COVID-19 Detection From Respiratory Sounds With Hierarchical Spectrogram Transformers

Idil Aytekin ⬩, Onat Dalmaz, Kaan Gonc ⬩, Haydar Ankishan, Emine Ulku Saritas ⬩, Ulas Bagci ⬩, Haydar Celik ⬩, and Tolga Çukur ⬩, *Senior Member, IEEE*

Abstract—Monitoring of prevalent airborne diseases such as COVID-19 characteristically involves respiratory assessments. While auscultation is a mainstream method for preliminary screening of disease symptoms, its utility is hampered by the need for dedicated hospital visits. Remote monitoring based on recordings of respiratory sounds on portable devices is a promising alternative, which can assist in early assessment of COVID-19 that primarily affects the lower respiratory tract. In this study, we introduce a novel deep learning approach to distinguish patients with COVID-19 from healthy controls given audio recordings of cough or breathing sounds. The proposed approach leverages a novel hierarchical spectrogram transformer (HST) on spectrogram representations of respiratory sounds. HST embodies self-attention mechanisms over local windows in spectrograms, and window size is progressively grown over model stages to capture local to global context. HST is compared against state-of-the-art conventional and deep-learning baselines. Demonstrations on crowd-sourced multi-national datasets indicate that HST outperforms competing methods, achieving over 90% area under the receiver operating characteristic curve (AUC) in detecting COVID-19 cases.

Index Terms—COVID-19, respiratory sound classification, auditory, spectrogram, transformer, auscultation.

I. INTRODUCTION

AUSCULTATION is a primary step in preliminary assessment of subjects for symptoms of respiratory disorders [1]. Assessment via stethoscope is non-invasive and inexpensive, but it must be performed by a healthcare professional during a hospital visit. As such, early-stage or continuous monitoring of symptoms might not be feasible across broad populations [2], as experienced during the COVID-19 pandemic that has significantly disrupted access to healthcare facilities across the globe [3]. A promising solution is remote monitoring of respiratory sounds based on audio recordings captured via portable equipment such as mobile or wearable devices [4]. While COVID-19 is clinically diagnosed with reverse-transcription polymer chain reaction (RT-PCR) tests [5] and/or radiological imaging [6], [7], [8], economic and time costs of these lab-administered procedures restrict patient access and elicit backlogs during periods of high transmission [9], [10], [11]. Remote screening of respiratory sounds can assist in preliminary assessment and risk stratification for potential COVID-19 cases under low-resource settings. By more-informed resource allocation, it can facilitate early referrals and timely interventions to deteriorating patients to help contain the spread of disease [12].

Auditory screening of respiratory disorders relies on the prevalence of disease-specific features in respiratory sounds [13]. Diverse pathology can be encountered in respiratory disorders ranging from inflammation and obstruction to consolidation and pleural effusion. While some commonalities exist among diseases, precise characteristics of pathology including location and severity typically show disease-specific patterns [14], [15]. Imaging studies report that common lung pathology in non-COVID-related pneumonia has central-peripheral distribution, air bronchograms, and pleural enlargement/effusion; whereas COVID-related pneumonia frequently elicits lower-peripheral distribution, enhanced ground glass opacity and vascular enlargement [16], [17]. These pathomorphological changes have been associated with increased prevalence of adventitious respiratory sounds in COVID-19 such as coarse breathing, wheezes, and crackles [18], which may carry distinctive cues compared to respiratory sounds in other diseases such as asthma, chronic obstructive pulmonary disease (COPD), bronchitis and pertussis [15], [19].

Literature suggests that normal respiratory sounds resemble filtered noise with a typical frequency range of 100–1500 Hz during inspiration/expiration segments of the respiration cycle, and are nearly inaudible during inter-segment intervals [20]. Meanwhile, adventitious sounds common in COVID-19 can show a degree of divergence from normal sounds in terms of their spectro-temporal characteristics. For instance, wheezes
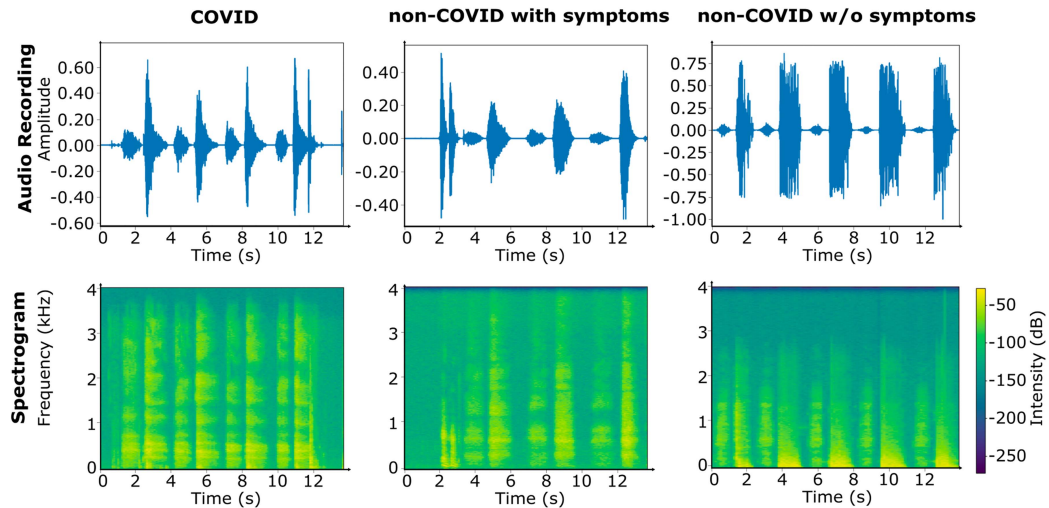
Fig. 1.   Audio recordings of breathing sounds depicted in time-domain (top row) and corresponding spectrogram representations (bottom row). Representative samples are displayed for a COVID case, a non-COVID case with symptoms, and a non-COVID case without symptoms.

resemble repeated sinusoids that last longer than ∼80 ms and that span across 100–5000 Hz [21], [22], and crackles resemble dampened sinusoids shorter than ∼15 ms and that span across 150–2000 Hz [18], [23]. Representative samples of breathing sounds for COVID and non-COVID cases are displayed in Fig. 1. The non-COVID case without symptoms follows a cyclic pattern with moderate frequency range and relatively silent inter-segment intervals. The non-COVID case with symptoms shows a degree of irregularity in time and frequency distribution, and modest elevation of intensity at moderate frequencies. In comparison, the COVID case shows a uniformly spread intensity distribution across both time and frequency dimensions, relatively stronger intensity at higher frequencies. Corroborating findings in literature, such apparent and other fine-grained differences in time-frequency characteristics bring forth the possibility of automated screening to identify potential COVID-19 cases.

Given recordings of respiratory sounds, remote screening leverages an algorithm to automatically infer respiratory conditions. Previous studies have successfully applied machine learning (ML) algorithms to detect a broad spectrum of conditions including bronchitis, bronchiolitis, emphysema, pertussis, pneumonia, rale and rhoscus [24], [25], [26], [27], [28]. Commonly, summary descriptors of audio data were extracted such as Mel-frequency cepstral coefficients (MFCC), tonal and chroma features [29]. Classifiers were then built via traditional methods such as support vector machines (SVM) and logistic regression, or via network models such as convolutional neural networks (CNN) [25] and recurrent neural networks (RNN) [26], [30]. Recent studies have followed similar approaches for COVID-19 detection from cough, breathing or speech sounds [31], [32], [33], [34]. Various sound descriptors were extracted including MFCC, tonal, chroma, spectral contrast, glottal flow and spectrogram features [32], [34], [35], [36]. Either traditional classifiers such as SVM, decision trees and random forests [31], [37], [38], or network models including CNNs or RNNs [26], [30], [39], [40] were then used for detection.

Despite the potential of learning-based approaches for preliminary COVID-19 screening, there remain avenues for further technical improvement. Many prior studies have employed shallow models –based on traditional classifiers or neural networks–

in conjunction with hand-crafted audio features. While this approach mitigates model complexity, latent representations in shallow models might have limited power to capture the rich information in respiratory sounds [38]. Few studies have considered additional input features extracted from pre-trained networks, which can be suboptimal compared against task-specific features derived via end-to-end learning [31]. Other studies have reported elevated detection performance with deep CNN or RNN models that capture a hierarchy of latent representations. Yet, CNNs perform local filtering with compact kernels so they show limited sensitivity to long-range contextual features [41]. Although RNNs can improve capture of temporal context, serial processing of input sequences can introduce computational burden and compromise feature learning over long time scales [41].

In this study, we introduce a novel deep learning method to automatically screen COVID-19 symptoms with short recordings of respiratory sounds. The proposed method first converts respiratory sounds onto a time-frequency (i.e., spectrogram) representation, and then classifies disease from spectrogram features using a novel hierarchical spectrogram transformer (HST). Unlike prior models that use a compact set of knowledge-based features, HST leverages a comprehensive characterization of respiratory sounds through high-resolution log-spectrogram features. Compared to CNNs, HST leverages self-attention mechanism for improved sensitivity to long-range context in spectrograms. Compared to RNNs and vanilla transformers, HST leverages a patch-based approach where the time-frequency extent of attended regions is progressively increased for computational efficiency.

Main Contributions:

- We introduce a novel hierarchical spectrogram transformer for screening COVID-19 symptoms from audio recordings of respiratory sounds.
- The proposed transformer model progressively captures local to global context in spectrogram representations for enhanced efficiency.
- We demonstrate improved performance in COVID-19 detection against state-of-the-art baselines including

traditional, CNN, RNN, transformer, and ensemble methods.

II. RELATED WORK

A. Shallow Classifiers

A first group of studies have proposed to detect COVID-19 via shallow models based on either traditional ML methods (e.g., SVM, random forests, decision trees), or neural networks of limited depth. In [37], MFCC features of cough sounds were analyzed with SVM to separate COVID-19 from other respiratory infections. In [32], a diverse set of audio features including the power spectrum, Mel spectrum, chroma, tonal, and MFCC features were extracted from cough sounds, and analyzed with SVM and logistic regression models. In [42], cough, breath, and voice recordings were analyzed with an ensemble of shallow CNN, gradient boosted trees and logistic regression models given Mel-spectrogram and cochleagram features along with network-based features from a pre-trained convolutional architecture for audio data (VGGish) [43]. In [33], hand-crafted features of cough sounds including MFCC, log energy and entropy were analyzed via a shallow five-layer model with gated linear units. Cough features were augmented with symptomatic and demographic features for improved performance. In [34], MFCC, tonal, chroma, spectral contrast, spectrogram features of cough sounds were input to an ensemble of decision-tree, logistic regression, random forest, boosting and multi-layer perceptron (MLP) models. In [38] and [44], COVID-19 screening was proposed based on spectrogram and wavelet features of respiratory sounds concatenated with network-based features obtained from pre-trained VGGish or CNN models. The compiled features were then processed with an SVM or logistic regression model. In [36], a new bio-inspired cepstral feature was proposed for COVID-19 detection, and demonstrated on speech, breathing, or cough sounds with SVM. In [31] acoustic and scattering features were augmented with pre-trained VGGish features of cough, breath, speech sounds. Decision trees, random forests or MLP models were then used. These previous studies commonly leveraged shallow models of low complexity to process either hand-crafted features or task-agnostic features from pre-trained networks. While this approach can improve learning behavior on limited datasets, the resultant classifiers are deprived from a diverse hierarchy of task-specific latent features that could be captured via end-to-end deep learning.

B. Deep Classifiers

A second group of studies have instead considered deep models typically based on CNN or RNN architectures for COVID-19 detection. In [37], a CNN was used to predict COVID-19 given MFCC features of cough sounds. In [39], short-time magnitude spectrogram of cough sounds were classified via a CNN. In [32], power spectrum, Mel spectrum, chroma, tonal, and MFCC features of cough sounds were classified using an ensemble of pre-trained CNN and traditional ML models. In [40], an ensemble of CNNs was used given as input MFCC features of cough sounds. Three different pre-trained CNNs were used to extract features related to lung and respiratory tract, vocal cord, and sentiment information. In [45] and [46], spectrogram features of cough and breathing sounds were analyzed with an ensemble of CNNs. In [47], spectrogram features of cough and breathing sounds were analyzed via a CNN. In [35], glottal flow features of speech sounds were extracted and classified via a CNN. In [48],

augmented mel-spectrogram features of breathing sounds were analyzed via a CNN. Although CNN models are powerful in capturing local features in time-frequency representations of audio data, the inductive bias introduced by filtering with local kernels limits sensitivity for long-range contextual features.

Several recent studies have instead proposed architectures designed for sequence modeling to improve capture of temporal context. In [49], MFCC, chroma and spectral features of cough sounds were analyzed with a long-short-time-memory (LSTM) model. In [50], spectral centroid, spectral roll-off, zero-crossing rate, MFCC, first and second derivatives of MFCC features of cough, breathing, and voice sounds were analyzed with an LSTM. In [51], both hand-crafted and learning-based features from spectrograms of cough sounds were analyzed with an ensemble of SVMs and LSTMs. In [52], an LSTM was proposed to detect COVID-19 from MFCC, power spectrum, and filter bank features of cough, breath and sneeze sounds. In [53], cepstral coefficients and filter bank features of breathing, cough, and speech sounds were analyzed with a bidirectional LSTM. In [54] and [55], hybrid CNN-LSTM architectures were proposed for COVID-19 detection, which received as input MFCC or spectrogram features of cough sounds. In [56], network-based features of speech and cough sounds were inspected via a gated recurrent unit (GRU) model that also received clinical features for improved accuracy. While RNNs have demonstrated success in capturing signal correlations over long time scales, they can suffer from suboptimal learning due to vanishing gradients.

Here, we introduce a novel hierarchical transformer model for COVID-19 detection from spectrogram features of respiratory sounds. Few recent studies have independently considered transformer models for analyzing respiratory sounds in COVID-19 patients. In [56], sound embeddings extracted from a pre-trained transformer are analyzed with an RNN for classification. Note that [56] use a vanilla transformer of quadratic complexity, and perform task-agnostic pre-training to extract sound embeddings that are frozen during the classification stage. Instead, HST employs a hierarchical patch-based attention to progressively capture local to global context while alleviating computational burden, and it is trained end-to-end for disease classification resulting in task-specific representations. In [57], spectrogram, MFCC and other clinical features are analyzed simultaneously with a nested multi-modal transformer. The nested transformer splits audio spectrograms into non-overlapping patches of growing size across a hierarchy. Yet, the nested transformer independently processes single patches in each stage, only permitting interactions among neighboring patches via pooling operations during block aggregation. In contrast, HST uses cyclic shifts to split spectrograms into partly overlapping patches. It further leverages self-attention operations to enable interactions between a broader set of patches at each stage, which can improve sensitivity to global context and model performance.

III. METHODS

A. Experimental Datasets

We performed demonstrations on two public datasets containing audio recordings of respiratory sounds: Cambridge (breathing and cough; https://www.covid-19-sounds.org/en/blog/data_sharing.html) [38], and COUGHVID (cough; https://zenodo.org/record/4048312) [58]. Vocal respiratory sounds were recorded via microphones on cellular phones or personal computers. Participants were instructed to record in a silent

environment, and breathe deeply through their mouth or cough directly to the microphones held within arm's length.

1) Cambridge Dataset: The dataset included three subject groups [38]: "COVID" group with positive test results within 14 days prior to the recording (141 samples), "non-COVID without symptom" group with clean medical records (298 samples), and "non-COVID with symptom" group with clean medical records albeit with a cough symptom (32 samples). The "non-COVID" participants were recruited from countries where the virus was not widespread during the time of data collection, had not tested positive for COVID-19.

Each participant provided a written report of symptoms, and uploaded a recording containing five breathing or three cough samples through either an Android application or a web-based platform. Each sample corresponded to a single audio recording. Recordings were imported at a sampling rate of 22050 Hz using the Librosa library [59]. Silent or noisy recordings were discarded. Silent periods at the beginning and end of the recording were trimmed. Only recordings that were longer than 2 s were analyzed. To alleviate data imbalance, augmentation was performed on the training set via amplification by a random scalar in [1.15 2], random change of pitch speed by [0.8 0.99], and addition of white noise without distorting the audibility of the original recordings significantly, following procedures in [38].

Two separate tasks were considered: discriminating "COVID" versus "non-COVID without symptom" groups (Task 1), and "COVID" versus "non-COVID with symptom" groups (Task 2). Both tasks were implemented given either cough or breath modalities as input [60]. In Task 1, we analyzed (137, 141) samples in COVID and non-COVID groups respectively for cough modality, (141, 144) samples in COVID and non-COVID groups for breath modality. In Task 2, we analyzed (54, 88) samples in COVID versus non-COVID groups for cough modality, (56, 89) samples in COVID versus non-COVID groups for breath modality.

2) COUGHVID Dataset: The dataset included participants of different ages, genders, geographic locations, and COVID-19 statuses [58]. Four experienced physicians labeled the recordings to diagnose any pulmonary abnormalities. Here, two groups of subjects were analyzed: "COVID" group with disease labels and cough symptoms (608 samples), and the "non-COVID with symptom" group with clean medical records and cough symptoms (1778 samples). Cough sounds were captured through a web-based platform. To improve data quality, poor recordings with signal-to-noise ratio (SNR) lower than 0.8 were discarded. A spectral peak detection algorithm was used to segment each recording into individual cough events. The recordings were imported using the Librosa library [59] at a sampling rate of 22050 Hz. Silent periods present at the start and end of the recordings were trimmed. Only recordings that were longer than 2 s were analyzed. Data augmentation was performed on the training set to alleviate data imbalance. A single task given cough modality as input was considered where "COVID" versus "non-COVID with symptom" groups were discriminated. Accordingly, we analyzed (1644, 1644) samples in COVID versus non-COVID groups.

B. Hierarchical Spectrogram Transformer

The proposed method leverages a novel architecture to detect COVID-19 from respiratory sounds. Audio recordings of cough or breathing sounds are first mapped onto a spectrogram that captures a two-dimensional time-frequency representation (Section III-B1). The spectrogram is then processed with a hierarchical patch-based attention mechanism to capture local to global context across network layers (Section III-B2). Here, we introduce a five-stage model where the first four stages derive latent audio representations with a cascade of local-windowed transformer blocks (Section III-B3). The spectrotemporal resolution of feature maps is progressively lowered across these stages, while the channel dimensionality is expanded. Meanwhile, the last stage maps latent representations onto a classification output via cross-entropy loss (Section III-B4). Code for implementing HST is available at https://github.com/icon-lab/HST.

1) Spectrogram Features: Many prior studies on disease detection from audio recordings suggest that spectrotemporal features can provide comprehensive description of respiratory sounds [61], [62], [63]. Inspired by this success, here we employ Mel-spectrogram representations of audio recordings. Given a recording of duration D, a the short-time Fourier transform (STFT) of audio recordings are calculated over windowed segments. The STFT ($A[k, n]$) of the sampled audio signal ($a[n]$) at time n with length N is taken as:

$$A[k,n] = \sum_{m=-\infty}^{\infty} a[m] \cdot w[m-n] \cdot \exp\left(-\frac{j2\pi km}{N}\right) \quad (1)$$

for $k = 0, 1, 2, \ldots, N-1$. The computed frequency components are $f = \frac{kf_s}{N}$ where the sampling frequency is f_s=22050 Hz, $N = 2048$ and $w[n]$ is the Hanning Window of length N.

$$w[n] = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi n}{N}\right) \quad (2)$$

The STFT is computed via a 2048-point fast Fourier transform (FFT) with 2048 points in each segment and 128 overlapping points between consecutive segments, for a favorable trade-off between spectral and temporal resolution. Lastly, the magnitude square of the STFT coefficients are used to derive the spectrogram representation:

$$\text{Mel-spectrogram}[mel, n] = \log\left(|A[f(mel)N/f_s, n]|^2\right) \quad (3)$$

where $f(mel)$ denotes frequency in Hz corresponding to mel-scale frequency mel. Two-dimensional (2D) spectrograms are log transformed to induce compressive nonlinearity and linearly downsampled onto a 224×224 matrix for efficiency.

2) Network Architecture: Here, we propose a five-stage architecture for COVID-19 detection (Fig. 2). In Stage 1, the 2D spectrograms are partitioned into non-overlapping patches of size $h \times h$, where $h = 4$. Thus, partitioning produces a grid of $P \times P = 56 \times 56$ patches. Each patch is flattened onto a feature dimensionality of $4 \times 4 = 16$, and then projected with a linear embedding layer onto a dimensionality d, where $d = 96$. This provides an input feature map of size $56 \times 56 \times d$ to the HST block in Stage 1. The input map is then processed with the transformer block equipped with local windowed attention mechanisms to achieve linear complexity as described in the following section, instead of global attention mechanisms in vanilla transformer models that suffer from quadratic complexity [41], [64].

HST comprises a hierarchical architecture for efficient capture of contextual features in spectrograms across multiple scales. To do this, the spectrotemporal resolution of features maps is progressively decreased while the embedding dimensionality is
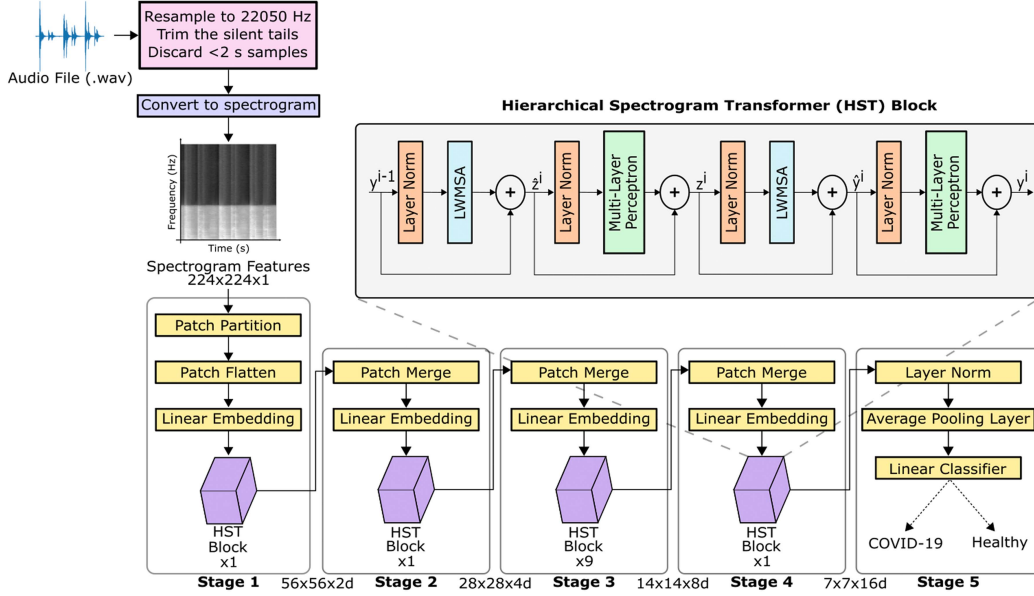
Fig. 2. Overall architecture for the hierarchical spectrogram transformer (HST) model for COVID-19 detection based on audio recordings of bodily sounds. To derive the model input, audio recordings are first mapped onto a spectrogram that captures a two-dimensional time-frequency representation. The spectrogram is then processed with HST that employs a hierarchical patch-based attention mechanism to capture local to global context. HST is a five-stage model where the first four stages derive latent audio representations with a cascade of local-windowed transformer blocks. The spectrotemporal resolution of the feature maps is progressively lowered across these stages, while the embedding dimensionality is increased. The last stage maps latent representations onto a classification output via cross-entropy loss.
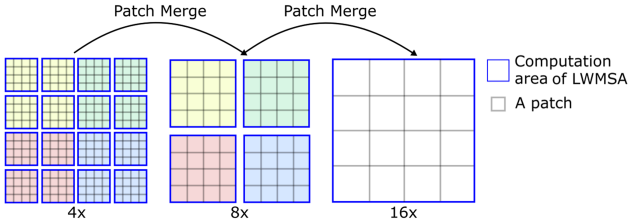
Fig. 3. Illustration of the local windowed multi-head self-attention (LWMSA) mechanism in HST. In a given stage, attention is computed among a local neighborhood of $M \times M$ patches ($M = 4$ in this example). During progression onto the next stage, feature maps are concatenated via patch merge across 2×2 grids of neighboring patches, resulting in 4-fold increase in patch size. LWMSA is again performed across $M \times M$ patches, but a broader receptive field is covered in the spectrogram due to growing patch size.

increased across stages (Fig. 3). In Stages 2–4, a patch merge layer is used across 2×2 grids of neighboring patches to lower the number of patches (i.e., sequence tokens) by a factor of 4. Afterwards, a linear embedding layer scales up the embedding dimensionality by a factor of 2. Therefore, the input feature maps to the stages are given as $(56/2^{S-2}) \times (56/2^{S-2}) \times (2^{S-2}d)$, and the input feature maps to the HST blocks in each stage are given as $(56/2^{S-1}) \times (56/2^{S-1}) \times (2^{S-1}d)$, where S is the stage number (see Fig. 2). A cascade of transformer blocks are then employed in each stage, where local windowed attention is computed over a broader scale due to merged patches.

In its final stage, HST processes latent representations of spectrograms extracted via the prior stages for disease detection. In Stage 5, an input feature map of size $7 \times 7 \times 8d$ is received. Afterwards, this feature map is passed through a cascade of a normalization layer, a one-dimensional adaptive average pooling layer [64], and a linear classification head with two output units.

Small, base and large variants of HST were implemented. Across Stages 1–4, the small variant had $(1,1,3,1)$ blocks, hidden

size of 96–768, MLP size of 384–3072, 3–24 attention heads, the base variant had $(1,1,9,1)$ blocks, hidden size of 96–768, MLP size of 384–3072, 3–24 attention heads, and the large variant had $(1,1,9,1)$ blocks, hidden size of 128–1024, MLP size of 512–4096, 4–32 attention heads.

3) Local Windowed Transformer Blocks: Each stage of HST except for the final stage includes a cascade of transformer blocks to derive attention-based latent representations. The transformer blocks are composed of multi-head self-attention (MSA) modules and MLPs, interleaved with normalization layers and residual connections as in [41]. However, unlike vanilla token-based [65] or patch-based transformers [64] proposed for auditory tasks, the MSA modules in HST leverage local windowed attention as inspired by the the success of restricted attention models in computer vision tasks [66], [67], [68]. Attention is restricted to a local neighborhood of $M \times M$ patches, where $M = 7$ in this work. Within a transformer block, a cyclic shift of $\left(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor\right)$ patches is enforced to improve diversity in window definition prior to the second MSA module. Given the input feature map y^{i-1} to the ith transformer block, latent representations are computed as:

$$\hat{z}^i = \text{LWMSA}\left(\text{LN}\left(y^{i-1}\right)\right) + y^{i-1},$$

$$z^i = \text{MLP}\left(\text{LN}\left(\hat{z}^i\right)\right) + \hat{z}^i,$$

$$\hat{y}^i = \text{LWMSA}\left(\text{LN}\left(z^i\right)\right) + z^i,$$

$$y^i = \text{MLP}\left(\text{LN}\left(\hat{y}^i\right)\right) + \hat{y}^i, \tag{4}$$

where LWMSA is the local windowed MSA module, LN denotes layer normalization, y^l is the output of the transformer block. A two-layer MLP is used with Gaussian error linear unit (GELU) activation functions.

Provided an input sequence of n_t tokens $X \in \mathbb{R}^{n_t \times d}$, the self-attention matrix in an LWMSA module is

calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \quad (5)$$

where $B \in \mathbb{R}^{n_t \times n_t}$ is a relative position bias matrix [41], and $Q, K, V \in \mathbb{R}^{n_T \times d}$ are the query, key, and value matrices obtained as learnable linear projections of X:

$$Q = XL^Q,$$
$$K = XL^K,$$
$$V = XL^V, \quad (6)$$

where $L^Q, L^K, L^V \in \mathbb{R}^{n_t \times d}$ are the corresponding projection matrices. Note that, in (5), calculation of softmax attention involves formation of an $\mathbb{R}^{n_t \times n_t}$ inter-token interaction matrix. A single sequence with $n_t = P^2$ patches are processed in vanilla MSA modules [41], [69], resulting in a quadratic complexity of $O(P^2 \times P^2)$. In contrast, LWMSA splits the overall sequence into $\left(\frac{P}{M}\right)^2$ sub-sequences of length $n_t = M^2$ each. As such, LWMSA enables a linearly scaled complexity of $O\left(\left(\frac{P}{M}\right)^2 \times M^2 \times M^2\right) = O(P^2 \times M^2)$ with respect to sequence length.

4) Loss Function: A binary cross-entropy loss function is employed to train the HST models for COVID-19 detection. The classification head at the output layer produces a predicted probability for each output class in the range [0, 1]. Cross-entropy loss is then expressed as:

$$\text{Loss} = -\frac{1}{J}\sum_{j=1}^{J}(x_j\log(\hat{x}_j) + (1 - x_j)\log(1 - \hat{x}_j)) \quad (7)$$

where x_j is the true label of the jth sample (1 denoting COVID-19, 0 denoting healthy control), $\hat{x}_j$ is the predicted probability of the jth sample, and J is the number of samples.

C. Competing Methods

Several state-of-the-art baselines were adopted including a traditional ML method along with CNN, RNN, transformer, and ensemble models for COVID-19 detection. All models analyzed grayscale spectrograms, yet the spectrogram for a given audio recording was replicated across the three color channels to provide inputs to CNN modules.

SVM: A total of 733 hand-crafted and data-driven features were taken as inputs [38]. Hand-crafted features including duration, onset, tempo, period, RMS energy, spectral centroid, roll-off frequency, zero crossing, MFCC, Δ-MFCC, Δ^2-MFCC were extracted via the Librosa library [59]. Data-driven features were obtained from intermediate layers of a pre-trained *VGGish* model. An SVM classifier with a radial basis function (RBF) kernel was built for each task.

CNN: A CNN model was built that received as input spectrogram features [39]. The ResNet34 architecture was adopted with image resolution 224×224, albeit a fully-connected (FC) layer with two output units was used for COVID-19 detection.

DeepShufNet: A CNN model was built based on the efficient architecture reported in [48]. The models received as input spectrogram features identical to the CNN baseline.

LSTM: An RNN model was built based on the LSTM architecture in [49]. Model inputs included 13 MFCC, 13 Δ-MFCC,

13 Δ^2-MFCC, 1 spectral center, 7 spectral contrast and 12 chroma features. A total of 59 features were compiled.

BiLSTM: A bidirectional RNN model was built based on the BiLSTM architecture reported in [70]. The input features were identical to those for the LSTM baseline.

CNN-LSTM: A hybrid model composed of CNN and LSTM layers was adapted as reported in [71]. The input features matched those provided to the LSTM baseline.

A-CNN-LSTM: An attention-based hybrid CNN-LSTM was implemented as reported in [54]. The model received as input spectrogram features.

Wav2Vec: An audio transformer model was adopted [65]. Wav2Vec processed raw audio signals resampled at 16 KHz with a feature encoder (7-layer CNN), a context encoder (12 transformer blocks), followed by a classification layer.

AST: An audio spectrogram transformer (AST) introduced for general audio tasks was adopted [64]. The model received as input spectrogram features as in HST. The architecture comprised vanilla transformer blocks processing the input with patch size 16×16. The last layer of AST was adapted to a classification layer with two output units.

HMT: A hierarchical multi-modal transformer (HMT) for COVID-19 detection was implemented [57]. The model fused representations from a transformer branch processing spectrograms and from an MLP branch processing MFCC features.

Ensemble: An ensemble model [72] was implemented that combined CNN and AST models described above. Output feature maps prior to classification layers in each model were concatenated, and input to an FC layer with two output units.

D. Modeling Procedures

Models were implemented in PyTorch except for LSTM, BiLSTM and CNN-LSTM that were implemented in Tensorflow. Models were executed on NVidia A4000 GPUs. Spectrogram intensities were normalized to a mean of 0.5 and standard deviation of 0.5. Modeling was performed separately on each dataset via 10-fold cross validation [38]. In each fold, data were partitioned into independent training, test and validation sets nearly of size (70%, 20%, 10%), with no participant-level overlap between the sets. Following common practice in literature, transformer models were initiated with pre-trained weights. AST and HST were pre-trained for object detection on ImageNet, Wav2Vec was pre-trained on sampled speech audio from the LibriSpeech dataset. Gradient clipping was used with an upper threshold of 0.1 for the gradient norm.

Model hyperparameters were selected based on F1 performance on the validation set (see Table I for representative results on learning rate, and Table II for selections). For SVM, regularization parameter C and kernel coefficient γ were selected via grid search. For network models, a common set of learning rate, regularization parameter for L_2 norm of model weights (λ_{L_2}), batch size, optimizer were selected across tasks. To avoid over-fitting, the number of epochs ranged between 1–100 and was selected separately for each task to stop training when a continual increase in validation loss or decrease in validation F1 was observed. Fig. 4 shows F1 and loss values of HST across epochs, where the trained models converge onto relatively high performance levels of around 0.85–0.90 F1 on the validation set. Model performance was measured using AUC, precision, recall, and F1 metrics on the test set. Individual metrics were averaged across the test set. Mean and standard deviation of metrics were

TABLE I
VALIDATION F1 SCORES OF COMPETING NEURAL NETWORK MODELS ACROSS LEARNING RATES. HIGHEST PERFORMING RATES FOR EACH MODEL ARE MARKED IN **BOLD** FONT

	10^{-5}	3×10^{-4}	10^{-3}	10^{-2}
CNN	0.77±0.10	0.79±0.10	**0.82±0.05**	0.75±0.13
DeepShufNet	0.68±0.11	0.71±0.13	0.67±0.15	**0.77±0.10**
LSTM	0.52±0.17	0.55±0.24	**0.56±0.10**	0.54±0.25
BiLSTM	0.47±0.08	0.58±0.07	**0.60±0.08**	0.57±0.08
CNN-LSTM	0.51±0.18	**0.69±0.09**	0.52±0.19	0.53±0.14
A-CNN-LSTM	0.54±0.28	0.68±0.08	0.71±0.10	**0.73±0.12**
Wav2Vec	**0.67±0.04**	0.61±0.05	0.57±0.03	0.58±0.03
AST	0.71±0.12	0.69±0.11	**0.74±0.14**	0.72±0.15
HMT	**0.77±0.09**	0.76±0.08	0.73±0.06	0.67±0.24
Ensemble	0.70±0.16	0.75±0.13	**0.77±0.16**	0.70±0.10
HST	**0.83±0.07**	0.74±0.08	0.70±0.10	0.72±0.15

TABLE II
HYPERPARAMETERS FOR NEURAL NETWORK MODELS

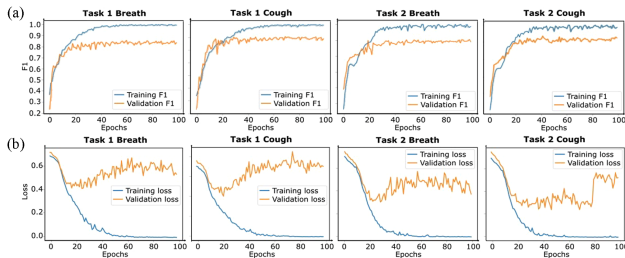|	Learning Rate	λ_{L_2}	Batch Size	Optimizer
CNN	10^{-3}	10^{-4}	8	Adam
DeepShufNet	10^{-2}	10^{-4}	8	Adam
LSTM	10^{-3}	-	8	Adam
BiLSTM	10^{-3}	-	8	Adam
CNN-LSTM	3×10^{-4}	-	8	Adam
A-CNN-LSTM	10^{-2}	10^{-4}	8	Adam
Wav2Vec	10^{-5}	10^{-6}	32	Adam
AST	10^{-3}	10^{-4}	8	Adam
HMT	10^{-5}	10^{-4}	5	Adam
Ensemble	10^{-3}	10^{-4}	6	Adam
HST	10^{-5}	10^{-8}	8	AdamW

Fig. 4. Training and validation. (a) F1 performance and (b) Loss values of HST across training epochs. Results shown separately for models based on breath and cough modalities, and for Task 1 (COVID vs. non-COVID without symptoms) and Task 2 (COVID vs. non-COVID with symptoms).

reported across cross-validation folds. Statistical significance of performance differences was assessed via non-parametric Wilcoxon signed-rank tests.

IV. RESULTS

A. COVID-19 Detection

We first demonstrated HST on the Cambridge dataset against state-of-the-art baselines including traditional (SVM), CNN (CNN, DeepShufNet), RNN (LSTM, BiLSTM), hybrid CNN-RNN (CNN-LSTM, A-CNN-LSTM), transformer (Wav2Vec, AST, HMT), and ensemble (Ensemble) methods. In Task 1, the COVID group was distinguished from the non-COVID group without symptoms based on either cough or breathing sounds. Performance for competing methods are reported in Table III with cough modality, in Table IV with breathing modality.

TABLE III
TEST PERFORMANCE OF COMPETING METHODS IN TASK 1 BASED ON COUGH SOUNDS FOR THE CAMBRIDGE DATASET. PERFORMANCE IN DISTINGUISHING THE COVID FROM THE non-COVID GROUP WITHOUT SYMPTOMS IS LISTED AS MEAN ± STD ACROSS CROSS-VALIDATION FOLDS

	AUC	Precision	Recall	F1
SVM	0.68±0.13	0.64±0.12	0.72±0.13	0.67±0.10
CNN	0.93±0.04	0.87±0.10	0.86±0.08	0.86±0.06
DeepShufNet	0.92±0.04	0.81±0.11	0.81±0.08	0.80±0.05
LSTM	0.71±0.10	0.64±0.07	0.68±0.09	0.66±0.07
BiLSTM	0.64±0.08	0.62±0.08	0.68±0.16	0.64±0.09
CNN-LSTM	0.66±0.12	0.64±0.09	0.61±0.19	0.61±0.10
A-CNN-LSTM	0.88±0.05	0.81±0.11	0.71±0.09	0.75±0.08
Wav2Vec	0.82±0.04	0.83±0.06	0.70±0.03	0.76±0.03
AST	0.86±0.04	0.89±0.11	0.70±0.13	0.77±0.05
HMT	0.89±0.07	0.77±0.11	0.77±0.13	0.76±0.05
Ensemble	0.90±0.04	0.84±0.12	0.74±0.16	0.77±0.08
HST	**0.97±0.03**	**0.92±0.08**	**0.94±0.07**	**0.93±0.04**

TABLE IV
PERFORMANCE IN TASK 1 BASED ON BREATHING SOUNDS, DISTINGUISHING THE COVID FROM THE non-COVID GROUP WITHOUT SYMPTOMS

|	AUC	Precision	Recall	F1
SVM	0.65±0.14	0.63±0.14	0.70±0.19	0.66±0.15
CNN	0.91±0.04	0.82±0.12	0.72±0.18	0.74±0.09
DeepShufNet	0.91±0.07	0.83±0.07	0.85±0.07	0.84±0.06
LSTM	0.68±0.14	0.65±0.19	0.60±0.16	0.62±0.16
BiLSTM	0.69±0.12	0.63±0.10	0.68±0.19	0.65±0.14
CNN-LSTM	0.72±0.09	0.68±0.08	0.75±0.13	0.71±0.08
A-CNN-LSTM	0.84±0.06	0.83±0.11	0.63±0.14	0.71±0.11
Wav2Vec	0.83±0.04	0.81±0.05	0.67±0.08	0.73±0.06
AST	0.85±0.01	0.82±0.10	0.70±0.19	0.75±0.13
HMT	0.86±0.09	0.82±0.09	0.71±0.17	0.75±0.13
Ensemble	0.87±0.05	0.80±0.06	0.80±0.07	0.80±0.04
HST	**0.97±0.02**	**0.94±0.06**	**0.95±0.04**	**0.94±0.04**
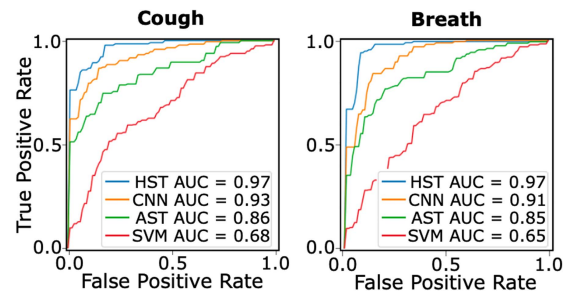
Fig. 5. ROC curves for Task 1 to distinguish the COVID group from the non-COVID group without symptoms, based on cough (left), or breathing sounds (right). AUC is listed for each model (see legend).

Receiver operating characteristics (ROC) curves for SVM, AST, CNN and HST are displayed in Fig. 5. Overall, HST significantly outperforms competing methods ($p < 0.05$). As seen in ROC curves, HST also maintains a more favorable trade-off between true and false positive rates. These results indicate that HST enables improved capture of contextual features in spectrograms to improve detection of COVID-19. Note that comparable detection performance is attained via HST with cough versus breathing modalities. This finding implies that both modalities carry discriminative information regarding respiratory symptoms of COVID-19. In Task 2, the COVID group was distinguished from the non-COVID group with symptoms. Performance metrics are listed in Table V for cough modality, and in Table VI

TABLE V

PERFORMANCE IN TASK 2 BASED ON COUGH SOUNDS, DISTINGUISHING THE COVID GROUP FROM THE non-COVID GROUP WITH SYMPTOMS

	AUC	Precision	Recall	F1
SVM	0.69±0.15	0.62±0.16	0.65±0.17	0.62±0.15
CNN	0.95±0.05	0.87±0.12	0.92±0.10	0.89±0.09
DeepShufNet	0.96±0.03	0.86±0.09	**0.98±0.04**	0.91±0.05
LSTM	0.57±0.20	0.56±0.18	0.71±0.27	0.62±0.20
BiLSTM	0.38±0.23	0.69±0.31	0.43±0.20	0.49±0.20
CNN-LSTM	0.55±0.21	0.67±0.17	0.43±0.12	0.49±0.13
A-CNN-LSTM	0.88±0.08	0.80±0.11	0.88±0.12	0.83±0.09
Wav2Vec	0.89±0.05	0.76±0.12	0.80±0.08	0.77±0.08
AST	0.83±0.11	0.86±0.06	0.86±0.07	0.86±0.05
HMT	0.87±0.07	0.80±0.09	0.85±0.07	0.82±0.07
Ensemble	0.94±0.06	0.78±0.15	0.96±0.09	0.85±0.09
HST	**0.98±0.03**	**0.94±0.08**	0.93±0.07	**0.94±0.07**

TABLE VI

PERFORMANCE IN TASK 2 BASED ON BREATHING SOUNDS, DISTINGUISHING THE COVID FROM THE non-COVID GROUP WITH SYMPTOMS

|	AUC	Precision	Recall	F1
SVM	0.71±0.18	0.74±0.20	0.58±0.19	0.62±0.13
CNN	0.92±0.06	0.85±0.08	0.91±0.09	0.87±0.06
DeepShufNet	0.96±0.04	0.88±0.06	0.95±0.06	0.91±0.05
LSTM	0.64±0.17	0.45±0.16	0.44±0.19	0.44±0.17
BiLSTM	0.61±0.21	0.56±0.33	0.48±0.32	0.49±0.28
CNN-LSTM	0.75±0.16	0.49±0.19	0.53±0.30	0.49±0.21
A-CNN-LSTM	0.87±0.08	0.80±0.11	0.92±0.06	0.85±0.06
Wav2Vec	0.92±0.05	0.80±0.17	0.88±0.04	0.83±0.09
AST	0.93±0.05	0.87±0.04	0.94±0.06	0.90±0.03
HMT	0.86±0.06	0.78±0.09	0.88±0.09	0.83±0.06
Ensemble	0.91±0.04	0.86±0.10	0.90±0.12	0.87±0.06
HST	**0.97±0.02**	**0.91±0.06**	**0.96±0.09**	**0.93±0.04**
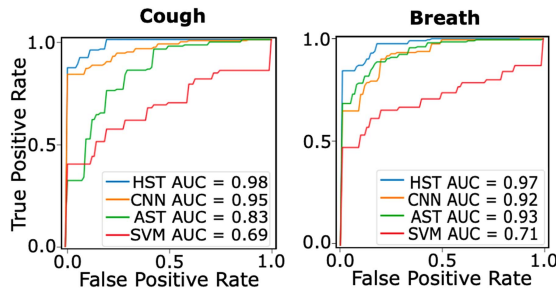
Fig. 6. ROC curves for Task 2 to distinguish the COVID-19 group from the non-COVID group with cough symptoms, based on cough sounds (left), and breathing sounds (right). AUC is listed for each model (see legend).

for breathing modality. ROC curves are displayed in Fig. 6. HST significantly outperforms competing methods ($p < 0.05$), except for DeepShufNet and Ensemble that yield higher recall. HST also maintains a modestly better trade-off between true and false positive rates. Taken together, these results indicate that contextual representations of audio spectrograms captured by HST enable discrimination of respiratory symptoms in COVID-19 patients versus healthy controls.

Next, we demonstrated HST on the separate COUGHVID dataset against the same set of traditional, CNN, RNN, hybrid CNN-RNN, transformer, and ensemble methods. A single task was implemented to distinguish the COVID group and non-COVID group with cough symptoms, based on cough sounds. Performance metrics for competing methods are reported in Table VII. All methods yield relatively lower performance

TABLE VII

PERFORMANCE IN DISTINGUISHING THE COVID FROM THE non-COVID GROUP BASED ON COUGH SOUNDS FOR THE COUGHVID DATASET

|	AUC	Precision	Recall	F1
SVM	0.55±0.01	0.59±0.01	0.56±0.01	0.57±0.02
CNN	0.75±0.03	0.69±0.04	0.68±0.04	0.69±0.04
DeepShufNet	0.76±0.04	0.74±0.06	0.70±0.06	0.72±0.05
LSTM	0.55±0.01	0.55±0.02	0.47±0.07	0.50±0.04
BiLSTM	0.56±0.01	0.56±0.02	0.55±0.08	0.55±0.01
CNN-LSTM	0.70±0.03	0.66±0.04	0.68±0.05	0.67±0.01
A-CNN-LSTM	0.83±0.02	**0.81±0.06**	0.76±0.04	0.78±0.01
Wav2Vec	0.64±0.07	0.61±0.10	0.65±0.09	0.61±0.08
AST	0.79±0.02	0.74±0.03	0.74±0.02	0.74±0.03
HMT	0.75±0.05	0.74±0.06	0.72±0.09	0.73±0.04
Ensemble	0.78±0.03	0.73±0.03	0.72±0.02	0.72±0.03
HST	**0.90±0.01**	0.80±0.03	**0.84±0.03**	**0.82±0.01**
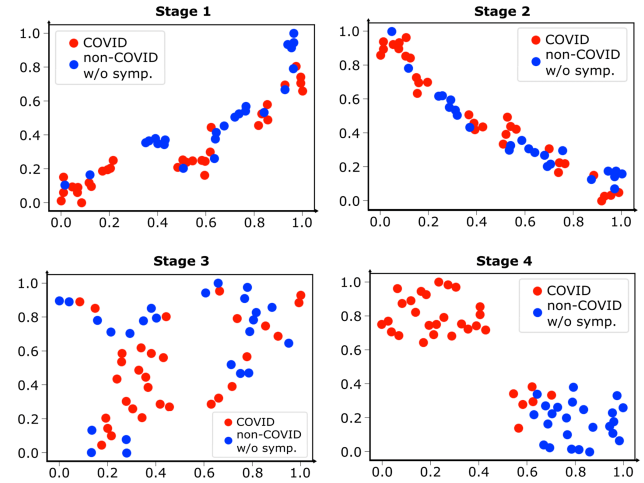
Fig. 7. Visualization of latent representations captured by HST across Stages 1–4. Embeddings of samples from the COVID group (red) and the non-COVID group without symptoms (blue) are displayed.

on the COUGHVID dataset, implying elevated task difficulty compared to the Cambridge dataset. That said, HST again outperforms competing methods significantly in all metrics ($p < 0.05$), except for A-CNN-LSTM that yields slightly higher precision. These results suggest that HST offers more reliable capture of COVID-related cues in respiratory sounds against competing methods.

B. Model Interpretation

Across multiple stages, the proposed HST model nonlinearly transforms spectrogram features of audio recordings to extract their latent representations. Ideally, these representations should capture hidden relationships among input features that serve to improve discrimination between output classes. To examine this issue, we visualized the latent representations extracted across separate stages in HST. A random subset of 55 audio recordings were projected through HST, and the evoked hidden unit responses for each sample were stored as stage-specific response vectors. At each stage, the response vectors of all samples were embedded via t-SNE onto two dimensions [73]. Fig. 7 displays representative embeddings of samples from the COVID and non-COVID without symptoms groups based on breath modality in the Cambridge dataset. The latent representations of COVID and non-COVID samples become progressively more distinct across stages, indicating that hierarchical transformations in
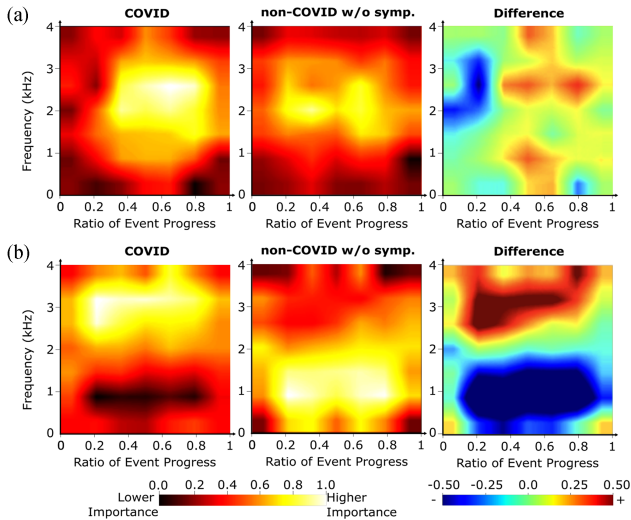
Fig. 8. Activation maps in HST for interpreting the spectrogram features that contribute to detection of the COVID group and the non-COVID group without symptoms. The first PC of activation maps prior to the final attention layer were computed. The difference maps between PCs for the two groups are also shown. Horizontal axis shows ratio of event progress, i.e. the fraction with respect to the total duration of the recording. Results are shown for models based on. (a) Cough sounds. (b) Breathing sounds in Task 1.

HST extract latent time-frequency features critical in disease detection.

Next, a post-hoc explanatory analysis was performed to interpret the spectrogram features that most significantly contribute to model decisions. On correctly classified test samples from the COVID and non-COVID groups, Grad-CAM was used to produce activation maps that indicate the relative importance levels of spectrogram features for detecting each group [74]. To obtain characteristic maps, principal component analysis (PCA) was performed on activation maps across samples within each group. The first PCs along with their difference between the two groups are shown in Fig. 8 for Task 1, and in Fig. 9 for Task 2. For the cough modality Fig. 8(a) and 9(a), HST attends to similar frequency bands for the two groups, centered around 1.5–3 kHz in Task 1 and 1–2 kHz in Task 2. However, across the audio recording, the COVID group shows more uniformly strong attention for the frequency band than the non-COVID group. As highlighted by the difference maps, the COVID group has predominantly higher activation at 0–1 kHz and 2–4 kHz bands for 0.4–0.9 event progress in Task 1, and at 2–3 kHz band for 0–0.1 event progress, 0–2 kHz band for 0.2–0.6 event progress in Task 2. For the breathing modality Fig. 8(b) and 9(b), HST generally shows elevated attention towards higher frequency bands in the COVID group (around 2–4 kHz in both tasks) versus the non-COVID group (around 0–2 kHz in both tasks). As visible in the difference maps, the COVID group has higher activation at 2–4 kHz band for 0.2-0.8 event progress in Task 1, and at 2.5–3.5 kHz band for roughly 0–0.1, 0.4–1.0 event progress and at 3.5–4 kHz band for 0.2–0.4, 0.6–0.8 event progress in Task 2. Note that the differences between the two groups are more accentuated for the breathing versus cough modality. This observation might suggest that there might be more diverse sound events during breathing as compared to cough. Although clinical research regarding the frequency features of respiratory sounds in COVID-19 is ongoing, a prior study has examined respiratory symptoms of COVID-19 patients in different stages
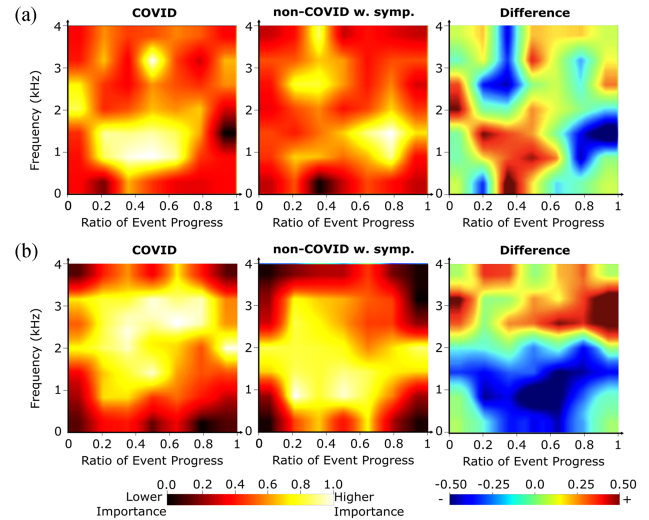

Fig. 9. Activation maps in HST for the COVID group and the non-COVID group with symptoms, along with the difference map between the two groups. Results are shown for models based on. (a) Cough sounds. (b) Breathing sounds in Task 2.

TABLE VIII
TEST F1 SCORES OF HST IN TASKS 1–2 BASED ON COUGH OR BREATHING SOUNDS WITH DIFFERENT TIME-FREQUENCY REPRESENTATIONS

	Mel	CQT	Gammatone	Bark	Linear
Task 1 Cough	**0.93±0.04**	0.84±0.06	0.87±0.07	0.85±0.03	0.86±0.06
Task 1 Breath	**0.94±0.04**	0.85±0.03	0.83±0.07	0.86±0.03	0.86±0.02
Task 2 Cough	**0.94±0.07**	0.92±0.07	0.93±0.02	0.92±0.01	0.92±0.02
Task 2 Breath	**0.93±0.04**	0.90±0.03	0.90±0.06	0.91±0.04	0.90±0.08

TABLE IX
TEST F1 SCORES OF HST IN TASKS 1–2 BASED ON COUGH OR BREATHING SOUNDS WITH DIFFERENT SPECTROGRAM WINDOW LENGTHS, IN NUMBER OF SAMPLES (N) AND MILLISECONDS (MS)

	N=1024 (46.4 ms)	N=2048 (92.8 ms)	N=4096 (185.7 ms)
Task 1 Cough	0.86±0.03	**0.93±0.04**	0.82±0.06
Task 1 Breath	0.85±0.05	**0.94±0.04**	0.84±0.06
Task 2 Cough	0.93±0.04	**0.94±0.07**	0.93±0.02
Task 2 Breath	0.89±0.07	**0.93±0.04**	0.92±0.07

of disease [75]. Frequency of respiration sounds was reported to be increased with disease development. While our results are consistent with this finding, future work on larger patient cohorts is warranted to examine the validity of the observations reported here.

C. Ablation Studies

Ablation studies were conducted to examine the influences of input spectrogram type, spectrogram window length, model size, and pre-training on the test performance of HST on the Cambridge dataset. The results for all four performance metrics (i.e., AUC, precision, recall, F1) are consistent, albeit only F1 scores for the variant models are listed for brevity. Table VIII lists performance for the proposed Mel-spectrogram and alternative CQT, Gammatone, Bark and Linear spectrograms. Table IX lists performance for the proposed 2048 window length and alternative 1024 and 4096 lengths. Table X lists performance

TABLE X

TEST F1 SCORES OF HST IN TASKS 1–2 BASED ON COUGH OR BREATHING SOUNDS. SMALL, BASE, AND LARGE VARIANTS WERE CONSIDERED

	Small	Base	Large
Task 1 Cough	0.85±0.09	**0.93±0.04**	0.86±0.03
Task 1 Breath	0.86±0.05	**0.94±0.04**	0.87±0.05
Task 2 Cough	0.89±0.05	**0.94±0.07**	0.89±0.05
Task 2 Breath	0.91±0.07	**0.93±0.04**	0.91±0.02

TABLE XI

TEST F1 SCORES OF HST IN TASKS 1–2 BASED ON COUGH OR BREATHING SOUNDS, WITHOUT AND WITH PRE-TRAINING

	No Pre-training	Pre-training
Task 1 Cough	0.77±0.08	**0.93±0.04**
Task 1 Breath	0.70±0.10	**0.94±0.04**
Task 2 Cough	0.83±0.07	**0.94±0.07**
Task 2 Breath	0.83±0.07	**0.93±0.04**

TABLE XII

TEST F1 SCORES OF HST IN TASKS 1–2. MODALITY-SPECIFIC MODELS BASED ON COUGH OR BREATHING SOUNDS, AND MODALITY-AGNOSTIC MODELS BASED ON A MIXTURE OF COUGH/BREATHING SOUNDS ARE SHOWN

	Task 1	Task 2
Modality: Cough	0.93±0.04	0.94±0.07
Modality: Breath	0.94±0.04	0.93±0.04
Modality-agnostic	0.91±0.04	0.92±0.04

TABLE XIII

COMPLEXITY OF HST AND A VANILLA TRANSFORMER VARIANT. PRE-TRAINING TIME, TRAINING TIME, PER-SAMPLE INFERENCE TIME ARE LISTED ALONG WITH FLOPS, MEMORY USE, AND NUMBER OF PARAMETERS

	Pre-training	Training	Inference	FLOPS	Memory	N_{params}
HST	114 h	5.5 m	19.4 ms	8.8 G	1.7 GB	50 M
Vanilla	455 h	13.7 m	22.7 ms	35 G	3.6 GB	171 M

for the proposed base model size and alternative small and large sizes. Table XI lists performance with and without pre-training, where HST is pre-trained for object classification on natural images. For all tasks and modalities, the proposed configuration of the HST model yields higher performance in all four tasks against variant models.

The datasets examined here contain recordings from participants instructed to produce either cough or breathing sounds. Yet, there can be practical scenarios in which the modality of a given recording might be unknown. To assess the influence of modality labels on HST, we compared modality-specific and modality-agnostic models. Modality-specific models were trained and tested on a single known modality (either cough or breathing). Meanwhile, modality-agnostic models were trained and tested on a mixture of breathing and cough sounds without labels. Table XII lists the resultant performance metrics. The performance differences between modality-specific and modality-agnostic models are modest, suggesting that HST shows reasonable reliability against missing modality labels.

Finally, we examined the computational complexity of HST against a vanilla transformer variant that replaced local-windowed MSA layers with global MSA layers. Table XIII lists total pre-training time on ImageNet, total training time and per-sample inference time on the Cambridge dataset, along with FLOPS, memory use, and number of parameters. Compared to the vanilla variant, HST requires lower FLOPS, memory use and fewer parameters to process spectrograms, and so it offers faster pre-training, training and inference.

V. DISCUSSION

Clinical testing of COVID-19 involves PCR and/or imaging procedures typically administered in centralized healthcare institutions [76], [77]. These cost-intensive procedures might not be broadly accessible in developing countries, and access might be delayed till relatively late stages of disease. In this context, remote monitoring via audio recordings of respiratory sounds can help pave the way to accessible preliminary screening [60]. Here, we introduce HST to detect COVID-19 with high accuracy from brief audio recordings of cough and breathing sounds. Similar to manual auscultation, the proposed method does not serve as a diagnostic test. Yet, it holds promise as a preliminary assessment tool for potential COVID-19 cases that can assist in informed allocation of limited resources for timely testing and interventions [78]. Rapid antigen tests are another low-cost alternative for COVID-19 screening, but they primarily detect high concentrations of viral load and typically miss early disease stages [79]. Instead, respiratory sounds are suggested to contain cues for infection starting at earlier stages [33]. It remains important future work to systematically compare the utility of respiratory sound analysis versus antigen tests in COVID-19 screening.

Transformers can capture contextual features in sound signals more effectively than CNN or RNN models [64]. Yet, vanilla transformers with global attention suffer from quadratic complexity [80], restricting their use under limited compute budgets on mobile devices. Instead, HST achieves high computational efficiency by leveraging local-windowed attention for linear complexity, and a hierarchical structure for progressively lowered spectrogram resolution. That said, analyses of respiratory sounds here were conducted offline for proof-of-concept demonstration. In practice, remote monitoring involves online processing of audio recordings either on mobile devices or via communication with a cloud computing platform [37], [45]. To reduce computational and communication load, network distillation or hybrid CNN-transformer architectures might be adopted [81], [82].

Here, demonstrations were performed on the Cambridge [38] and COUGHVID [58] datasets. Across competing methods, we find generally higher detection performance on Cambridge versus COUGHVID. Several differences between the datasets might have contributed to this pattern. First, single audio recordings in Cambridge of duration 10.0 ± 6.0 s contain repeated respiratory events (5 repeats for breathing, 3 repeats for cough), whereas those in COUGHVID of duration 6.6 ± 2.5 s contain a single cough event per recording. Acoustic signatures of COVID-19 might manifest not only over short but also over longer time intervals, and prolonged recordings might facilitate separation of foreground-background signals. Thus, analyses based on multiple events per sample can improve detection sensitivity. Second, while Cambridge uses participant-reported COVID-19 status labels based on positive PCR test results, COUGHVID relies on expert-provided labels based on inspection of audio recordings. Thus, analyses on COUGHVID might suffer from higher label inaccuracy that can limit model performance. Third, native differences in microphone type,

microphone positioning, recording environment, and compliance to recording instructions might have influenced the data quality, and thereby detection sensitivity. Future studies are warranted to identify ideal recording procedures for COVID detection from respiratory sounds.

The Cambridge and COUGHVID datasets primarily comprise COVID and non-COVID groups (with or without cough symptoms), so the binary classification tasks examined here concerned segregation of these groups. While we find high detection accuracy for the COVID group against the non-COVID group with symptoms, a clinical characterization of respiratory disorders that underlie the cough symptoms is unavailable in the datasets. A practical concern regarding the adoption of a screening technology is its reliability on patient cohorts with varying types of disease [79], [83]. Recent studies suggest that respiratory sounds carry symptomatic cues to distinguish COVID-19 from other conditions such as asthma, chronic obstructive pulmonary disease (COPD), bronchitis and pertussis [18], [19]. It remains important future work to examine whether HST can discriminate among a broad spectrum of respiratory diseases based on cough or breathing sounds alone. In cases where these two modalities do not carry sufficient information to detect COVID, speech or wheeze modalities might also be incorporated to boost model performance [26], [31], [35], [50]. Furthermore, additional clinical information such as patient demographics can be integrated to boost sensitivity of HST [33], [56].

Here, we find that HST outperforms competing methods based on analyses conducted on relatively modest-to-moderate sized datasets. An essential next step for validation of HST is demonstration of its reliability on broader patient cohorts. In practice, cross-checking the specific input features of an audio recording that drive the model output against known respiratory sound markers of COVID can be critical to avoid erroneous decisions when using black-box deep learning models. Explanatory analyses presented here indicate that the Grad-CAM algorithm applied on HST can identify important input features for COVID detection that closely match the expected discriminating features of respiratory sounds in COVID-19 [18]. Yet, it remains important future work to evaluate the efficacy of gradient-based explanatory methods such as Grad-CAM and uncertainty characterization methods in interrogating the decisions provided by HST.

VI. CONCLUSION

In this study, we proposed a deep learning approach to detect COVID-19 from respiratory sounds. A novel hierarchical transformer model, HST, was introduced to extract contextual features from spectrogram representations of audio signals. HST leverages local attention mechanisms over progressively growing windows to capture long-range context without excessive computational burden. While our demonstrations focused on COVID-19 screening, the contextual sensitivity of HST might also be helpful in detection of other pervasive respiratory disorders such as pneumonia, bronchitis or obstructive pulmonary disease.

REFERENCES

[1] T. H. Falk et al., "Spectro-temporal analysis of auscultatory sounds," *New Dev. Biomed. Eng.*, vol. 5, pp. 93–104, 2010.

[2] L. Sun et al., "Clinical impact of multi-parameter continuous non-invasive monitoring in hospital wards: A systematic review and meta-analysis," *J. Roy. Soc. Med.*, vol. 113, no. 6, pp. 217–224, 2020.

[3] Y. Zhang et al., "COVID-DA: Deep domain adaptation from typical pneumonia to COVID-19," 2020, *arXiv:2005.01577*.

[4] A. Aliverti, "Wearable technology: Role in respiratory health and disease," *Breathe*, vol. 13, no. 2, pp. e27–e36, 2017.

[5] Y. Bai et al., "Presumed asymptomatic carrier transmission of COVID-19," *JAMA*, vol. 323, no. 14, pp. 1406–1407, 2020.

[6] S. A. Harmon et al., "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nature Commun.*, vol. 11, no. 1, pp. 1–7, 2020.

[7] S. Tabik et al., "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3595–3605, Dec. 2020.

[8] L. Sun et al., "Adaptive feature selection guided deep forest for COVID-19 classification with chest CT," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2798–2805, Oct. 2020.

[9] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[10] Z. Wang, Q. Liu, and Q. Dou, "Contrastive cross-site learning with redesigned net for COVID-19 CT classification," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2806–2813, Oct. 2020.

[11] Q. Zhu et al., "GACDN: Generative adversarial feature completion and diagnosis network for COVID-19," *BMC Med. Imag.*, vol. 21, no. 1, pp. 1–13, 2021.

[12] T. Goel et al., "Automatic screening of COVID-19 using an optimized generative adversarial network," *Cogn. Comput.*, vol. 4, pp. 1–16, 2021.

[13] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," *Physiol. Meas.*, vol. 40, no. 3, 2019, Art. no. 035001.

[14] K. Jackson et al., "The use of lung ultrasound during the COVID-19 pandemic," *Postgraduate Med. J.*, vol. 97, no. 1143, pp. 34–39, 2020.

[15] W.-C. Dai et al., "CT imaging and differential diagnosis of COVID-19," *Can. Assoc. Radiol. J.*, vol. 71, no. 2, pp. 195–200, 2020.

[16] H. X. Bai et al., "Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT," *Radiol.*, vol. 296, no. 2, pp. E46–E54, 2020.

[17] W. Zhao et al., "The importance of distinguishing COVID-19 from more common respiratory illnesses," *Epidemiol. Infect.*, vol. 148, 2020, Art. no. e195.

[18] B. Wang et al., "Characteristics of pulmonary auscultation in patients with 2019 novel coronavirus in China," *Respiration*, vol. 99, no. 9, pp. 755–763, 2020.

[19] G. Rudraraju et al., "Cough sound analysis and objective correlation with spirometry and clinical diagnosis," *Inform. Med. Unlocked*, vol. 19, 2020, Art. no. 100319.

[20] A. Bohadana et al., "Fundamentals of lung auscultation," *New England J. Med.*, vol. 370, no. 8, pp. 744–751, 2014.

[21] A. Roy et al., "A novel melspectrogram snippet representation learning framework for severity detection of chronic obstructive pulmonary diseases," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 4003311.

[22] C. Habukawa et al., "A wheeze recognition algorithm for practical implementation in children," *PLOS One*, vol. 15, 2020, Art. no. e0240048.

[23] G. Serbes et al., "Pulmonary crackle detection using time-frequency analysis," *Digit. Signal Process.*, vol. 23, pp. 1012–1021, 2012.

[24] R. X. A. Pramono et al., "A cough-based algorithm for automatic diagnosis of pertussis," *PLoS One*, vol. 11, no. 9, 2016, Art. no. e0162128.

[25] M. Aykanat et al., "Classification of lung sounds using convolutional neural networks," *EURASIP J. Image Vid. Process.*, vol. 2017, no. 1, pp. 1–9, 2017.

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

[27] C. Bales et al., "Can machine learning be used to recognize and diagnose coughs," in *Proc. IEEE Int. Conf. E-Health Bioeng.* 2020, pp. 1–4.

[28] V. Basu and S. Rana, "Respiratory diseases recognition through respiratory sound with the help of deep neural network," in *Proc. IEEE 4th Int. Conf. Comput. Intell. Netw.*, 2020, pp. 1–6.

[29] A. Srivastava et al., "Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease," *PeerJ Comput. Sci.*, vol. 7, 2021, Art. no. e369.

[30] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust deep learning framework for predicting respiratory anomalies and diseases," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 164–167.

[31] V. Despotovic et al., "Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results," *Comput. Biol. Med.*, vol. 138, 2021, Art. no. 104944.

[32] E. A. Mohammed et al., "An ensemble learning approach to digital corona virus preliminary screening from cough sounds," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[33] A. Pal et al., "Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing," in *Proc. ACM Symp. Appl. Comput.*, 2021, pp. 620–628.

[34] N. K. Chowdhury et al., "Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method," *Comput. Biol. Med.*, vol. 145, 2022, Art. no. 105405.

[35] S. Deshmukh, M. Al Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting COVID-19 from voice," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1055–1059.

[36] T. K. Dash et al., "Detection of COVID-19 from speech signal using bio-inspired based cepstral features," *Pattern Recognit.*, vol. 117, 2021, Art. no. 107999.

[37] A. Imran et al., "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Inform. Med. Unlocked*, vol. 20, 2020, Art. no. 100378.

[38] C. Brown et al., "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 3474–3484.

[39] P. Bagad et al., "Cough against COVID: Evidence of COVID-19 signature in cough sounds," 2020, *arXiv:2009.08790*.

[40] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[42] A. Ponomarchuk et al., "Project Achoo: A practical model and application for COVID-19 detection from recordings of breath, voice, and cough," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 2, pp. 175–187, Feb. 2022.

[43] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.

[44] Y. E. Erdogan et al., "COVID-19 detection with traditional and deep features on cough acoustic signals," *Comput. Biol. Med.*, vol. 136, 2021, Art. no. 104765.

[45] M. E. Chowdhury et al., "QUCoughScope: An artificially intelligent mobile application to detect asymptomatic COVID-19 patients using cough and breathing sounds," *Diagn.*, vol. 12, no. 4, 2022, Art. no. 920.

[46] B. W. Schuller et al., "Detecting COVID-19 from breathing and coughing sounds using deep neural networks," 2020, *arXiv:2012.14553*.

[47] H. Coppock et al., "End-2-end COVID-19 detection from breath & cough audio," 2021, *arXiv:2102.08359*.

[48] O. O. Abayomi-Alli et al., "Detection of COVID-19 from deep breathing sounds using sound spectrum with image augmentation and deep learning techniques," *Electron.*, vol. 11, no. 16, 2022, Art. no. 2520.

[49] M. Pahar et al., "COVID-19 cough classification using machine learning and global smartphone recordings," *Comput. Biol. Med.*, vol. 135, 2021, Art. no. 104572.

[50] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 detection system using recurrent neural networks," in *Proc. IEEE Int. Conf. Commun. Comput. Cybersecurity Inform.*, 2020, pp. 1–5.

[51] V. Karas et al., "Recognising COVID-19 from coughing using ensembles of SVMs and LSTMs with handcrafted and deep audio features," in *Proc. Interspeech*, 2021, pp. 911–915.

[52] L. Khriji, A. Ammari, S. Messaoud, S. Bouaafia, A. Maraoui, and M. Machhout, "COVID-19 recognition based on Patient's coughing and breathing patterns analysis: Deep learning approach," in *Proc. IEEE 29th Conf. Open Innov. Assoc.*, 2021, pp. 185–191.

[53] M. R. Kamble, J. Patino, M. A. Zuluaga, and M. Todisco, "Exploring auditory acoustic features for the diagnosis of the COVID-19," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 566–570.

[54] S. Hamdi et al., "Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound," *J. Intell. Inf. Syst.*, vol. 23, pp. 1–23, 2022.

[55] Y. Bhanusree et al., "A deep learning approach to predict COVID-19 through cough analysis using CNN-bidirectional LSTM," in *Proc. Int. Conf. Adv. Technol.*, 2022, pp. 1–5.

[56] G. Pinkas et al., "SARS-CoV-2 detection from voice," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 268–274, 2020.

[57] S. Tang et al., "Hierarchical multi-modal transformer for automatic detection of COVID-19," in *Proc. Int. Conf. Signal Process.*, 2022, pp. 197–202.

[58] L. Orlandic et al., "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Sci. Data*, vol. 8, no. 1, pp. 1–10, 2021.

[59] B. McFee et al., "librosa: Audio and music signal analysis in python," in *Proc. Python Sci. Conf.*, 2015, pp. 18–24.

[60] X. Ni et al., "Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients," *Proc. Nat. Acad. Sci.*, vol. 118, no. 19, 2021, Art. no. e2026610118.

[61] R. X. A. Pramono et al., "Evaluation of features for classification of wheezes and normal respiratory sounds," *PLoS One*, vol. 14, no. 3, 2019, Art. no. e0213659.

[62] M. Bahoura, "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes," *Comput. Biol. Med.*, vol. 39, no. 9, pp. 824–843, 2009.

[63] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.

[64] Y. Gong et al., "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.

[65] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[66] J. Yang et al., "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[67] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[68] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1747–1763, Jul. 2022.

[69] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2598–2614, Oct. 2022.

[70] B. Balamurali et al., "Deep neural network-based respiratory pathology classification using cough sounds," *Sensors*, vol. 21, no. 16, 2021, Art. no. 5555.

[71] M. Z. Islam et al., "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Inform. Med. Unlocked*, vol. 20, 2020, Art. no. 100412.

[72] S. D. Deb and R. K. Jha, "COVID-19 detection from chest X-ray images using ensemble of CNN models," in *Proc. Int. Conf. Power Instrum. Control Comput.*, 2020, pp. 1–5.

[73] L. van der Maaten et al., "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[75] C. M. Davidson, D. Kaushik, H. Van Truong, S. Jain, and K. Guha, "Acoustic features of cough in COVID-19," in *Proc. 1st Int. Conf. Adv. Comput. Future Commun. Technol.*, 2021, pp. 33–38.

[76] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2376–2387, Jul. 2021.

[77] S. Mohagheghi, M. Alizadeh, S. M. Safavi, A. H. Foruzan, and Y. -W. Chen, "Integration of CNN, CBMIR, and visualization techniques for diagnosis and quantification of COVID-19 disease," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 1873–1880, Jun. 2021.

[78] B. H. Crotty et al., "Hospitalization outcomes among patients with COVID-19 undergoing remote monitoring," *JAMA Netw. Open*, vol. 5, no. 7, 2022, Art. no. e2221050.

[79] M. Akay et al., "Healthcare innovations to address the challenges of the COVID-19 pandemic," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3294–3302, Jul. 2022.

[80] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[81] G. Hinton et al., "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[82] A. Güngör, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, "TranSMS: Transformers for super-resolution calibration in magnetic particle imaging," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3562–3574, Dec. 2022.

[83] T. Liu et al., "Deep learning and medical image analysis for COVID-19 diagnosis and prediction," *Annu. Rev. Biomed. Eng.*, vol. 24, pp. 179–201, 2022.