.cn).
[†]These authors contributed equally to this work.

Respiratory diseases pose a significant threat to human health worldwide [2]. As the details about the cough frequency, intensity, and sound help physicians in their diagnostics of respiratory diseases, the field of automatic cough sensor research has been established with various systems achieving high precision and sensitivity. The Leicester cough monitor consists of an audio recorder and a microphone and detects the time-varying spectral features of cough sound based on hidden Markov models [3]. Costa et al. [4] applied a mechanomyography sensor on the abdominal region to detect cough events. Doddabasappla et al. [5] achieved cough detection using the multiband spectral summation features of acceleration signal measured by a portable accelerometer. Most of these automatic cough sensors are considered to be uncomfortable to wear during daily activities, and their expensive cost hindered the application of large-scale cough screening.

Large-scale cough detection plays a pivotal role in epidemiological research, disease screening, and epidemic control efforts. The global impact of the COVID-19 pandemic, which primarily affects the respiratory system, has resulted in a significant number of confirmed cases worldwide, as reported by the World Health Organization (WHO) by March 2, 2022 [6]. The emergence and dominance of the Omicron variant, along with other COVID-19 variants, have further exacerbated the health crisis and posed immense challenges to human health and the global economy [7]. Given that cough is a prominent symptom of COVID-19, there has been a growing focus on developing large-scale and contactless cough detection systems in research initiatives. These systems hold immense potential in facilitating early detection, monitoring, and effective control of infectious diseases. By enabling non-invasive and convenient screening, they can contribute significantly to mitigating disease transmission, informing public health strategies, and supporting timely interventions. The development and implementation of such systems are crucial steps toward safeguarding public health

and minimizing the impact of future epidemics and pandemics.

In response to the distinct acoustic characteristics associated with cough events, several studies have made significant strides in developing models that utilize cough audio features for remote and contactless detection of cough events. For instance, Islam et al. proposed an algorithm that leverages acoustic features extracted from cough sound samples, combined with a deep neural network, for automated and noninvasive diagnosis of COVID-19 [8]. Tena et al. focused on extracting time-frequency cough features from audio signals and applied a supervised machine-learning algorithm to identify the most relevant features for COVID-19 diagnosis [9]. Another notable study by Monge-Álvarez et al. involved the construction of a machine hearing system specifically designed for robust cough detection, incorporating short-term spectral features and the standard deviation of short-term descriptors [10]. These advancements demonstrate the potential of using cough audio analysis in developing efficient and accurate diagnostic tools, providing valuable insights for the detection and management of respiratory diseases.

The widespread use of smartphones has made large-scale cough detection possible. Patients and medical professionals now have the convenience of capturing cough audio signals using the built-in microphone and voice recorder on their smartphones, eliminating the need for additional specialized cough assessment devices. Notably, Hoyos-Barcelo et al. proposed a cough detector on smartphones that leverages local Hu moments as robust features, combined with an optimized k-NN classifier [11]. Imran et al. built a COVID-19 diagnosis app analyzing cough sound by an Artificial Intelligence (AI)-based engine [12]. Most of the aforementioned studies utilizing AI algorithms failed to accurately measure cough frequency and intensity.

One significant challenge in utilizing smartphones for cough detection is the limited battery consumption of these devices. Executing complex machine learning or deep learning algorithms directly on smartphones can quickly drain the battery and hinder their practicality. Consequently, alternative approaches are necessary to overcome this limitation and enable efficient cough analysis [13, 14]. One feasible solution is to leverage the capabilities of external servers for audio signal processing. With the widespread application of 5G technology, it becomes increasingly feasible to use smartphones solely for audio collection while offloading computationally intensive tasks to remote servers [2]. The high-speed and low-latency characteristics of 5G networks facilitate seamless and efficient transmission of cough audio signals from smartphones to external servers.

In this paper, we propose a novel cough detection system called CASCO, which can calculate the number of coughs by an audio division algorithm. This system combines STFT with CNN, achieving an impressive accuracy rate of 97.0% in classifying cough sounds and non-cough sounds. The integration of this system into the WeChat mini program enables the deployment of cough detection on smartphones, enabling widespread and contactless screening for coughs at a large scale. Furthermore, by processing the audio recorded on smartphones in an external server, we alleviate the issue of high battery consumption associated with complex algorithms, ensuring a smoother user experience. The subsequent sections of this paper are organized as follows: In Section 2, we outline the methodology and the specific procedure of cough detection system. In Section 3, we explain the dataset we used and the training process, compare the experimental results in terms of performance metrics, and discuss the potential and limitations of the study. We conclude the paper by summarizing the key findings and contributions of the research.

2 Cough Detection System

The overall system architecture is shown in Fig. 1. The WeChat mini program in a smartphone records sound when the "Start recording" and "Finish recording" buttons are pressed. When the "Detect cough" button is pressed, the recorded sounds are transmitted to the server for further processing. At the server, the audio division algorithm extracts the high parts above the threshold from a long piece of audio. This process divides the long audio into shorter segments, each containing only a single suspicious sound. Then STFT is applied to the short audio to generate a spectrogram that serves as the feature of the audio. Subsequently, the spectrogram is forwarded to CNN, classifying cough samples and non-cough samples. The server performs cough detection and counts the number of coughs in the long audio. Finally, the output results are displayed in the WeChat mini program for user accessibility.

The details of detection and diagnosis classifiers are presented below.

2.1 Audio Division

To select a suitable threshold for extracting a single suspicious sound, we use Otsu's thresholding method in the audio division algorithm. Otsu's method is a global thresholding algorithm, which can automatically generate the optimal segmentation threshold based on the input signal [15]. For the input audio signal, we suppose the number of points is denoted as N, which are dichotomized into two classes: the low part C_0 and the high part C_1, using a threshold at level T. The proportion of points belonging to the low part in the whole audio is denoted by ω_0 and its average amplitude level is μ_0. Similarly, the proportion of points belonging to the

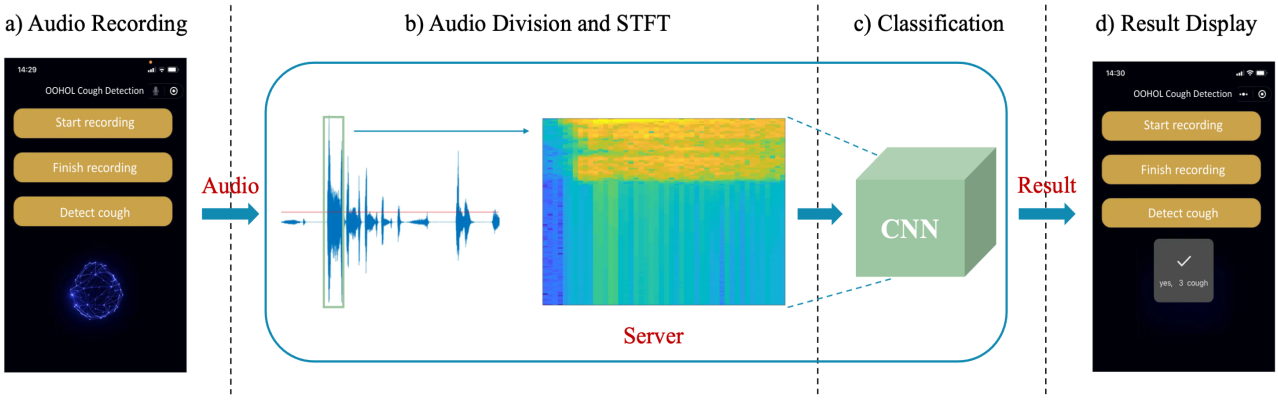

Fig. 1: Pipeline of the CASCO cough detection system: a) record a long piece of audio to be detected via the WeChat mini program; b) divide the long audio into several short sounds and extract features using STFT; c) classify cough sounds and non-cough sounds through CNN; d) display the result and the number of coughs on the WeChat mini program.

high part in the whole audio is denoted by ω_1 and its average amplitude level is μ_1. Then the total average amplitude level of the audio is given by:

$$\mu_T = \omega_0\mu_0 + \omega_1\mu_1 \tag{1}$$

We can easily verify the following relation for any choice of T:

$$\omega_0 + \omega_1 = 1 \tag{2}$$

To evaluate the class separability of the threshold at level T, we introduce the following between-class variance used in the discriminant analysis:

$$\begin{aligned}\sigma_B^2 &= \omega_0\left(\mu_0 - \mu_T\right)^2 + \omega_1\left(\mu_1 - \mu_T\right)^2 \\ &= \omega_0\omega_1\left(\mu_1 - \mu_0\right)^2\end{aligned} \tag{3}$$

In equation (3), it can be observed that the farther the two means μ_0 and μ_1 are from each other, the larger the between-class variance is, which indicates that the between-class variance serves as an effective measure of differentiability between classes. To determine the optimal threshold T^* that maximizes the between-class variance, we employ the following equation:

$$\sigma_B^2\left(T^*\right) = \max \sigma_B^2(T) \tag{4}$$

After selecting a suitable threshold, we extract the high parts above the threshold from a long piece of audio. This process involves splitting the long audio into shorter segments, each containing only a single suspicious sound. The length of the short segments is not fixed but adaptively determined based on the threshold and characteristics of the audio signal. It is expected to implement the function of counting the number of coughs. The processing steps of the audio division algorithm can be seen in Fig. 2.

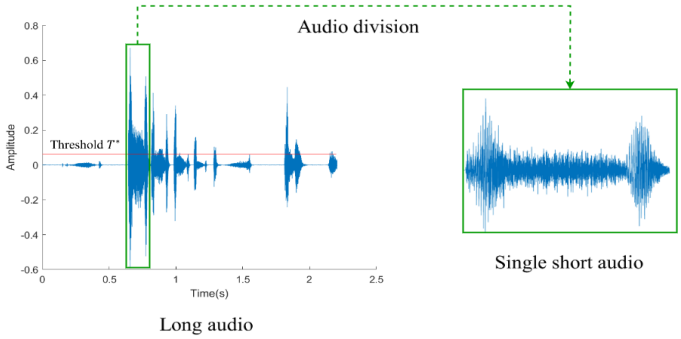

Fig. 2: Processing steps of the audio division algorithm.

2.2 Feature Extraction

For automatic speech recognition, STFT has been considered to be an effective feature extraction method. The spectrogram generated by STFT can show the relationship of time and frequency of audio signals, thus extracting features of audio and then differentiating different audio signals [16]. The process of STFT is illustrated in Fig. 3. The STFT form of signal $x(t)$ can be defined by the following equation:

$$\mathrm{STFT}(t,f) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)\mathrm{e}^{-j2\pi f\tau}\mathrm{d}\tau \tag{5}$$

where $h(\tau - t)$ means window function.

The audio signal is first pre-emphasized by a first-order high-pass filter to improve the signal-to-noise ratio in the high-frequency portion of the signal. After the audio signal is framed and windowed, Fast Fourier Transform (FFT) is applied to all the frames and generates spectrums. The amplitude values of the spectra are quantified and mapped to different colors, providing a visual representation of the fre-

quency content. Finally, the transformed multi-frame spectrums are stitched together in the time dimension to form a final spectrogram of the audio signal.

We apply the STFT to the short audio to generate a spectrogram with time on the horizontal axis, frequency on the vertical axis and color indicating amplitude as the feature of the audio. The spectrogram generated through STFT enables visual representation and facilitates accurate differentiation of cough sounds.

2.3 Classification

Mapping the time-frequency spectrogram into a color representation as the input to a CNN serves two main purposes: 1) Channel dimensions: By converting the time-frequency spectrogram to a color image with three channels (R, G, B), it aligns with the common CNN input format, such as RGB images. This allows the CNN to process the spectrogram as image-like input, with height, width, and three color channels, enabling the use of standard image-based CNN architectures. CNNs excel in learning hierarchical features, progressing from local patterns to global representations. 2)Feature representation: The STFT output consists of the real and imaginary parts (or magnitude and phase), representing different aspects of the audio signal. By mapping it to a color spectrogram, the CNN can potentially learn distinct features from different parts of the spectrogram. For instance, the CNN can capture spatial patterns, temporal changes, and frequency content from the color representation, leading to a more comprehensive and distinctive feature representation for cough detection.

The generated spectrogram is then fed into the CNN to decide whether the audio corresponds to a cough or not. An overview of the used CNN structure is illustrated in Fig. 4. The CNN consists of eight layers: 5 convolutional layers, 2 fully connected layers, and a softmax classification layer. In each convolutional layer, the Rectified Linear Unit (ReLU) is utilized as the activation function. The first, second, and fifth convolutional layers are connected to a 3×3 max-pooling layer, which is performed with a stride of 2. The first convolutional layer takes in the $224\times224\times3$ spectrogram as inputs and consists of 96 filters of kernel size 11×11, a stride of 4, and padding of 2. It is followed by a 5×5 convolutional layer with a padding of 2. The last three convolutional layers all have filters of size 3×3 and padding of 1. The features are then passed to two fully connected layers with 4,096 neurons each, which also employ 0.5 dropout regularization to avoid overfitting. Finally, the last layer, comprising two neurons, takes the outputs from the second fully connected layer and employs the softmax function to classify the spectrograms as either cough or non-cough. By utilizing this CNN architecture, we aim to capture and learn the distinguishing patterns and characteristics of cough events, enabling accurate classification of cough and non-cough spectrograms. The combination of convolutional layers, max-pooling layers, fully connected layers, and the softmax classification layer provides the necessary capacity for CNN to effectively differentiate between cough and non-cough audio samples.

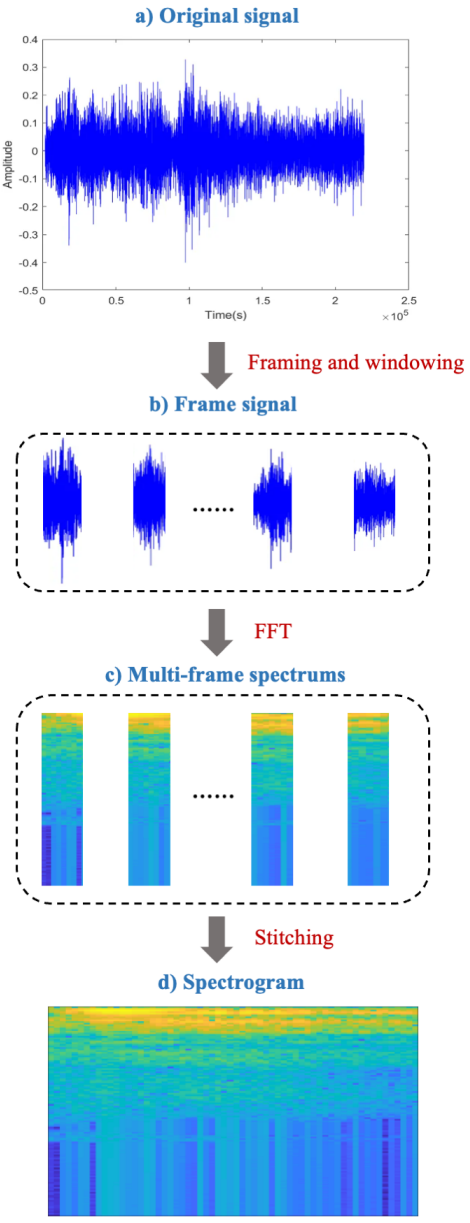

Fig. 3: Processing steps of generating spectrogram by STFT: a) the amplitude-time graph of the original signal; b) the separated signals after framing and windowing; c) multi-frame spectrums with amplitude values mapped into colors; d) the final spectrogram.

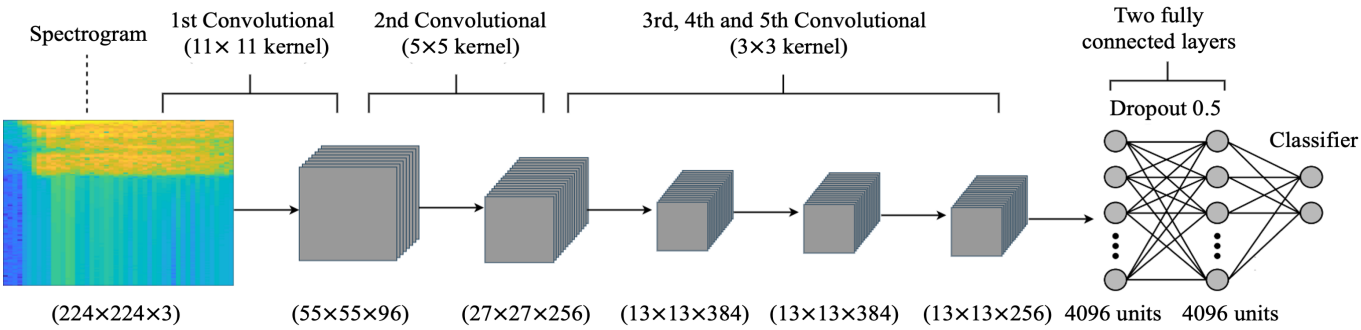

Fig. 4: Structure of the used CNN classifier: the network consists of eight layers, including five convolutional layers, two fully connected layers and a softmax classification layer.

3 Experiments

3.1 Dataset Explanation

To train and build the proposed system, we create a dataset consisting of cough samples and non-cough samples. The dataset of cough samples comprises self-recorded coughing audio, cough recordings sourced from the Environmental Sound Classification (ESC-50) dataset [17], and inpatient cough recordings collected from individuals diagnosed with respiratory diseases. The inpatient cough recordings were collected from 24 patients, including 15 males and 9 females, in the respiratory disease department at Ruijin Hospital [18]. The inpatient's cough samples ranged in age from 48 to 85 years old and suffered from respiratory diseases with symptoms of cough. The non-cough samples contain self-recorded environmental audio and labeled environmental recordings from the ESC-50 dataset, including interior sounds, exterior noises, natural sounds, and human (non-speech) sounds [17]. Both the self-recorded samples and inpatient cough recordings were recorded using mobile phone microphones, incorporating background noise in these samples.

Before training, we apply the audio division algorithm to split the long audio from the previously collected dataset into shorter audio clips, each containing a single suspicious sound. The preprocessed audio has durations ranging from 230ms to 670ms. In total, we obtain 13,529 samples, comprising 6,985 cough samples and 6,544 non-cough samples. To facilitate the training process, we divide the dataset into a training dataset and a testing dataset. The training dataset consists of 5,588 cough samples and 5,235 non-cough samples, while the test dataset comprises 1,397 cough samples and 1,309 non-cough samples.

3.2 Network Training

The convolutional network employed in this study is trained using the Adam optimizer, which is a first-order gradient-based optimization algorithm for stochastic objective functions [19]. The Adam optimizer leverages adaptive estimates of lower-order moments, making it well-suited for handling large datasets and sparse gradients. To train the convolutional network, the cross entropy loss function is utilized. This loss function is commonly employed in classification tasks and measures the dissimilarity between the predicted probabilities and the true labels. An initial learning rate of 0.0002 is set, allowing the network to gradually adjust its weights based on the optimization process.

During the training process, a batch size of 32 is utilized. The batch size determines the number of samples processed in each iteration, allowing for efficient utilization of computational resources and improved generalization performance. By carefully selecting these hyperparameters and leveraging the capabilities of the Adam optimizer, we optimize the performance of the convolutional network and enhance its ability to accurately classify cough and non-cough samples.

3.3 Experimental Results

To evaluate and compare the performance of the system in cough detection, two experiments are undertaken.

In the initial experiment, we assess the system's performance using True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) values obtained from the confusion matrix presented in Table 1. From this matrix, we derive several performance metrics including accuracy, precision, sensitivity/recall, and F1-score. These metrics provide valuable insights into the effectiveness and reliability of the model when applied to the test dataset. The calculation of these metrics is as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{sensitivity/recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-score} = 2 * \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \tag{9}$$

Table 1: Confusion matrix for cough detection.

True Class	**Predicted Class**	
	Cough	**Non-cough**
Cough	TP: 1379	FN: 18
Non-cough	FP: 63	TN: 1246

Based on the classification results presented in Table 2, the cough detection model demonstrates excellent performance in distinguishing between cough events and non-cough samples. The accuracy of the model is measured at 97.0%, indicating a high level of overall correct classification. Additionally, the recall (also known as sensitivity) is calculated at 98.7%, which signifies the model's ability to correctly identify the majority of actual cough events. The precision of the model stands at 95.6%, indicating the proportion of correctly identified cough events among the total number of predicted cough events. Moreover, the F1-score, which combines both precision and recall, is calculated as 0.97. The F1-score is a measure that balances the trade-off between precision and recall, providing an overall assessment of the model's performance. These results demonstrate its potential as a reliable tool for large-scale and contactless cough screenings.

In the second experiment, we conduct a comprehensive comparative analysis to assess the effectiveness of the cough detection model in comparison to machine learning and other deep learning algorithms for cough classification. Considering the widely-used Mel-Frequency Cepstral Coefficients (MFCC) as classical speech recognition features, we construct a Support Vector Machine (SVM) model using MFCC features to examine how a more complex machine learning classifier performs in comparison to the CNN-based model. Additionally, we train a model using the softmax (SM) function on the MFCC features to directly compare with CNN's classification layer. Furthermore, we build comparative deep learning models utilizing ResNet [20] and VGG16 [21] architectures.

The performance metrics of these comparative models are summarized in Table 2, providing insights into their respective accuracy, recall, precision, and F1-score. The results clearly demonstrate that the cough detection model outperforms the SVM model employing MFCC features, exhibiting a remarkable improvement of 2.1% in terms of accuracy.

The main reasons for the superiority of STFT+CNN over other combinations lie in its feature representation, deep learning architecture, and data representation. STFT extraction captures rich time-frequency information, providing a comprehensive description of cough sounds, whereas MFCC only considers Mel frequency information, and SM and VGG16 may not effectively utilize time-frequency information. CNN, designed for image processing, excels in handling color spectrograms and can learn spatial, temporal, and frequency features, facilitating accurate cough sound classification and detection. The color spectrogram leverages the three channels to enhance feature diversity, while other combinations might not fully exploit audio data characteristics. This comparative experiment highlights the superiority of the CNN-based approach over traditional machine learning algorithms and other deep learning architectures in accurately detecting and classifying cough sounds.

3.4 Discussion

The remarkable accuracy achieved by the cough detection model, surpassing that of the comparative models, clearly demonstrates the immense potential of the system. However, the system's capabilities extend beyond accurate detection. It also offers the ability to count the number of coughs through the implementation of the audio division algorithm, providing valuable information such as cough frequency and intensity. This has various applications, including respiratory health monitoring, identifying cough outbreaks, and tracking medical interventions. For instance, in a healthcare setting, the system could be deployed in hospitals or clinics to monitor the cough frequency and intensity of patients with respiratory conditions, enabling healthcare professionals to gain insights into the severity and progression of their conditions. One of the key advantages of the system is its integration with a WeChat mini program, enabling the implementation of cough detection on smartphones and facilitating large-scale, contactless cough screenings.

By processing the audio recordings on an external server, the issue of high battery consumption typically associated with continuous audio processing on mobile devices is mitigated. This opens up a wide range of possibilities for deploying the system in various public settings, including hos-

Table 2: Comparison of SM, SVM, CNN, ResNet, and VGG for cough classification.

Model	Accuracy (%)	Sensitivity/Recall (%)	Precision (%)	F1-Score
MFCC+SM	85.7	87.4	83.9	0.86
MFCC+SVM	94.9	97.1	93.1	0.95
STFT+CNN	**97.0**	**98.7**	**95.6**	**0.97**
STFT+ResNet	94.2	95.6	92.9	0.94
STFT+VGG16	95.4	97.8	94.1	0.95

pital wards, subway stations, and classrooms, where monitoring the frequency of coughs is essential. With the ability to accurately assess coughing incidents, the system can contribute to proactive measures in maintaining public health and safety. Furthermore, the non-intrusive nature of the system, coupled with its ease of deployment, allows for efficient monitoring and analysis of coughing patterns in real time. This information can aid in identifying potential outbreaks, tracking the effectiveness of preventive measures, and providing early warnings in situations where the spread of respiratory illnesses is a concern.

The performance of the cough detection model is subject to certain limitations, which we acknowledge and aim to address in future improvements. Two key factors affecting model performance are given below:

Feature extraction method: In real-world environments, noise poses a challenge to system accuracy, especially in cases of confusion between cough and speech sounds. Low-amplitude cough signals may be masked or overlooked by high-amplitude background noise, affecting threshold selection and segmentation. To overcome these limitations, future research can employ novel noise suppression techniques to reduce the impact of background noise on threshold selection. Utilizing multiple Otsu's thresholds for multi-scale analysis can detect cough signals with different amplitudes, reducing the likelihood of missing low-amplitude cough signals [22]. Data augmentation by adding various noise and low-amplitude cough signals can enhance the model's adaptability to different audio conditions and improve cough signal detection accuracy. Moreover, considering the fusion of other signals, such as video or sensor data, can provide comprehensive information about cough events, aiding in more accurate cough signal detection and enhancing the model's robustness to various types of background noise.

Furthermore, the distance between the smartphone and the user during cough sound recording can affect the volume of the recorded signal. If users are at a considerable distance from the smartphone while recording cough sounds, the cough volume may be lower, resulting in the potential masking or reduced detectability of cough signals, especially in environments with higher background noise. To address this issue, multiple microphones or microphone arrays can be employed to capture sound from different angles, and adaptive volume control or dynamic gain adjustment techniques can be introduced in the system. These measures ensure that cough sounds can be effectively captured under various distances and environmental noise conditions, thereby enhancing the reliability and robustness of cough signal detection.

Additionally, we will continue to refine the models by incorporating new feature extraction methods and exploring advanced deep learning architectures. By combining multiple feature extraction techniques, such as MFCCs and other spectral or temporal features, we can capture a broader range of characteristics related to cough events. This will help us improve the discrimination between coughing and other sounds, further enhancing the precision and reliability of the system.

Limited types of signals collected: The current system solely relies on audio signals for cough detection. However, in certain situations where cough waveforms densely overlap, distinguishing individual cough events becomes challenging, leading to inaccurate cough counting. To enhance system performance, it is crucial to incorporate additional signals. Coughing is often accompanied by specific movements and physical cues, which can be valuable in understanding coughing events comprehensively. By incorporating additional signals, we can gain a more comprehensive understanding of coughing events. The integration of motion or video data can provide valuable insights into the physical manifestations of coughing, such as body movements, hand gestures, or facial expressions. These cues can contribute to more accurate and reliable detection of cough events, reducing both false negatives and false positives.

In future research, we will explore the fusion of audio and image sequence data to develop a more robust and comprehensive cough detection system. By leveraging the complementary nature of these modalities, we aim to achieve even higher accuracy and reliability in detecting and analyz-

ing cough events, contributing to the advancement of large-scale and contactless cough screenings in various fields.

4 Conclusion

Cough detection plays a vital role in epidemiological research, disease screening, and epidemic control. In this paper, we present CASCO, an advanced cough detection system that combines CNN with a WeChat mini program. The system accurately detects cough events in real time and provides an automated count of the number of coughs. To train the robust cough detection model, we construct a comprehensive dataset comprising self-recorded audio samples, labeled environmental recordings from the ESC-50 dataset, and inpatient cough recordings from respiratory disease patients.

The WeChat mini program integrated into smartphones serves as the primary interface for the system, allowing users to record audio and view the cough detection results. The recorded audio is then processed on an external server using the sophisticated cough detection model. An audio division algorithm is employed to extract high-intensity segments from the audio, isolating individual cough events. The extracted segments are subsequently converted into spectrograms using STFT, capturing the distinctive time-frequency patterns of cough sounds. These spectrograms are then fed into the CNN model, which categorizes them as either cough or non-cough samples.

Extensive evaluations demonstrate the outstanding performance of the cough detection model, achieving an accuracy, recall, precision, and F1-score of 97.0%, 98.7%, 95.6%, and 0.97, respectively. The integration of the system with the WeChat mini program allows for large-scale and contactless cough screenings, overcoming the limitations of traditional detection methods. Additionally, processing audio on an external server reduces battery consumption while leveraging the server's computational power for faster and more accurate detection.

In future work, we will focus on improving the noise robustness of the model and exploring new application scenarios for the CASCO cough detection system. The goal is to develop a versatile and user-friendly solution that enhances public health monitoring through reliable and scalable cough detection.

References

1. McCool F D, Global physiology and pathophysiology of cough: ACCP evidence-based clinical practice guidelines. Chest, vol.129.1, p.48S-53S, 2006.
2. Alqudaihi K S, Aslam N, Khan I U, et al., Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. Ieee Access, vol.9, p.102327-102344, 2021.
3. Matos S, Birring S S, Pavord I D, et al., An automated system for 24-h monitoring of cough frequency: the leicester cough monitor. vol.54.8, p.1472-1479, 2007.
4. Costa T D, Nogueira-Neto G N, Nohama P, Cough detection through mechanomyographic signal in synchronized respiratory electrical stimulation systems. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.
5. Doddabasappla K, Vyas R, Spectral summation with machine learning analysis of tri-axial acceleration from multiple wearable points on human body for better cough detection. IEEE Sensors Letters, vol.5.9, p.1-4, 2021.
6. WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/.
7. Weekly epidemiological update on COVID-19 - 1 March 2022, https://www.who.int/publications/m/item/weeklyepidemiological-update-on-covid-19—1-march-2022.
8. Islam R, Abdel-Raheem E, Tarique M, A study of using cough sounds and deep neural networks for the early detection of COVID-19. Biomedical Engineering Advances, vol.3, p. 100025, 2022.
9. Alberto Tena, Francesc Clarià, Francesc Solsona, Automated detection of COVID-19 cough. Biomedical Signal Processing and Control, vol.71, p.103175, 2022.
10. Monge-Álvarez J, Hoyos-Barceló C, San-José-Revuelta L M, et al, A machine hearing system for robust cough detection based on a high-level representation of band-specific audio features. vol.66.8, p.2319-2330, 2019.
11. Hoyos-Barceló C, Monge-Álvarez J, Shakir M Z, et al, Efficient k-NN implementation for real-time detection of cough events in smartphones. IEEE Journal of Biomedical and Health Informatics, vol.22.5, p.1662-1671, 2018.
12. Imran A, Posokhova I, Qureshi H N, et al, AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked, vol.20, p.100378, 2020.
13. Agu E, Pedersen P, Strong D, et al, The smartphone as a medical device: Assessing enablers, benefits and challenges. 2013 IEEE International Conference on Sensing, Communications and Networking (SECON), 2013.
14. Chen Xinru, Hu Menghan, Zhai Guangtao, Cough Detection Using Selected Informative Features from Audio Signals. 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2021.
15. Nobuyuki Otsu, A threshold selection method from gray level histograms. IEEE Transactions on Systems, Man, and Cybernetics, vol.9, p.62-66, 1979.
16. Griffin D, Jae Lim, Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.32.2, p.236-243, 1984.
17. Piczak, Karol J, ESC: Dataset for environmental sound classification. Proceedings of the 23rd ACM international conference on Multimedia, 2015.
18. Jiang Zheng, Hu Menghan, Gao Zhongpai, Detection of respiratory infections using RGB-infrared sensors on portable device. IEEE Sensors Journal, vol.20.22, p.13674–13681, 2020.
19. Kingma D P, Ba J, Adam: A method for stochastic optimization. International Conference on Learning Representations, 2015.
20. He K, Zhang X, Ren S, et al, Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
21. Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, vol.1409.1556, 2015.
22. Wu, Zongwei and Allibert, Guillaume and Meriaudeau, Fabrice and Ma, Chao and Demonceaux, Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE Transactions on Image Processing, vol.32, p.2160–2173, 2023.