# Detecting COVID-19 from Respiratory Sound Recordings with Transformers

Idil Aytekin[a], Onat Dalmaz[a], Haydar Ankishan[b], Emine U Saritas[a], Ulas Bagci[c], Tolga Cukur[a], and Haydar Celik[d]

[a]Dept. of Electrical and Electronics Eng., Bilkent University, Ankara
[b]Vocational School of Technical Sciences, Baskent University, Ankara
[c]Dept. of Radiology and BME, Northwestern University
[d]Children's National Hospital, Washington, DC

## ABSTRACT

Auscultation is an established technique in clinical assessment of symptoms for respiratory disorders. Auscultation is safe and inexpensive, but requires expertise to diagnose a disease using a stethoscope during hospital or office visits. However, some clinical scenarios require continuous monitoring and automated analysis of respiratory sounds to pre-screen and monitor diseases, such as the rapidly spreading COVID-19. Recent studies suggest that audio recordings of bodily sounds captured by mobile devices might carry features helpful to distinguish patients with COVID-19 from healthy controls. Here, we propose a novel deep learning technique to automatically detect COVID-19 patients based on brief audio recordings of their cough and breathing sounds. The proposed technique first extracts spectrogram features of respiratory recordings, and then classifies disease state via a hierarchical vision transformer architecture. Demonstrations are provided on a crowdsourced database of respiratory sounds from COVID-19 patients and healthy controls. The proposed transformer model is compared against alternative methods based on state-of-the-art convolutional and transformer architectures, as well as traditional machine-learning classifiers. Our results indicate that the proposed model achieves on par or superior performance to competing methods. In particular, the proposed technique can distinguish COVID-19 patients from healthy subjects with over 94% AUC.

**Keywords:** COVID-19, respiratory, sound, breathing, cough, transformer

## 1. INTRODUCTION

Auscultation via hand-held or digital stethoscopes is a common first step in clinical examination of patients for signs of respiratory or cardiac disorders.[1] While it is a relatively old technique, it is still preferred due to its safety, non-invasiveness and low economic costs. That said, there are clinical scenarios in which a scheduled hospital visit might be difficult or where continuous monitoring of health signs might be critical.[2] A promising solution to this fundamental problem is the emergent use of mobile or wearable devices for continuous monitoring of health signs outside the clinic. These digital devices hold the potential to become a key component in healthcare by enabling automated screening of respiratory symptoms.[3]

Several recent studies have adopted machine learning algorithms for automated detection of respiratory diseases based on bodily sounds. In Ref. 4, the distinctness of pathomorphological alterations in the respiratory system induced by COVID-19 infection when compared to other respiratory infections was investigated based on cough sounds. An ensemble model containing support vector machine (SVM) and convolutional neural network (CNN) classifiers were trained to detect COVID-19 among other respiratory infections. In Ref. 5, cough sounds related to bronchitis, bronchiolitis and pertussis were examined. CNN classifiers were built to discriminate among the three disease categories. In Ref. 6, cough and whoop sounds during pertussis were examined. A logistic regression model with MFCC features was built. In Ref. 7, cough sounds in COVID patients were examined.

Further author information: (Send correspondence to U.B., T.C., or H.C.)
U.B.: E-mail: ulas.bagci@northwestern.edu, Telephone: +1 240-383-8587
T.C.: E-mail: cukur@ee.bilkent.edu.tr, Telephone: +90 312 290 1164
H.C.: E-mail: haydari@gmail.com, Telephone: +1 202 476 5024

Medical Imaging 2022: Computer-Aided Diagnosis, edited by Karen Drukker, Khan M. Iftekharuddin,
Hongbing Lu, Maciej A. Mazurowski, Chisako Muramatsu, Ravi K. Samala, Proc. of SPIE
Vol. 12033, 1203306 · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2611490

A CNN model was built to classify disease given short-time magnitude spectrogram of audio recordings. More recently, Brown et.al.[8] have proposed to use breathing and cough sounds from COVID-19 patients for disease screening. Spectrogram features were combined with pre-trained visual embeddings from a VGG-type network, and logistic regression or SVM classifiers were built. While promising results were reported when simultaneously leveraging breathing and cough sounds, the study was limited to traditional machine learning models.

Here, we propose a novel deep learning technique to automatically detect COVID-19 patients based on brief audio recordings of their cough and breathing sounds.[8] The proposed technique first extracts spectrogram features of respiratory recordings, and then classifies disease state via a vision transformer architecture. Demonstrations are provided on a crowdsourced database of respiratory sounds from patients and healthy controls.[8] The proposed transformer model is compared against alternative methods based on state-of-the-art convolutional and transformer architectures, as well as traditional machine-learning classifiers. Our results indicate that the proposed model achieves on par or superior performance to competing methods. In particular, the proposed technique can distinguish COVID-19 patients from healthy users with cough symptoms with over 94% AUC.

## 2. METHODS

In this study, we propose to detect COVID-19 using audio recordings of respiratory sounds, specifically cough and breathing sounds. To do this, the proposed technique processes spectrogram features of respiratory sounds with a vision transformer model. The remainder of this section explains details regarding the dataset curation, pre-processing and modeling procedures.

### 2.1 Dataset

We performed demonstrations on a public dataset containing audio recordings of respiratory sounds on cellular phones or personal computers.[8] This dataset was crowdsourced with volunteer submissions across multiple countries, and further curated to focus on only cough and breathing sounds while discarding silent or noisy recordings. Three subject groups were compiled as described in:[8] 'COVID' group with positive test results within 14 days of the recording (141 samples), 'non-COVID' group with clean medical record (298 samples), and 'non-COVID with cough' group with clean medical history albeit with a symptom (32 samples). Each sample corresponds to an audio recording from a unique subject. The digital recordings were imported at a sampling rate of 22050 Hz using the librosa library in Python.[9] The silent periods at the beginning and end of the recording were trimmed. Only recordings that were longer than 2 s were subjected to further analyses.

### 2.2 Classification Tasks

Two separate tasks were considered in distinguishing COVID-19 patients from healthy controls. In Task 1, the aim was to separate patients in the COVID group from healthy subjects without any symptoms. Within each subject, both cough and breathing sounds were recorded. Task 1 was implemented using the two modalities simultaneously, and each modality individually to assess their relative contributions. In Task 2, the aim was to separate patients in the COVID group from healthy controls with reported cough symptoms. Since cough is a common symptom across groups in this case, Task 2 was implemented using only the cough modality.

### 2.3 Proposed Model

The proposed model predicts COVID-19 presence given audio recordings of cough and/or breathing sounds (Figure 1). Many prior studies suggest that spectro-temporal features in audio recordings of respiratory sounds can carry discriminative information regarding disease symptoms.[10–12] Following this convention, we computed spectrogram representations of audio recordings via a 2048-point FFT (capturing 2048 data points in each frame), and with 128 overlapping points between consecutive frames. The resulting spectrograms were log transformed to induce compressive nonlinearity, and then converted to 224x224 grayscale images. The spectrogram images were given as inputs to a deep classifier, based on a recent hierarchical vision transformer architecture (Swin Transformer) reported to yield state-of-the-art performance in general computer vision tasks.[13] Given the moderate dataset size, a compact variant of Swin Transformer, Swin-Small, was implemented with 4 blocks of (2,2,18,2) layers, 96 channels, a window size of 7, 4 attention heads with 32-long queries each. The Swin-S model was pre-trained for object classification on ImageNet-1k database with 1000 classes. The model's classification head was later modified to project onto 2 output units for the COVID detection tasks implemented here. The pre-trained Swin-S model was then fine-tuned on respiratory recordings for each task, where it was optimized to predict disease presence given spectrogram images.

## 2.4 Baselines

Several state-of-the-art competing models were implemented to perform the same classification tasks. These included traditional machine learning methods, convolutional neural networks along with recent transformer architectures.

**SVM** In this traditional machine learning baseline, a large array of hand-crafted (e.g., MFCC, spectrogram) and data-driven (VGG-based) features were taken as inputs as described in.[8] An SVM classifier was then built for each task to detect COVID patients.

**ResNet** This deep-learning (DL) baseline followed the same overall strategy as the proposed technique, where spectrogram images of cough and/or breathing sounds were taken as inputs. However, the classifier was based on the ResNet34 architecture,[14] pretrained on ImageNet-1k. The last fully-connected (FC) layer within ResNet34 was replaced with a fully-connected layer with two output units. The model was fine-tuned respiratory recordings.

**ViT** In this DL baseline, the classifier was based on the Vision Transformer architecture,[15] pretrained on ImageNet-21k and later fine-tuned on ImageNet-1k. The ViT model employed 16x16 patches and assumed an image resolution of 224x224. The classification head was modified from 1000 to 2 classes. The model was fine-tuned on respiratory recordings.

**CvT** In this DL baseline, the classifier was based on the Convolutional Vision Transformer architecture,[16] pretrained on ImageNet-1k. The CvT-13 model assumed an image resolution of 224x224. The last linear layer in the model was modified from 1000 to 2 outputs. The model was fine-tuned on respiratory recordings.

**Ensemble 1** Ensemble models were also considered by fusing the predictions of convolutional and transformer architectures. A first baseline was created by combining the pre-trained ResNet and ViT models. The final classification layers of both models were removed, the features in the preceding layers were concatenated, and two output units were inserted. The ensemble model was fine-tuned on respiratory recordings.

**Ensemble 2** A second ensemble baseline was created by combining the pre-trained ResNet and Swin-S models. The final classification layers of both models were removed, the features in the preceding layers were concatenated, and two output units were inserted. The ensemble model was fine-tuned on respiratory recordings.

## 2.5 Modeling Procedures

Given the relatively modest size of the datasets, data augmentation procedures were performed on the spectrogram images.[8] Accordingly, each image was randomly rotated by an angle up to 20 degrees, randomly flipped across the horizontal axis. Images were finally normalized to a mean of 0.5 and a standard deviation of 0.5. A binary cross-entropy loss function was used to build classification models. A 10-fold cross-validation procedure was employed where data were randomly split into an 80% training set and 20% test set. The training set was further split to reserve a separate validation set to select model hyperparameters. The selected hyperparameters for each model included number of epochs, learning rate, and regularization parameter for $L_2$ norm of model weights ($\lambda_{L_2}$). Performance metrics were taken as the area under the receiver operating characteristic curve (AUC), precision, recall, and F1. Metrics were averaged across test sets, and mean and standard deviation across subjects were reported.

The selection of hyperparameters for each model are summarized here. **ResNet:** 100 epochs, $10^{-3}$ learning rate, $\lambda_{L_2}$=$10^{-4}$. **ViT:** 100 epochs, $10^{-3}$ learning rate, $\lambda_{L_2}$=$10^{-4}$. **CvT:** 100 epochs, $3x10^{-4}$ learning rate, $\lambda_{L_2}$=$10^{-4}$. **Ensemble 1:** 100 epochs, $10^{-3}$ learning rate, $\lambda_{L_2}$=$10^{-4}$. **Ensemble 2:** 100 epochs, $3x10^{-3}$ learning rate, $\lambda_{L_2}$=$10^{-5}$. **Proposed:** 100 epochs, $10^{-5}$ learning rate, $\lambda_{L_2}$=$10^{-8}$. Batch size was taken as 8 samples for all models, except ensemble models where it was reduced to 6. Gradient clipping was used with an upper threshold of 0.1 for the gradient norm. ResNet, ViT and Ensemble 1 models were trained via the Adam optimizer, whereas Proposed, CvT and Ensemble 2 models were trained via the AdamW optimizer.

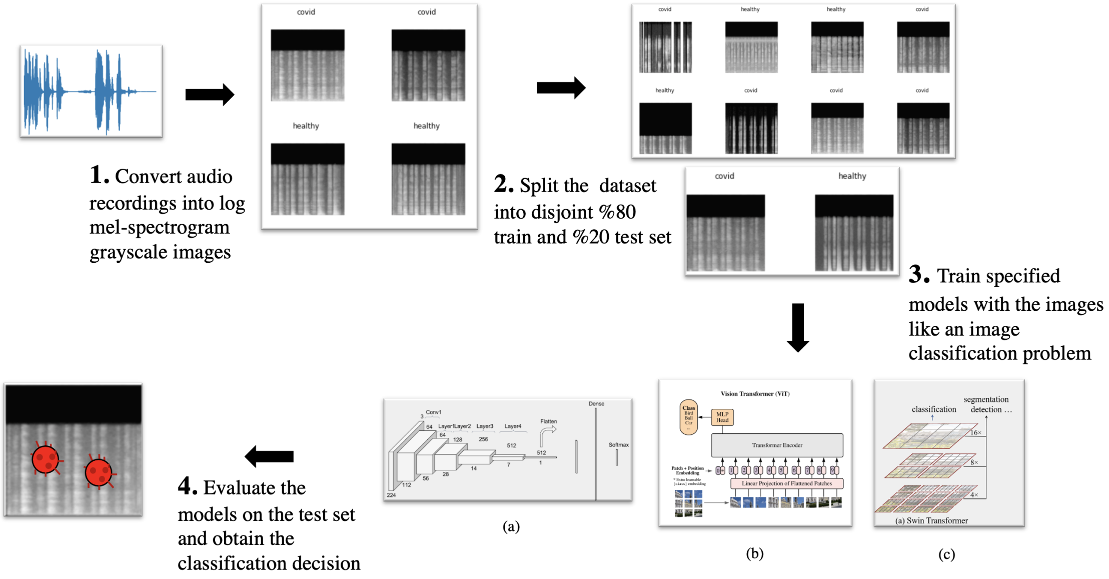

covid
covid
healthy
healthy
1. Convert audio recordings into log mel-spectrogram grayscale images
2. Split the dataset into disjoint %80 train and %20 test set
covid
healthy
3. Train specified models with the images like an image classification problem
4. Evaluate the models on the test set and obtain the classification decision
Vision Transformer (ViT)
Transformer Encoder
MLP Head
Linear Projection of Flattened Patches
classification
segmentation detection ...
(a) Swin Transformer
(a)
(b)
(c)

Figure 1: Flowchart of the proposed methodology for COVID detection. **1)** Auditory recordings of respiratory sounds (e.g., cough, breathing) are transformed into spectrogram representations. **2)** Data are split into non-overlapping training and test sets. **3)** Deep classification models based on (a) ResNet,[14] (b) ViT[16] and (c) Swin-S[13] (proposed) are trained to detect disease based on input spectrogram images of respiratory sounds. **4)** Model performance is evaluated on the test sets.

Table 1: Performance in Task 1 based on both cough and breathing sounds, distinguishing subjects in the COVID group from those in the non-COVID group without symptoms.

| Models | Precision | Recall | F1 |
|---|---|---|---|
| SVM | 0.686 ± 0.089 | 0.748 ± 0.204 | 0.703 ± 0.130 |
| ResNet | 0.835 ± 0.074 | 0.831 ± 0.080 | 0.831 ± 0.063 |
| ViT | 0.890 ± 0.091 | 0.785 ± 0.128 | 0.821 ± 0.058 |
| CvT | 0.817 ± 0.116 | 0.724 ± 0.134 | 0.755 ± 0.089 |
| Ensemble 1 | 0.780 ± 0.102 | 0.830 ± 0.072 | 0.797 ± 0.053 |
| Ensemble 2 | 0.729 ± 0.180 | 0.753 ± 0.221 | 0.688 ± 0.064 |
| **Proposed** | 0.888 ± 0.076 | 0.936 ± 0.041 | 0.908 ± 0.032 |

## 3. RESULTS

We comparatively demonstrated the proposed transformer model (see Fig. 1) against the baselines described in Sec. 2.4. The two classification tasks detailed in Sec. 2.2 were considered. Task 1 is aimed at distinguishing COVID patients from healthy subjects without any respiratory symptoms. Classification performance of competing models in this task are reported in Tab. 1 with cough and breathing modalities as input, in Tab. 2 for only breathing modality as input, in Tab. 3 for only cough modality as input. Please note that Brown et.al.[8] proposed to build separate SVMs for breathing and cough modalities and only report the best performing one, so the SVM baseline follows their convention for consistency. The proposed method outperforms the other competing methods ($p < 0.05$) except ResNet in Task 2 where both ResNet and the proposed model are saturated.

Task 2 is aimed at distinguishing COVID patients from healthy subjects with cough symptom. Classification performance of competing models in this task are reported in Tab. 4. Overall, all deep-learning models achieve

Table 2: Performance in Task 1 based on only breathing sounds, distinguishing subjects in the COVID group from those in the non-COVID group without symptoms.

| Models | Precision | Recall | F1 |
|---|---|---|---|
| SVM | 0.659 ± 0.089 | 0.765 ± 0.118 | 0.680 ± 0.081 |
| ResNet | 0.856 ± 0.070 | 0.874 ± 0.071 | 0.863 ± 0.057 |
| ViT | 0.823 ± 0.099 | 0.705 ± 0.191 | 0.747 ± 0.131 |
| CvT | 0.768 ± 0.184 | 0.685 ± 0.189 | 0.690 ± 0.105 |
| Ensemble 1 | 0.804 ± 0.058 | 0.796 ± 0.069 | 0.797 ± 0.045 |
| Ensemble 2 | 0.738 ± 0.140 | 0.718 ± 0.210 | 0.692 ± 0.075 |
| **Proposed** | 0.937 ± 0.059 | 0.953 ± 0.042 | 0.944 ± 0.040 |

Table 3: Performance in Task 1 based on only cough sounds, distinguishing subjects in the COVID group from those in the non-COVID group without symptoms.

| Models | Precision | Recall | F1 |
|---|---|---|---|
| SVM | 0.679 ± 0.082 | 0.771 ± 0.153 | 0.696 ± 0.078 |
| ResNet | 0.807 ± 0.117 | 0.895 ± 0.081 | 0.844 ± 0.081 |
| ViT | 0.894 ± 0.111 | 0.701 ± 0.128 | 0.771 ± 0.054 |
| CvT | 0.806 ± 0.130 | 0.771 ± 0.079 | 0.778 ± 0.066 |
| Ensemble 1 | 0.843 ± 0.121 | 0.745 ± 0.156 | 0.771 ± 0.079 |
| Ensemble 2 | 0.782 ± 0.175 | 0.676 ± 0.189 | 0.680 ± 0.090 |
| **Proposed** | 0.924 ± 0.082 | 0.936 ± 0.066 | 0.926 ± 0.045 |

relatively high performance in this task compared to Task 1, with almost saturated AUC levels. Here, the proposed model performs competitively with the ResNet and CvT models, while outperforming the SVM model by a large margin.

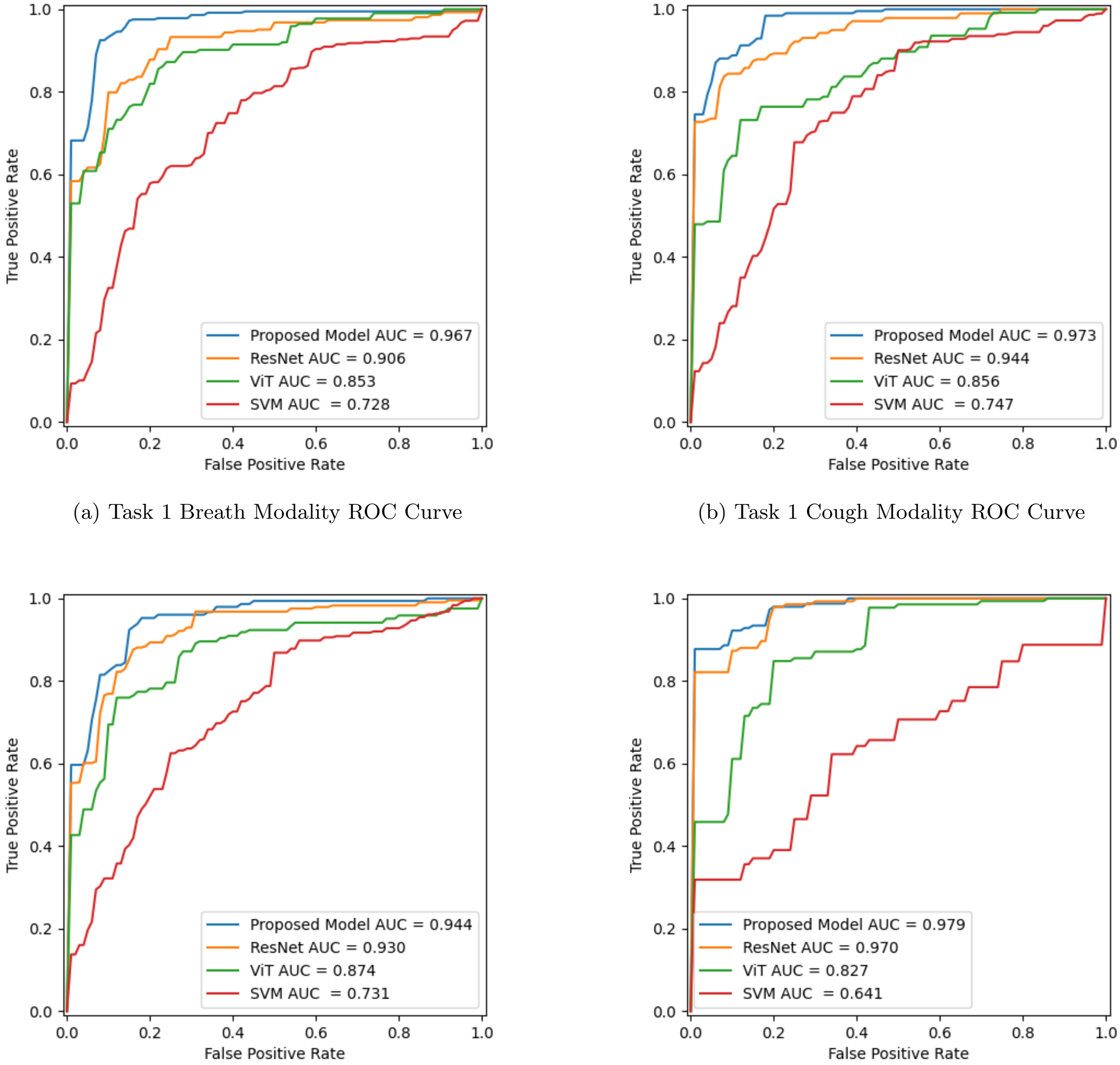
The ROC curves of the proposed model, ResNet, ViT and SVM are shown in Figure 2 for all four tasks. In Figure 2d, the ROC curves are steep as they reach their highest AUC value rapidly due to the size of the dataset for Task 2. The proposed model beats ResNet and ViT in all modalities of Task 1 while our proposed model, ResNet and ViT outperform SVM by a wide margin.

## 4. DISCUSSION

Prior studies in the literature have proposed the traditional classifiers or CNNs to classify respiratory diseases given spectro-temporal features of respiratory sounds.[4–7] Yet, traditional classifiers often have suboptimal sensitivity, and CNNs have suboptimal capture of long-range dependencies in their inputs. In this study, we proposed a novel method based on hierarchical vision transformers to detect COVID-19 from cough and breathing sounds. Our results indicate that the proposed method yields on par or superior performance to CNNs and traditional classifiers. While we have primarily focused on spectrogram representations of auditory recordings in this study, it is possible to perform classification directly on audio signals via recurrent neural network architectures. Further work is warranted to evaluate the relative benefits of using spectrogram features versus raw auditory signals in disease detection.

## ACKNOWLEDGMENTS

This study is partially supported by the NIH grants R01-CA246704 and R01-CA240639, and by TUBA GEBIP 2015 and BAGEP 2017.



(a) Task 1 Breath Modality ROC Curve

(b) Task 1 Cough Modality ROC Curve

(c) Task 1 Breath and Cough Modality ROC Curve

(d) Task 2 Cough Modality ROC Curve

Figure 2: ROC Curves of the proposed model, ResNet, ViT and SVM classifiers.

Table 4: Performance in Task 2, distinguishing subjects in the COVID group from those in the non-COVID group with cough symptoms.

| Models | Precision | Recall | F1 |
|---|---|---|---|
| SVM | $0.579 \pm 0.157$ | $0.683 \pm 0.253$ | $0.686 \pm 0.168$ |
| ResNet | $0.900 \pm 0.071$ | $0.960 \pm 0.046$ | $0.927 \pm 0.047$ |
| ViT | $0.860 \pm 0.065$ | $0.863 \pm 0.075$ | $0.859 \pm 0.051$ |
| CvT | $0.881 \pm 0.071$ | $0.921 \pm 0.079$ | $0.897 \pm 0.047$ |
| Ensemble 1 | $0.785 \pm 0.152$ | $0.964 \pm 0.086$ | $0.851 \pm 0.088$ |
| Ensemble 2 | $0.842 \pm 0.126$ | $0.814 \pm 0.137$ | $0.809 \pm 0.064$ |
| **Proposed** | $0.896 \pm 0.067$ | $0.960 \pm 0.053$ | $0.924 \pm 0.045$ |

## REFERENCES

[1] Falk, T. H., Chan, W.-Y., Sejdic, E., and Chau, T., "Spectro-temporal analysis of auscultatory sounds," *New Developments in Biomedical Engineering* , 93–104 (2010).

[2] Sun, L., Joshi, M., Khan, S. N., Ashrafian, H., and Darzi, A., "Clinical impact of multi-parameter continuous non-invasive monitoring in hospital wards: a systematic review and meta-analysis," *Journal of the Royal Society of Medicine* **113**(6), 217–224 (2020).

[3] Aliverti, A., "Wearable technology: role in respiratory health and disease," *Breathe* **13**(2), e27–e36 (2017).

[4] Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., John, C. N., Hussain, M. I., and Nabeel, M., "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked* **20**, 100378 (2020).

[5] Bales, C., Nabeel, M., John, C. N., Masood, U., Qureshi, H. N., Farooq, H., Posokhova, I., and Imran, A., "Can machine learning be used to recognize and diagnose coughs?," in [*2020 International Conference on e-Health and Bioengineering (EHB)*], 1–4, IEEE (2020).

[6] Pramono, R. X. A., Imtiaz, S. A., and Rodriguez-Villegas, E., "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one* **11**(9), e0162128 (2016).

[7] Bagad, P., Dalmia, A., Doshi, J., Nagrani, A., Bhamare, P., Mahale, A., Rane, S., Agarwal, N., and Panicker, R., "Cough against covid: Evidence of covid-19 signature in cough sounds," *arXiv preprint arXiv:2009.08790* (2020).

[8] Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., and Mascolo, C., "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in [*Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*], 3474–3484 (2020).

[9] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O., "librosa: Audio and music signal analysis in python," in [*Proceedings of the 14th python in science conference*], **8** (2015).

[10] Pramono, R. X. A., Imtiaz, S. A., and Rodriguez-Villegas, E., "Evaluation of features for classification of wheezes and normal respiratory sounds," *PloS one* **14**(3), e0213659 (2019).

[11] Bahoura, M., "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes," *Computers in biology and medicine* **39**(9), 824–843 (2009).

[12] Song, I., "Diagnosis of pneumonia from sounds collected using low cost cell phones," in [*2015 International Joint Conference on Neural Networks (IJCNN)*], 1–8 (2015).

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030* (2021).

[14] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR* (2021).

[16] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L., "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808* (2021).