Title	Wave2Graph: Integrating spectral features and correlations for graph-based learning in sound waves
Authors	Van Truong, Hoang;Tran, Khanh-Tung;Vu, Xuan-Son;Nguyen, Duy-Khuong;Bhuyan, Monowar;Nguyen, Hoang D.
Publication date	2024-09-03
Original Citation	Hoang, V.-T., Tran, K.-T., Vu, X.-S., Nguyen, D.-K., Bhuyan, M. and Nguyen, H.D. (2024) 'Wave2Graph: Integrating spectral features and correlations for graph-based learning in sound waves', AI Open, 5, pp. 115–125. Available at: https://doi.org/10.1016/j.aiopen.2024.08.004
Type of publication	Article (peer-reviewed);journal-article
Link to publisher's version	10.1016/j.aiopen.2024.08.004
Rights	© 2024 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) - http://creativecommons.org/licenses/by/4.0/
Download date	2026-08-10 12:28:59
Item downloaded from	https://hdl.handle.net/10468/16306

AI Open 5 (2024) 115–125

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Full length article

Wave2Graph: Integrating spectral features and correlations for graph-based learning in sound waves

Van-Truong Hoang [a,1], Khanh-Tung Tran [b,1], Xuan-Son Vu [c], Duy-Khuong Nguyen [a], Monowar Bhuyan [c], Hoang D. Nguyen [b,*]

[a] AI Center, FPT Software Company Limited, Viet Nam
[b] School of Computer Science and Information Technology, University College Cork, Ireland
[c] Department of Computing Science, Umeå, University, Sweden

ARTICLE INFO

Keywords:
Wave2Graph
Graph neural network
Correlation
Sound signal processing
Neural network architecture

ABSTRACT

This paper investigates a novel graph-based representation of sound waves inspired by the physical phenomenon of correlated vibrations. We propose a Wave2Graph framework for integrating multiple acoustic representations, including the spectrum of frequencies and correlations, into various neural computing architectures to achieve new state-of-the-art performances in sound classification. The capability and reliability of our end-to-end framework are evidently demonstrated in voice pathology for low-cost and non-invasive mass-screening of medical conditions, including respiratory illnesses and Alzheimer's Dementia. We conduct extensive experiments on multiple public benchmark datasets (ICBHI and ADReSSo) and our real-world dataset (IJSound: Respiratory disease detection using coughs and breaths). Wave2Graph framework consistently outperforms previous state-of-the-art methods with a large magnitude, up to 7.65% improvement, promising the usefulness of graph-based representation in signal processing and machine learning.

1. Introduction

In nature, sounds are produced by the vibrations of multiple physical objects, creating unique characteristics perceivable to humans or animals. Therefore, in the digital era, the mathematical representations of sounds such as Mel frequency spectral coefficients (MFCCs) or Mel-scaled spectrograms (Mel-Spectrogram) have been extensively investigated for various downstream tasks in signal processing and machine learning (Tang et al., 2021; Madanian et al., 2023). These transformations have shown their advantages in providing more useful information in both the time and frequency domains for various downstream tasks, including sound classification and voice generation. Nevertheless, conventional methods (e.g., Holi et al., 2013; Sharma et al., 2022; Mahanta et al., 2021) typically adopt these 2D representations as is, neglecting the underlying information regarding multiple vibrations co-occurring, which may reveal fine-grain signals of objects' structures and conditions along the sound pathway for machine learning. For instance, in voice pathology, human sounds are generated by vibrations traveling through internal organs (e.g., lung and throat) that co-operate, creating correlation patterns between sources. These sound correlation patterns may be beneficial as indicators of medical conditions or symptoms in detecting illnesses and other applications (Al-nasheri et al., 2017; Lv et al., 2021). To this end, we propose a novel approach to learning spectral features and correlations which is capable of extending the boundaries of existing sound classification methods and, in particular, voice pathology detection for better understanding and performance of acoustic waves.

In detail, Fig. 1 illustrates the differences in human coughs between COVID-19 and non-COVID-19 sound samples using various representations. The figure shows how MFCC or Mel-Spectrogram features can distinguish between positive and negative COVID-19 cases, in which temporal and spatial trends are observable (e.g., cough lengths, strengths and gaps). These features have been explored with convolutional neural network-based (CNN-based) models (He et al., 2016) to achieve the current state-of-the-art performance (Imran et al., 2020; Fakhry et al., 2021; Ding et al., 2021). Apart from them, we discover non-linear patterns in cepstral correlations beyond the temporal and spatial perspectives of the differences between disease-related and healthy cases. The correlations of MFCC or Mel coefficients can be seen to vary over time or between different frame-windows, explaining the disease pathology. For example, in the time range from 50 to 125 of the MFCC feature or in frames 50 to 125 of the Mel-Spectrogram

* Corresponding author.
E-mail address: hn@cs.ucc.ie (H.D. Nguyen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.aiopen.2024.08.004
Received 30 December 2023; Received in revised form 19 July 2024; Accepted 21 August 2024
Available online 3 September 2024

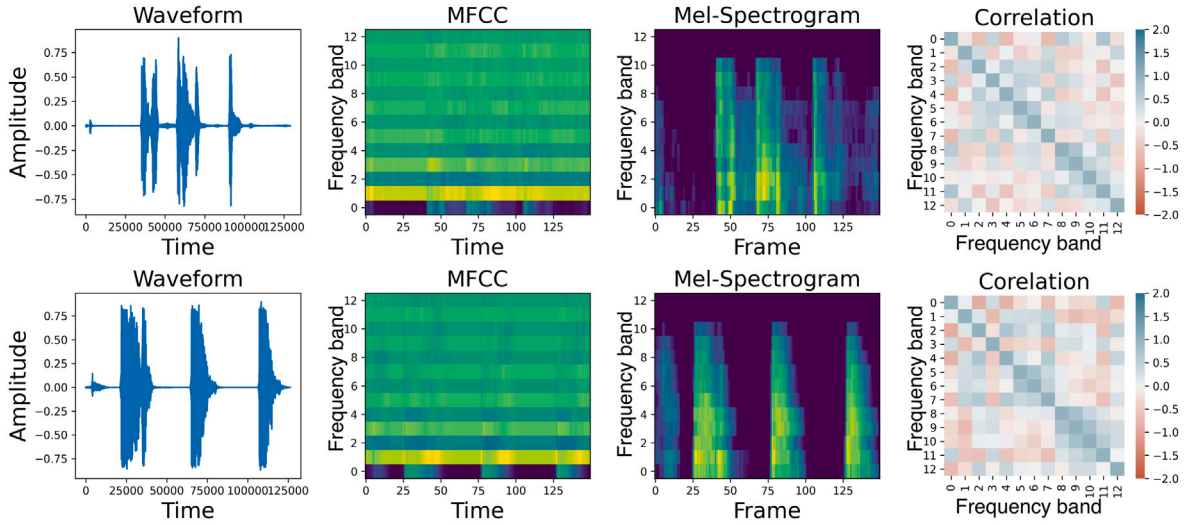

Fig. 1. Visualizations of human coughs in Waveform, MFCC, Mel-Spectrogram and Correlation representations. The first row shows the features of a *positive* case, whereas the second row is that of a *negative* case for COVID-19.

feature, the green color pattern is closer for the positive case than for the negative case. The negative case has a simpler pattern than the positive case, and the correlation of cepstral coefficient 3 with 4, 5, 6, and 7 is higher for the negative case than for the positive case. MFCC or Mel-Spectrogram representations capture the local patterns of sound in time or frequency domains, but they are not capable of capturing the long-range correlations between different parts of the spectrum. For example, the 3rd cepstral coefficient may be strongly related to the 4, 5, 6, and 7th coefficients, but not to the 0 or 11th coefficients. In spectrograms, CNNs are capable of learning these local patterns well by using convolution filters, but not the long-range dependencies. The local patterns of sound contain noises and reduce the quality of features. Thus, the use of a correlation-based presentation can be useful to overcome this issue by employing graph structures that connect distant nodes in the spectrum. Based on these analyses, we propose a novel enhanced graph representation of audio signals. To learn about this type of representation, we investigate graph neural networks (GNNs) to capture the relationships between spectral features as a continuous adjacency matrix. We present Wave2Graph, a novel method capable of learning potent signals from graph-structured representations of sound waves.

To the best of our knowledge, our proposed approach is one of the foundational works that enable the conversion from waveform to a novel graph representation based on the sound frequency spectrum. The graph representation is straightforward and yet extremely helpful as input to a variety of network architectures, especially GNNs. More importantly, we also integrate all temporal, spatial and graph representations into an end-to-end framework. This framework enables neural computing models to process diverse sound signals, where the correlation of frequency coefficients enriches the original signal and enhances the performance of earlier and state-of-the-art approaches.

Our key contributions are summarized as follows:

- We propose a method for converting sound signals to graph representation that combines frequency spectrum and correlation. This novel graph representation allows for the use of graph neural networks to strengthen the classification method for sound classification and overcome the limitations of state-of-the-art approaches.
- We introduce an end-to-end framework named Wave2Graph, to handle and demonstrate the integration of temporal, spatial and graph representations for automatic detection tasks across a range of healthcare contexts and diseases, including respiratory illnesses, Dementia, and COVID-19.
- We carry out extensive experiments to prove Wave2Graph enhances the performance on the *sound tasks* - i.e., respiratory illnesses classification, Dementia detection, and COVID-19 detection. Particularly, our method achieves better performance than approaches claimed SoTA on ADDReSSo and ICBHI dataset, with a large gap in benchmarking metrics, up to ▲7.65% of improvements.
- We also show Wave2Graph works well when combined with other modalities. Even in a multi-modal setting, where signals from multiple domains are leveraged with high accuracy, Wave2Graph still provides a beneficial impact on the final model's performance across all metrics. This shows that our method can provide useful signals for fusion tasks or multi-modal learning.

Our source codes and documentation are made publicly available at https://github.com/ReML-AI/Wave2Graph for future research and benchmarking purposes.

2. Related work

In this Section, we survey existing methods related to our works and summarize their strengths and weaknesses.

2.1. Representation learning for voice pathology detection

Representation learning in sound refers to the process of learning a compact and meaningful representation of sound signals that can be used for various tasks. The task of automatic screening and diagnosing various respiratory diseases has seen a large amount of machine learning and deep learning-centric approaches. Jin et al. (2014) proposed feature extraction methods based on instantaneous kurtosis, discriminating function, and histogram distortion of sample entropy applied to respiratory sound classification. Yadav et al. (2021) built a cardiovascular disease diagnosis system from heartbeat sound by initial spectrograms in the form of images extracted from phonocardiogram sound and feeding into Regularized Convolutional Neural Network. Pham et al. (2022) and Moummad and Farrugia (2023) applied on both Mel-scaled spectrograms with data augmentation methods before feeding into back-end deep learning models for classifying respiratory cycles into certain categories. For COVID-19 cough classification, the dominant deep learning architecture used is Convolutional Neural Network. Transferring the knowledge from Computer Vision,

Table 1
Comparison among existing methods and our proposed method.

Literature	Features	Augmentation	Spectral correlation	Classifier
Jin et al. (2014)	Kurtosis, entropy	✗	✗	SVM
Yadav et al. (2021)	Mel-spectrogram	✗	✗	CNN
Pham et al. (2022)	Mel-spectrogram	✓	✗	CNN
Fakhry et al. (2021)	Mel-spectrogram, Metadata	✓	✗	CNN
Pappagari et al. (2021)	ASR embeddings, BERT	✗	✗	Logistic regression
Perna and Tagarelli (2019)	MFCC	✗	✗	RNN
Ying et al. (2022)	Wav2vec 2.0 + OpenSmile	✗	✗	SVM
Moummad and Farrugia (2023)	Mel-spectrogram	✓	✗	CNN
Lang et al. (2020)	Statistical feature	✗	✗	Graph-based OCSVM
Shirian et al. (2022)	Low level descriptors, MFCCs	✗	✗	GNN
Wave2Graph (OURS)	MFCC, Mel-spectrogram, Correlation	✗	✓	CNN, GNN

typical CNNs such as ResNet are applied directly to 2D matrix representation of original cough signals obtained through pre-processing techniques, such as Mel Frequency Cepstral Coefficients (MFCCs) or Mel-scaled spectrogram. Related metadata (e.g., other health-related issues, sex, age, or hand-crafted features) are also used in parallel with the cough signals in Fakhry et al. (2021). To detect Alzheimer's disease, Pappagari et al. (2021) and Ying et al. (2022) combined various acoustic and linguistic models. The authors suggested that acoustic models can provide better results than linguistic models in certain circumstances due to errors in automatic speech recognition transcription.

Concurrently, instead of 2D matrix representation, cough signal can be viewed as a type of temporal sequence data, which Recurrent Neural Network and its variants Long Short Term Memory networks, Gated Recurrent Units, are designed for modeling the recurrence relationships. With this intuition, Perna and Tagarelli (2019) leverage LSTM-based methods to support the clinician in detecting respiratory diseases, at either level of abnormal sounds or pathology classes.

Table 1 summarizes the main differences between previous works in comparison to ours. In detail, our work differs from the above approaches in the way we handle input data. To be more specific, instead of transforming sound signals to other types of features used in architectures that are well-known for other data domains, we hypothesized that important information in sound signals, the correlation information, is not yet explored. Therefore, we propose to leverage this information for training to detect respiratory diseases. We perform experiments on multiple different architectures and propose Wave2Graph which is a graph neural network-based architecture specifically built for sound signals. We also proposed a method to convert all waveform data to graph data incorporating correlation information.

2.2. Graph learning

Graph learning, in particular graph neural networks (Scarselli et al., 2009; Zhou et al., 2020), is a general term for machine learning on graphs. It aims to extract the desired features from graphs, including node features and edge features, or adjacency matrix between nodes. The features of graphs are projected to feature vectors in continuous space using graph learning methods. As a result, the graph representations can be easily used by downstream tasks. An example of graph neural network-based architecture is illustrated in Fig. 2.

Graph learning methods have gradually increased in popularity (Wu et al., 2020). An advantage of graph learning is that it can capture complex and essential relationships between vertices. For example, in biology, by inferring protein interactions with each other, new treatments for complex diseases can be found (Yang et al., 2020). In sound, Lang et al. (2020) constructed a directed weight spectral graph with k-nearest-neighbors (Cover and Hart, 1967). The adjacent and distributive information of the lung sound samples: labeled normal samples and unlabeled samples. The properties in frequency domain of the lung sounds are applied as the features. The weight of a directed edge measures the similarity between its two vertices using Gaussian kernel function. Shirian et al. (2022) applied a graph neural network-based architecture on multiple audio tracks, considering each audio sample as a graph node to learn inter-dependent audio representations. Other applications of graph learning are widely studied, including image recognition (Nguyen et al., 2021), information systems (Vlachos et al., 2019), and communication networks (Suarez-Varela et al., 2022).

To date, this study lays the first brick to propose a method for transforming all different types of spectral sound signals to graph data for learning and applying the learned representation to multiple downstream tasks and datasets. Here, in this work, correlation information, viewed as continuous adjacency matrix, is used to understand how frequency bandas, viewed as node embeddings, are changing in the sound signal. There is a number of well-known correlation calculation techniques, such as Pearson Correlation (Pearson, 1895) and Coherence Correlation (Bendat and Piersol, 1971), to construct graphs for learning.

3. Methodology

In this Section, we investigate the spectral features and correlations in frequency bands which are obtained by applying the Fourier transform function to the raw audio waveform. Whereas existing approaches (Sharma et al., 2022; Mahanta et al., 2021) typically utilize spectrum as visual representations to a machine learning model, our approach further exploits the connections between sound vibrations forming novel correlation characteristics for representation learning. By analyzing the correlations between frequency bands that vibrate together, it is promising to reflect the hidden and interconnected structures along the sound pathway. We discover that the retrieved correlation features are a natural fit for constructing a graph representation of the data. This representation is then useful with the employment of graph neural networks (GNNs) for learning high-level features in an end-to-end manner, accompanied by elements from the spectrogram. Our proposed method, Wave2Graph, aims to transform any type of sound signals, in their raw waveform, into graph-structured representations, facilitating effective feature learning through graph neural networks.

The first part of this section describes how to convert and visualize audio waveform into a graph. Then, the latter part delves into details on the architecture of Wave2Graph, our end-to-end framework for exploiting spectral features and correlations in sound signals.

3.1. Correlation analysis of the spectrum of sound frequencies

In our method, the original waveform $\mathcal{W}^{(D,1)}$ with D dimension, is transformed into a time and frequency domain representation to decompose the raw signal into individual frequencies and the frequency's amplitude. More specifically, we utilize Mel Frequency Cepstral Coefficients (MFCC) as a nonlinear spectrum of Mel scale of frequencies. The shape of spectrum is (F, T), where F denotes the number of frequency bands, which can be set to 13, 26, 39, 64, 96, 128 or 256, and T denotes the number of time frames. Next, to calculate the pair-wise

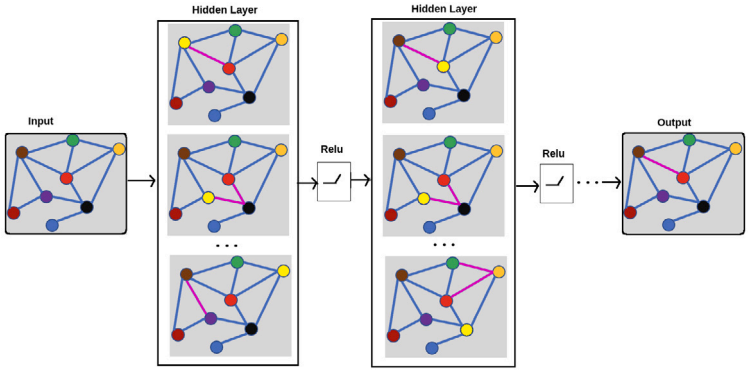

Fig. 2. Graph neural network architecture using message passing mechanism to encode network's structures and patterns (Kipf and Welling, 2017).

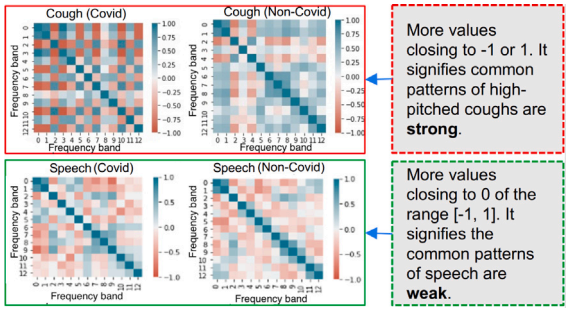

Fig. 3. Comparing spectral correlations of *Cough*, and *Speech* sounds of four random samples in IJSound dataset.

correlation values between frequency bands of spectrum, we consider either Pearson's correlation equation,

$$\mathcal{R}_{ff'} = \frac{\sum_{t=1}^{T}(\mathcal{M}_{ft} - \bar{\mathcal{M}}_f)(\mathcal{M}_{f't} - \bar{\mathcal{M}}_{f'})}{\sqrt{\sum_{t=1}^{T}(\mathcal{M}_{ft} - \bar{\mathcal{M}}_f)^2}\sqrt{\sum_{t=1}^{T}(\mathcal{M}_{f't} - \bar{\mathcal{M}}_{f'})^2}} \tag{1}$$

where:

- f, f' represent the two difference frequencies,
- $\bar{\mathcal{M}}_f = \frac{1}{T}\sum_{t=1}^{T}\mathcal{M}_{ft}$ is mean of the values of the f-variable and analogously for $\bar{\mathcal{M}}_{f'}$.
- $\mathcal{R}_{ff'}$ is correlation coefficient between two different frequencies f, f'

or the Coherence equation,

$$\mathcal{R}_{ff'} = \frac{|Q_{ff'}|^2}{Q_{ff}Q_{f'f'}} \tag{2}$$

where:

- $Q_{ff'} = \sum_{t=1}^{T}(FFT^*(\mathcal{M}_{ft} - \bar{\mathcal{M}}_f) * FFT(\mathcal{M}_{f't} - \bar{\mathcal{M}}_{f'}))$ is the Cross-spectral density between frequency f and frequency f'. FFT is function computes the one-dimensional discrete Fourier Transform (Cooley and Tukey, 1965), FFT^* is complex conjugate of FFT
- $Q_{ff} = \sum_{t=1}^{T}(FFT^*(\mathcal{M}_{ft} - \bar{\mathcal{M}}_f) * FFT(\mathcal{M}_{ft} - \bar{\mathcal{M}}_f))$ is the spectral density of f and f' respectively,
- $|Q|$ is the magnitude of the spectral density.

The output will be normalized to the appropriate range of $(0, 1)$. Illustrations of computed correlation matrices for different types of human signals can be found in Fig. 3. Notably, the visualization highlights the presence of stronger correlation signals in cough data compared to speech signals. This is due to the distinct characteristics of cough, where cough signals are typically short and concise, exhibiting a concentrated burst of energy within a narrow frequency range. In contrast, speech signals encompass a broader frequency distribution due to the inherent variability in vocal production. As a result, the correlation signals observed in speech data tend to be less prominent and display lower relationship between frequency bands compared to cough signals. It is also interesting to note that any other methods of calculating correlation can be used. Our choice for experimenting with the above methods is that they are the most widely used and dominant methods for correlation calculation.

The next step would be to map the correlation features into a graph representation, specifically, a weighted un-directed graph, where each node is a specific frequency band, and the weight of the edge connecting every two nodes is their corresponding frequency band's correlation value. This graph is presented by an *adjacency matrix*, with the values of every cell in the main diagonal line being 1. This matrix, along with the frequency bands, will be used as the Edge feature matrix and Node feature matrix for the input data in GNN.

3.2. Wave2Graph

Fig. 4 depicts our proposed end-to-end framework for learning high-level features from spectral coefficients and correlations and then applying them to classification tasks. It consists of three components: (1) Our proposed approach for learning high-level features from correlation analysis with Graph Neural Network; (2) *Conventional Feature Learning (Backbone)* as in other previous approaches for spectrogram modeling $\mathcal{M}_c$, viewing the representation as image (CNN approaches), temporal sequence (Recurrent Neural Network), or sub-images (Transformer approaches); (3) *Classifier* for performing classification task, based on fusion of output features of both branches.

To learn the conventional representations from input spectrogram, any type of neural network architectures can be used for the *Backbone* branch. An example would be the Resnet-50-based model with the same hyper-parameters as in Table 1 of Fakhry et al. (2021). The input has a shape of (F_c, T_c), with F_c is the number of frequency bands to retain, and T_c is the number of time frames extracted from the audio. The learned representation output $\mathcal{E}^{\mathbf{c}}$ is calculated by:

$$\mathcal{E}^{\mathbf{c}} = \mathbf{c}(\mathcal{M}_c) \tag{3}$$

with $\mathbf{c}$ denotes the *Backbone* neural network and is differentiable. The shape of $\mathcal{E}^{\mathbf{c}}$ is set to be $(H^{\mathbf{c}}, 1)$.

In contrast to existing approaches that primarily utilize the spectrum as visual representations for machine learning models, our proposed approach explores the connections between sound vibrations to discover novel correlation characteristics for representation learning. By analyzing the correlations between spectral bands that vibrate together, it becomes possible to capture the hidden and interconnected structures along the sound pathway. We propose adopting a GNN architecture to learn and leverage high-level characteristics from the

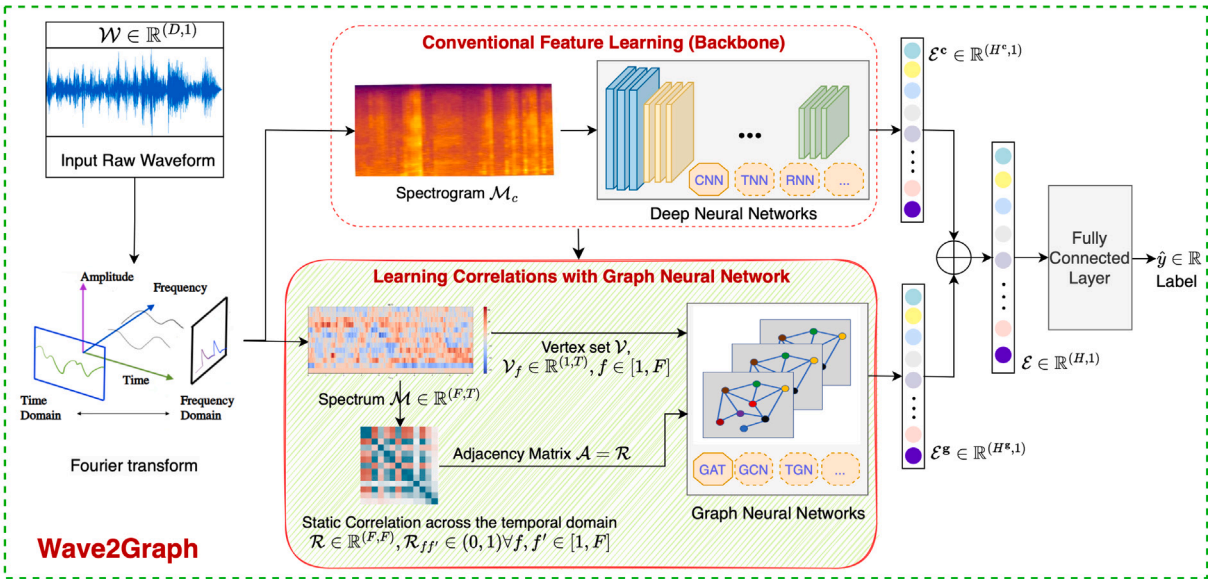

Fig. 4. Summary of Wave2Graph framework. Fourier transform filters a complex signal (time) by decomposing it into a simpler representation (frequency). The first (upper) branch leverages a common approach in sound classification, typically using a CNN or RNN-based model. The second (lower) branch focuses on learning correlation. The final output of the two branches are fused before being passed to a classification layer.

underlying structure of the input matrices to utilize the newly computed correlation features. Here, we leverage graph attention networks (GATs) (Veličković et al., 2018) as the minimum benchmark for learning spectral and correlation features in sound data, ensuring greater extensibility and potential improvements for future work.

We defined the graph presentation $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ denotes the vertex set, and $\mathcal{A}$ is the adjacency matrix. To obtain $\mathcal{V}$, we defined each frequency band of spectrum $\mathcal{M}$ as a node of the correlation graph. This means the graph has a total of F nodes, and each node is a vector of shape $(1, T)$. The adjacency matrix $\mathcal{A}$ defines the weights for the edges between every node in the graph. Concurrently, the correlation matrix $\mathcal{R}$ also defines the correlation of the frequency bands across the temporal domain. This makes the correlation features a natural fit for the adjacency matrix $\mathcal{A}$. $\mathcal{A}$ is a square matrix, $\mathcal{A} \in \mathbb{R}^{F \times F}$, and:

$$\mathcal{A}_{ff'} = \mathcal{R}_{ff'} \quad \text{for } f, f' \in [1, F] \tag{4}$$

Denotes the function of GNN as $\mathbf{g}$, then the output $\mathcal{E}^{\mathbf{g}}$ of the second branch is:

$$\mathcal{E}^{\mathbf{g}} = \mathbf{g}(\mathcal{V}, \mathcal{A}) \tag{5}$$

The final shape of the output high-level features is $(H^{\mathbf{g}}, 1)$. In our experiments, we control the output embedding dimensions for both branches to be equal. This allows for fair analysis of the contributions of each branch, either conventional learning or graph learning algorithms to the downstream task.

To obtain the final representation $\mathcal{E}$, $\mathcal{E}^{\mathbf{c}}$ and $\mathcal{E}^{\mathbf{g}}$ are fused together. In this groundwork, we use the concatenation function $\oplus$, but more advanced fusion techniques can also be further explored.

$$\mathcal{E} = \mathcal{E}^{\mathbf{c}} \oplus \mathcal{E}^{\mathbf{g}} \tag{6}$$

As a result, the shape of $\mathcal{E}$ is $(H^{\mathbf{c}} + H^{\mathbf{g}}, 1) = (H, 1)$. Finally, the extracted high-level features are passed to a fully connected layer for the downstream task. For instance, in Alzheimer's Dementia classification task, the two labels are Alzheimer's disease and healthy control. The final output is $\hat{y}$.

The framework is trained in an end-to-end manner through a cross-entropy loss function:

$$\mathcal{L}oss = y \log(\sigma(\hat{y})) + (1 - y) \log(1 - \sigma(\hat{y})) \tag{7}$$

where y is the true label, $y = 1$ means the current sample is labeled positive with the disease, and $y = 0$ means the sample is negatived with the disease. $\sigma(\cdot)$ is known as the activation function.

4. Experiments

This section describes a comprehensive list of experiments and datasets employed in this study. Experimental results and analysis are explained and discussed in this section. We intend to provide a deeper understanding of the research problem and to reach relevant conclusions based on our analysis by carefully examining these experiments and datasets. Standard deviation is reported in all of our experiments after ± symbols in all tables. For related work, we note that most of previous SoTA approaches (Pham et al., 2022; Pappagari et al., 2021; Nguyen and Pernkopf, 2022; Li et al., 2021) only report the mean values, without taking into consideration the standard deviations across training runs, we include these metrics wherever available.

4.1. Datasets

Datasets investigate in this work are listed as follows:

• *ICBHI (Rocha et al., 2019)* is a public respiratory sound database associated with the International Conference on Biomedical Health Informatics — ICBHI 2017. The database includes 920 annotated audio samples from 126 patients, totaling 5.5 h of recordings with 6898 breathing cycles, of which 1864 have crackles, 886 have wheezes, and 506 manifest both crackles and wheezes. The dataset is composed of a diverse collection of patients, including patients with chronic diseases (e.g., COPD, Bronchiectasis, and Asthma) or non-chronic diseases (e.g., Upper and Lower respiratory tract infection, Pneumonia, and Bronchiolitis), as well as healthy control subjects. We perform four-class (*normal, crackle, wheeze, both*) breathing cycle classification task with the official 60%–40% data split for train–test. The main benchmarking metric ICBHI score is computed as the mean of Sensitivity $S_e = \frac{P_c + P_w + P_b}{N_c + N_w + N_b}$ and Specificity $S_p = \frac{P_n}{N_n}$, where P and N denote the amount of correctly classified and total number of samples for each class, respectively.

• *ADReSSo (Luz et al., 2021)* is an English dataset from the INTERSPEECH 2021 Alzheimer's Dementia Recognition through Spontaneous Speech Challenge. 165 audio samples from 87 patients and 79 cognitively normal persons are included in the ADReSSo dataset. All of the participants in the Boston Diagnostic Aphasia Assessment were asked to explain the Cookie Theft image, and the dataset is made up of recordings of their descriptions. All audios are acoustically enhanced with stationary noise removed and level balanced across all speech segments in order to reduce variance brought on by recording conditions, such as microphone positioning. The audio's duration ranges from 22 to 235 s. Given the complex multi-modality characteristic of this dataset, we employ SoTA approaches such as leveraging pretrained models and toolkit, including OpenSmile (Eyben et al., 2010) and Wav2vec 2.0 (Baevski et al., 2020) for extracting speech acoustic signals, and fine-tuned BERT (Kenton and Toutanova, 2019) for extracting deep linguistic features. Moreover, we incorporate Wave2Graph to integrate powerful signals from frequency-based correlations and graph-structured audio input in combination with other multi-modality features. We follow the pipeline of 10-fold cross-validation (Ying et al., 2022) for fair benchmarking on this dataset, with accuracy as the primary evaluation metric. We also report other important metrics, including precision, recall, and F1-score.

• *IJSound* is a private real-world dataset for respiratory disease detection from sounds. The data is collected through a web-based and mobile application from August 2021 to the end of October 2021. IJSound consists of 2620 samples, including cough (denotes $IJSound_{cough}$) and breathing (denotes $IJSound_{breathe}$) sound signals. The participants were asked to provide their cough signals and the corresponding illnesses (e.g., Related-Flu, Pneumonia, and COVID-19) along with medical-related habits (e.g., smoking, drinking). We conduct experiments on detecting COVID-19 diseases from different type of sound signals. Data is split into 5 folds with stratify for cross-validation, and AUC score is our main focused metric as this dataset is highly imbalanced. Our dataset is available upon request.

4.2. Experimental details

In this research, Pytorch was our framework of choice for implementing models. All experiments are carried out on an Ubuntu Server (20.04 LTS) with a single RTX 3090 GPU. This work acts as a foundational work in the application of graph neural networks to the domain of sound data. To this end, we propose Wave2Graph, which incorporates a basic variant of Graph Attention (GAT) Layer (Veličković et al., 2018). The base version of Wave2Graph comprises GAT layers with carefully selected embedding dimensions. For ICBHI, we employed four layers with embedding dimensions of 512, 512, 512, and 256. For our experiments on ADReSSo, we utilized two GAT layers with embedding dimensions of 2048 and 512. Similarly for $IJSound_{cough}$, we employed three layers with embedding dimensions of 1024, 1024, and 32. Finally, for $IJSound_{breathe}$, we used two layers with dimensions of 1024 and 256. These settings are found through architecture search, represent the minimum requirements for achieving optimal performance on each dataset, ensuring that Wave2Graph extracts meaningful and informative features from the input data. We train the model for 100 or 200 epochs, depending on the dataset used and the model's size. Nevertheless, the model tends to converge before completing the training process. All models are optimized with AdamW optimizer (Loshchilov and Hutter, 2017), a learning rate of $1e{-}4$, and a batch size of 64. We use the original sampling rate of 4000 Hz on ICBHI and 16000 Hz on ADReSSo and IJSound. To normalize the original audio signal y, we use a scaling factor of 0.9 and divided the signal by its maximum value $max(y)$: $y = 0.9 \times \frac{y}{\|max(y)\|}$. This normalization process helps ensure that the audio signal is consistent and suitable for further analysis and processing. The trim of silence in start and end of audio was implemented with Librosa library[2] in all cases. To extract mel-spectrogram representation for learning on backbone branch, window length and hop size are set to 256 and 12 on ICBHI or 1024 and 48 on IJSound, with respect to their original sampling rate. This pre-processing step is not necessary for experiments on ADReSSo, as we leveraged pre-trained embeddings on audio provided by Ying et al. (2022).

In this study, the pre-processing step of extracting audio signals and representing them as graph structures is essential for employing graph learning algorithms to capture powerful signal from the frequency-based correlations. To achieve this, we employ MFCCs as the temporal-spatial representation of input audio waveforms. The number of cepstral coefficients M was set to 39 to capture the relevant and robust spectral characteristics of sound waves. To ensure the MFCC representation is effective, we optimize the hop length and window size hyper-parameters based on the specific audio length of each dataset. For ICBHI dataset, the hop length and window size are set to 20 and 256, respectively. Similarly on ADReSSo and IJSound datasets, the hop length and window size are set to 200 and 1024, respectively. These specific hyper-parameters are chosen so that after the pad and crop operations, the embedding lengths of input nodes to graph learning algorithms are approximately 800. We also conduct ablation studies on this pre-processing step to verify the hypothesis.

4.3. Results

Overall performances.

Wave2Graph framework consistently performs better across all datasets, including ICBHI, ADReSSo, and IJSound, in comparison to SoTA methods and baselines. This shows that Wave2Graph can reliably work on different types of sound data for classification tasks with extensibility.

Results on ICBHI. Results on ICBHI dataset are demonstrated in Table 2 where our method, Wave2Graph, incorporating graph learning algorithm with correlation matrix between frequency bands, achieves SoTA result for solutions without additional training data by an improvement of 3.89 in ICBHI score (▲7.65%). It also obtains comparable performance compared to solutions using extra training data, without additional training data and fewer parameters. It is worth noting that we only employ a simple transformer-based model for the backbone branch (Fig. 4) without any specific modifications seen in previous SoTA methods. This also indicates the effectiveness of our Wave2Graph framework.

Results on ADReSSo. With the multi-modality characteristic of ADReSSo dataset, it may pose as challenging for graph learning methods on spectrum representation of sound data to encode informative signals when being fused with other modalities. Nevertheless, we achieve a substantial improvement of up to 3.00 improvement in Accuracy score (▲3.58%) compared to SoTA method (Ying et al., 2022) when Wave2Graph is plugged in and trained in an end-to-end manner. This is illustrated in Table 3 (Exp#4 compared to Exp#3), proving that even in a multi-modal environment, when signals from multiple domains are leveraged with high accuracy, Wave2Graph is still able to have a beneficial impact on final model's performances across all metrics. We note that the amount of available recordings on ADReSSo is limited means that precise predictions on each sample can significantly influence the performance metrics, leading to the observed relatively high standard deviation across folds. However, Wave2Graph still successfully enhances the performance of previous method and achieve a new SoTA result on the dataset. This outcome underscores the robustness and effectiveness of our proposed neural architecture.

Results on IJSound. To further understand the strength of our proposed method, we perform experiments on our private real-world

[2]

Table 2
Performances of proposed methods on ICBHI dataset (official 60-40 split). ↑ means higher number is better.

Exp	Method		Additional	With	% ICBHI↑	% Sens.↑	% Spec.↑
	Previous SoTA	Wave2Graph (Ours)	training data	correlation	Mean ± std	Mean ± std	Mean ± std
#1	Pham et al. (2022)	✗	✓	✗	57.3	30.0	**85.6**
#2	Moummad and Farrugia (2023)	✗	✓	✗	57.55	*39.15*	75.95
#3	Nguyen and Pernkopf (2022)	✗	✓	✗	*58.29*	37.24	*79.34*
#4	Li et al. (2021)	✗	✗	✗	53.90	36.36	71.44
#5	Moummad and Farrugia (2023)	✗	✗	✗	54.74	33.84	75.35
#6	Transformer-based	✓	✗	✓	**58.93 ± 0.48**	**40.48 ± 0.91**	77.34 ± 0.92

Table 3
Performances of proposed methods on ADReSSo dataset with 10-fold cross validation. ↑ means a higher number is better.

Exp	Method		With	% Accuracy↑	% Precision↑	% Recall↑	% F1-score↑
	Previous SoTA	Wave2Graph (Ours)	correlation	Mean ± std	Mean ± std	Mean ± std	Mean ± std
#1	Wang et al. (2021)	✗	✗	77.2	78.7	74.0	76.3
#2	Pappagari et al. (2021)	✗	✗	81.3	–	–	–
#3	Ying et al. (2022)	✗	✗	83.7 ± 8.6	85.3 ± 8.6	84.0 ± 8.4	83.6 ± 8.7
#4	Ying et al. (2022)	✓	✓	**86.7 ± 7.1**	**87.6 ± 6.7**	**86.9 ± 6.9**	**86.6 ± 7.1**

Table 4
Performance comparison on IJSound dataset employing various backbone and SoTA network architectures from other benchmark datasets. Here, we explore the efficacy of Wave2Graph when integrated with different architectures for the backbone branch. Specifically, we leverage the widely-adopted CNN and transformer architectures, as well as SoTA methods (Ying et al., 2022; Moummad and Farrugia, 2023) on ICBHI and ADReSSo datasets. ↑ means a higher number is better. Standard deviations are provided after ± symbols.

Exp	Method		With	% AUC on	% AUC on
	Previous SoTA	Wave2Graph (Ours)	Correlation	IJSound$_{cough}$ ↑	IJSound$_{breathe}$ ↑
#1	CNN-based	✗	✗	71.83 ± 0.93	74.27 ± 0.93
#2	CNN-based	✓	✓	**73.90 ± 1.01**	**76.27 ± 0.28**
#3	Ying et al. (2022)	✗	✗	65.97 ± 1.37	66.25 ± 2.93
#4	Ying et al. (2022)	✓	✓	**66.18 ± 0.81**	**67.00 ± 3.29**
#5	Moummad and Farrugia (2023)	✗	✗	74.69 ± 1.84	77.26 ± 2.86
#6	Moummad and Farrugia (2023)	✓	✓	**75.35 ± 1.89**	**77.87 ± 3.03**
#7	Transformer-based	✗	✗	87.13 ± 0.66	83.63 ± 0.61
#8	Transformer-based	✓	✓	**88.22 ± 1.10**	**85.50 ± 1.11**

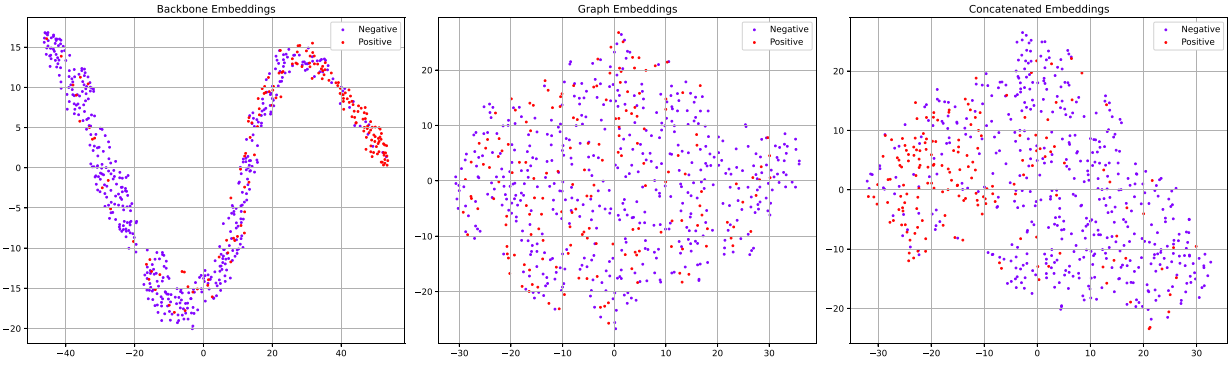

Fig. 5. t-SNE of embeddings learned through Wave2Graph architecture (right) and previous method (left) on IJSound$_{breathe}$.

dataset, IJSound. As baselines for the dataset, we leverage the widely-adopted CNN and Transformer architectures, as well as SoTA methods (Ying et al., 2022; Moummad and Farrugia, 2023) on ICBHI and ADReSSo datasets. Subsequently, we conduct experiments using the same backbone networks and plugging in Wave2Graph and analyze the performances of the enhanced system compared to without Wave2Graph. Results in Table 4 consistently demonstrate the efficacy of our method, across all backbones and sound types. Notably, Wave2Graph helps achieve the highest AUC scores of 88.22 and 85.50 (▲1.25% and ▲2.24%, respectively) for IJSound$_{cough}$ and IJSound$_{breathe}$ when integrated with Transformer-based networks, as can be seen in Exp#7 and Exp#8. These findings align with recent studies that have highlighted the excellent performances of Transformer-based models in multiple domains.

Embeddings visualization. To visually analyze and better understand the learned embeddings of each branch and their concatenation, we employed t-distributed stochastic neighbor embedding (t-SNE) to obtain a 2-dimensional visualization, as shown in Fig. 6 for IJSound$_{cough}$ and Fig. 5 for IJSound$_{breathe}$. In both cases, for the backbone embeddings, we observe that while there is some level of clustering for the two classes, there is also significant overlap with high-density regions. This hinders at overfitting behavior for the backbone branch. On the other hand, the embeddings learned by Wave2Graph exhibit a different and more generalized trend, suggesting that it captures distinct information compared to the backbone. The distances between points from different classes are bigger, allowing a boundary line to be drawn between points of each class easier. Finally, the concatenated embeddings combined the strengths of both branches, with an adequate clustering effects and

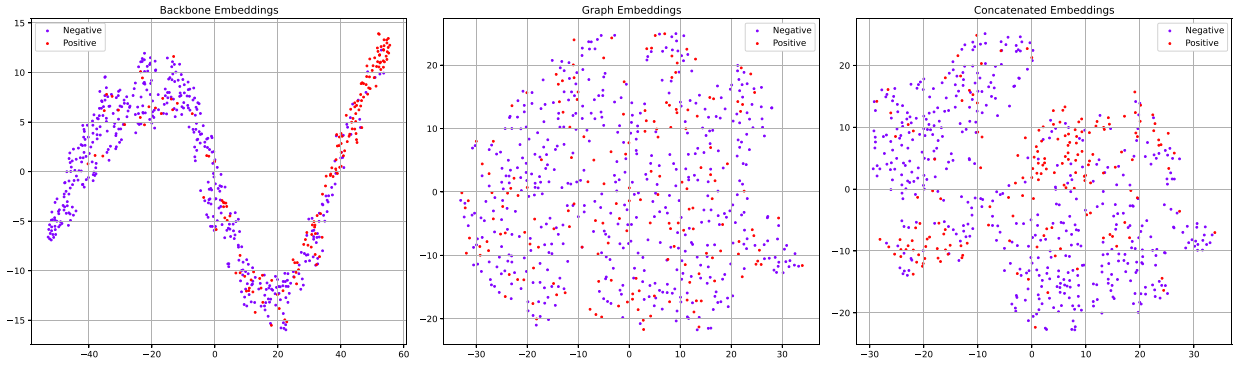

Fig. 6. t-SNE of embeddings learned through Wave2Graph architecture (right) and previous method (left) on IJSound$_{cough}$.

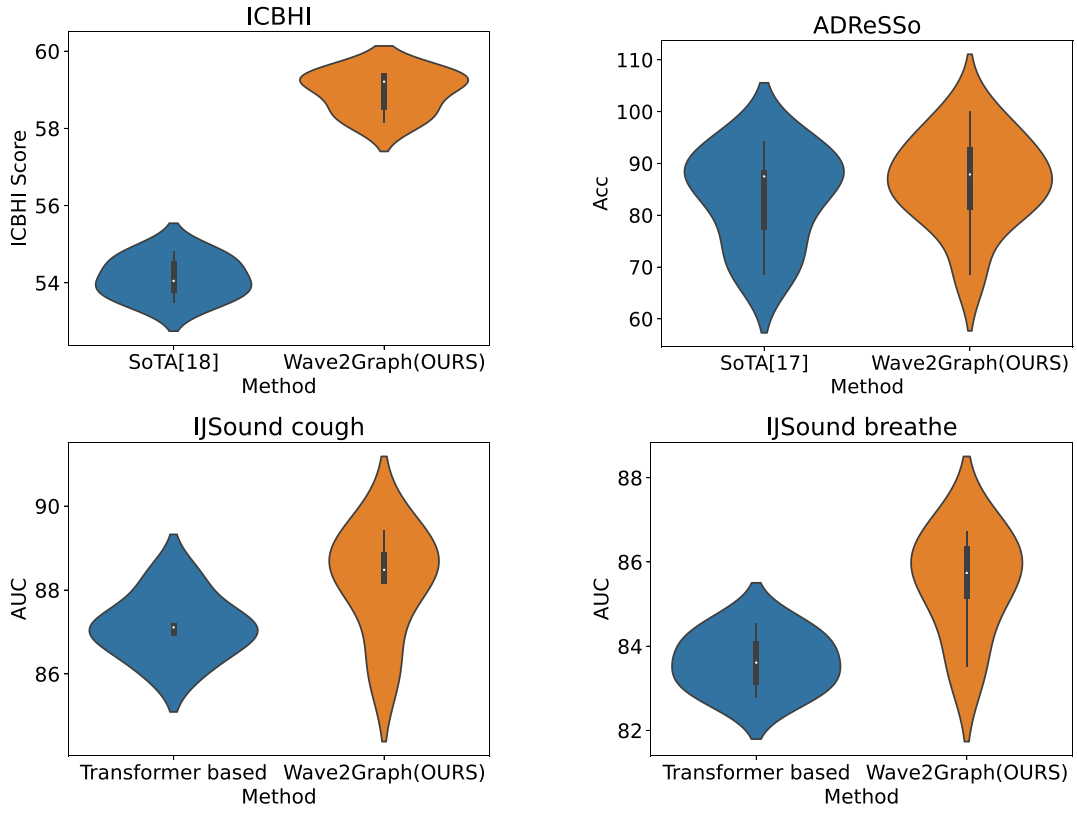

Fig. 7. Performance analysis across multiple seeds/folds on different datasets. From left to right, top to bottom: ICBHI, ADReSSo, IJSound$_{cough}$, and IJSound$_{breathe}$.

a looser density indicating that the concatenated embeddings capture a more comprehensive representation of the data. This combination of embeddings may provide a more robust feature space for classification tasks.

Performance analysis. This performance analysis study is designed to verify the impact of random seeds on the final performance of our Wave2Graph framework in comparison to the previous SoTA and baseline methods. Fig. 7 evidently shows that Wave2Graph framework reliably outperforms previous approaches up to the following average improvements: ▲7.65% ($t[4] = 13.95$, $p = 0.00015$) on ICBHI, ▲3.58% ($t[9] = 1.32$, $p = 0.2209$) on ADReSSo, ▲1.25% ($t[4] = 3.56$, $p = 0.0235$) on IJSound$_{cough}$ with Transformer, and ▲2.24% ($t[4] = 3.84$, $p = 0.0184$) on IJSound$_{breath}$ with Transformer. Overall, with the use of Wave2Graph, we observe a consistent and statistically significant improvement of ▲3.76% ($t[24] = 2.72$, $p = 0.0119$) in average across all datasets. Furthermore, the stability of our approach is also evident by the standard deviations between training runs, we include these metrics wherever available. As shown in experiments across datasets and baselines in Table 2, 3, and 4, our approach not only outperforms the baselines but also exhibits comparable or even higher stability. For instance, on the ADReSSo dataset, our method achieves a standard deviation of 7.1, compared to 8.6 of the SoTA method (Ying et al., 2022).

Efficiency analysis. While incorporating Wave2Graph into the backbone introduces additional complexity in terms of the number of parameters, this is partly mitigated by the fact that both branches operate and execute in parallel. When a larger backbone is chosen, the total running time during training and inference should remain comparable to using the backbone alone. This is because the backbone, being the more complex component, dominates the overall computation. On the other hand, if a small backbone is selected, the total time complexity would be governed by Wave2Graph. For instance, in our experiments on IJSound, with the transformer-based backbone, we observed an additional overhead of 16% in training and inference speed. Specifically, the backbone alone achieved a throughput of 424.83 samples/s (with a batch size of 16), compared to 372.48 samples/s of the final solution

Table 5
Ablation study of Wave2Graph under different settings on ICBHI, ADReSSo and IJSound. ↑ means higher number is better.

Exp	Wave2Graph	ICBHI (% ICBHI Score↑)	ADReSSo (% Accuracy↑)	IJSound$_{cough}$ (% AUC↑)	IJSound$_{breathe}$ (% AUC↑)
#1	13 Frequency bands, MFCC, Pearson	58.15 ± 0.69	84.89 ± 8.58	88.38 ± 0.78	86.00 ± 1.02
#2	64 Frequency bands, MFCC, Pearson	58.63 ± 0.35	84.82 ± 10.05	88.39 ± 0.90	85.70 ± 0.55
#3	128 Frequency bands, MFCC, Pearson	58.69 ± 0.37	84.85 ± 9.91	88.41 ± 0.93	85.82 ± 1.08
#4	39 Frequency bands, MFCC, Pearson	**58.93 ± 0.48**	**86.65 ± 7.09**	88.22 ± 1.10	85.49 ± 1.11
#5	39 Frequency bands, Mel-spectrogram, Pearson	58.08 ± 0.54	84.93 ± 9.06	86.34 ± 0.89	**86.19 ± 0.67**
#6	39 Frequency bands, MFCC, Coherence	58.62 ± 0.60	85.48 ± 7.74	88.30 ± 1.09	85.74 ± 0.59

Table 6
Ablation study of Wave2Graph under different Signal-to-Noise Ratio (SNR) where Gaussian noise is injected accordingly. We report the AUC scores (%) on IJSound$_{cough}$ and IJSound$_{breathe}$, where higher value is better.

Exp	Dataset	Transformer-based	Wave2Graph (w/ transformer-based backbone)
	IJSound$_{cough}$		
#1	+5.0 SNR	56.05 ± 0.85	**56.76 ± 1.53**
#2	0.0 SNR	52.87 ± 1.74	**53.28 ± 1.96**
#3	−5.0 SNR	50.15 ± 1.29	**50.42 ± 1.18**
	IJSound$_{breathe}$		
#4	+5.0 SNR	65.71 ± 1.95	**66.99 ± 1.25**
#5	0.0 SNR	60.55 ± 1.57	**62.29 ± 1.09**
#6	−5.0 SNR	55.78 ± 1.48	**58.27 ± 1.96**

with Wave2Graph. It is important to note that the impact on efficiency can vary depending on the specific backbone chosen and the hardware utilized.

Ablation study with different input graph representation. The ablation study presents in Table 5, is designed to explore the role of input graph-structured representation for Wave2Graph and how they affect final performances across multiple benchmarking datasets. The implications of the results are two folds. First, our hypothesis of using MFCCs as the representation of sound signals, and our novel view on this representation as graph structure for input to graph neural networks has been verified. We consider each cepstral coefficient as a node, compute the continuous correlation matrix, and utilize graph learning algorithms. This approach yields significant improvements in AUC score, with a gap of up to 2.07 in AUC score (▲2.40%) on IJSound$_{cough}$. Secondly, our choice of leveraging graph neural network for learning on this new type of sound data representation has proven effective for downstream tasks and is robust with regard to different hyper-parameters in the pre-processing step. While we recommend using 39 cepstral coefficients for MFCC as the default setting, our models are able to learn from the graph structure transformed from various representations. This suggests that our approach is reliable and can be used across different datasets and scenarios.

Ablation study on robustness of Wave2Graph against noise. We further investigate the sensitivity of Wave2Graph framework to noisy data. To this end, we inject Gaussian noise into the corresponding evaluation sets of IJSound$_{cough}$ and IJSound$_{breathe}$ at +5.0, 0.0, or −5.0 Signal-to-Noise ratios (SNR). The models, trained on the clean training sets of IJSound and presented in Table 4, are then evaluated on each noisy evaluation set. The results, reported in Table 6, highlight the robustness of our framework. The transformer-based model achieves higher AUC scores on noisy settings when plugged into the framework, even though it was trained only on clean data. Notably, there is an absolute improvement of up to 2.49 (▲4.46%) in AUC with an SNR of −5.0 on IJSound$_{breathe}$. Moreover, the performance gains are consistent across both datasets and different noise levels, indicate the robustness of Wave2Graph against various noise scenarios.

5. Conclusion

This study contributes Wave2Graph, an end-to-end framework for understanding and learning spectral characteristics and correlations in acoustic waves. Moving beyond conventional sound representations, Wave2Graph draws inspiration from physical insights into how sound travels through objects and structures. Wave2Graph considers spectral correlation signals and establishes an unprecedented graph-structured representation of sound signals. It enables the extraction and utilization of high-level characteristics through graph neural networks, providing consistent and reliable state-of-the-art performances. Our approach facilitates feasible, low-cost training and inference for mass-screening of downstream sound-related tasks, as well as the usage of several backbones for better representation and adaption in various use cases. The efficacy and generalization of the proposed approach are validated by conducting extensive experiments on three real-world benchmark datasets in the medical domain.

With its suitability and compatibility, any existing neural network-based system can take advantage of Wave2Graph for learning and investigating the spectral correlation patterns of sensors and sound waves, gaining significant performance improvements. There are many practical implications for Wave2Graph across various domains with the collection of waveform data, including manufacturing, environmental monitoring, and healthcare applications. We list several scenarios below:

- **Manufacturing**. Industrial machines require synchronous and smooth operation (Truong et al., 2021; Peng et al., 2022), noticeable from their produced day-to-day noises and sounds. Based on the fact that a smooth-running machine makes regular sound frequencies, Wave2Graph can effectively analyze abnormal frequencies in vibrating sounds or sensors, predicting potential breakdowns. Moreover, its applicability extends to other manufacturing-related tasks, such as predictive maintenance, quality control, and anomaly detection.
- **Environmental Monitoring**. We can analyze the different aspects and distinguish the objects that cause sounds by vibrating in the soundscape ecology. For example, researchers use population monitoring as a strategy to understand and study endangered bird populations. Wave2Graph can be employed to encode the correlation feature based on multiple birds' sounds when they sing/call for multi-label bird classification tasks.
- **Healthcare**. By leveraging neural networks, human sounds can provide reliable diagnostic suggestions for various health-related diseases and disorders, including dementia, flu, pneumonia, and bronchitis (Lei et al., 2014; Yang et al., 2022). Our experiments have shown the effectiveness of Wave2Graph in analyzing correlational patterns of sound for medical conditions. It is promising to explore further the effectiveness of Wave2Graph on other wave types of medical signals, including ECG and EEG data.

Although Wave2Graph achieves outstanding performance on datasets of medical sounds, there are several potential directions for future works. Our study only focuses on medical sound, and it is intriguing to further investigate Wave2Graph on new benchmark datasets of sound or sensors in various domains such as industry, education, and environment for diversity. Moreover, more advanced GNN architectures can be explored for leveraging the correlation information to achieve even better performances. Modern fusion methods for traditional features and the newly computed correlations, such as Hazarika et al. (2020), is also worth further investigation. As a foundational work, Wave2Graph is the beginning of graph-based learning on spectral features and correlations for advancements in neural computing, signal processing, and artificial intelligence.

CRediT authorship contribution statement

Van-Truong Hoang: Data curation, Methodology, Visualization, Writing – original draft. **Khanh-Tung Tran:** Methodology, Visualization, Writing – original draft. **Xuan-Son Vu:** Conceptualization, Supervision, Validation, Writing – original draft, Writing – review & editing. **Duy-Khuong Nguyen:** Validation, Writing – review & editing. **Monowar Bhuyan:** Validation, Writing – review & editing. **Hoang D. Nguyen:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available upon request.

Acknowledgments

This research work has emanated from research conducted with financial support from Science Foundation Ireland (SFI) under Grant 12/RC/2289-P2 and 18/CRT/6223.

References

Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., 2017. Investigation of voice pathology detection and classification on Different Frequency Regions using correlation functions. J. Voice (ISSN: 0892-1997) 31 (1), 3–15. http://dx.doi.org/10.1016/j.jvoice.2016.01.014.

Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: International Conference on Neural Information Processing Systems. NIPS '20, 33, Curran Associates Inc., Red Hook, NY, USA, ISBN: 9781713829546, pp. 12449–12460.

Bendat, J.S., Piersol, A.G., 1971. Random Data: Analysis and Measurement Procedures. Wiley-Interscience.

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine calculation of complex Fourier series. Math. Comp. 19, 297–301.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13 (1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Ding, W., Nayak, J., Swapnarekha, H., Abraham, A., Naik, B., Pelusi, D., 2021. Fusion of intelligent learning for COVID-19: A state-of-the-art review and analysis on real medical data. Neurocomputing (ISSN: 0925-2312) 457, 40–66. http://dx.doi.org/10.1016/j.neucom.2021.06.024.

Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – the munich versatile and fast open-source audio feature extractor. In: MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. pp. 1459–1462. http://dx.doi.org/10.1145/1873951.1874246.

Fakhry, A., Jiang, X., Xiao, J., Chaudhari, G., Han, A., 2021. A multi-branch deep learning network for automated detection of COVID-19. In: International Speech Communication Association. INTERSPEECH, pp. 4139–4143. http://dx.doi.org/10.21437/Interspeech.2021-378.

Hazarika, D., Zimmermann, R., Poria, S., 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: ACM International Conference on Multimedia. Association for Computing Machinery, ISBN: 9781450379885, pp. 1122–1131. http://dx.doi.org/10.1145/3394171.3413678.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Holi, M.S., et al., 2013. Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks. In: IEEE Point-of-Care Healthcare Technologies. pp. 76–79.

Imran, A., Posokhova, I., Qureshi, H.N., Masood, U., Riaz, M.S., Ali, K., John, C.N., Hussain, M.I., Nabeel, M., 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform. Med. Unlocked (ISSN: 2352-9148) 20, 100378. http://dx.doi.org/10.1016/j.imu.2020.100378.

Jin, F., Sattar, F., Goh, D., 2014. New approaches for spectro-temporal feature extraction with applications to respiratory sound classification. Neurocomputing (ISSN: 0925-2312) 123, 362–371. http://dx.doi.org/10.1016/j.neucom.2013.07.033.

Kenton, J.D.M.-W.C., Toutanova, L.K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Vol. 1, p. 2.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. Int. Conf. Learn. Represent..

Lang, R., Lu, R., Zhao, C., Qin, H., Liu, G., 2020. Graph-based semi-supervised one class support vector machine for detecting abnormal lung sounds. Appl. Math. Comput. 364, 124487. http://dx.doi.org/10.1016/j.amc.2019.06.001.

Lei, B., Rahman, S.A., Song, I., 2014. Content-based classification of breath sound with enhanced features. Neurocomputing (ISSN: 0925-2312) 141, 139–147. http://dx.doi.org/10.1016/j.neucom.2014.04.002.

Li, J., Yuan, J., Wang, H., Liu, S., Guo, Q., Ma, Y., Li, Y., Zhao, L., Wang, G., 2021. LungAttn: advanced lung sound classification using attention mechanism with dual TQWT and triple STFT spectrogram. Physiol. Meas. 42 (10), 105006.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. Int. Conf. Learn. Represent..

Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2021. Detecting cognitive decline using speech only: The adresso challenge. In: International Speech Communication Association. INTERSPEECH, pp. 3780–3784. http://dx.doi.org/10.21437/Interspeech.2021-1220.

Lv, G., Hu, Z., Bi, Y., Zhang, S., 2021. Learning unknown from correlations: Graph neural network for inter-novel-protein interaction prediction. In: International Joint Conferences on Artificial Intelligence. pp. 3677–3683. http://dx.doi.org/10.24963/ijcai.2021/506.

Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., Schneider, S.L., 2023. Speech emotion recognition using machine learning — A systematic review. Intell. Syst. Appl. (ISSN: 2667-3053) 20, 200266. http://dx.doi.org/10.1016/j.iswa.2023.200266.

Mahanta, S.K., Kaushik, D., Van Truong, H., Jain, S., Guha, K., 2021. COVID-19 diagnosis from cough acoustics using ConvNets and data augmentation. In: International Conference on Advances in Computing and Future Communication Technologies. pp. 33–38. http://dx.doi.org/10.1109/ICACFCT53978.2021.9837350.

Moummad, I., Farrugia, N., 2023. Learning audio features with metadata and contrastive learning. arXiv:2210.16192.

Nguyen, T., Pernkopf, F., 2022. Lung sound classification using co-tuning and stochastic normalization. IEEE Trans. Biomed. Eng. 69 (9), 2872–2882. http://dx.doi.org/10.1109/TBME.2022.3156293.

Nguyen, H.D., Vu, X.-S., Le, D.-T., 2021. Modular graph transformer networks for multi-label image classification. Proc. AAAI Conf. Artif. Intell. 35 (10), 9092–9100.

Pappagari, R., Cho, J., Joshi, S., Moro-Velázquez, L., Żelasko, P., Villalba, J., Dehak, N., 2021. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In: International Speech Communication Association. INTERSPEECH, pp. 3825–3829. http://dx.doi.org/10.21437/Interspeech.2021-1850.

Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. In: Proceedings of the Royal Society of London. pp. 240–242.

Peng, X., Li, H., Yuan, F., Razul, S.G., Chen, Z., Lin, Z., 2022. An extreme learning machine for unsupervised online anomaly detection in multivariate time series. Neurocomputing (ISSN: 0925-2312) 501, 596–608. http://dx.doi.org/10.1016/j.neucom.2022.06.042.

Perna, D., Tagarelli, A., 2019. Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In: IEEE 32nd International Symposium on Computer-Based Medical Systems. CBMS, pp. 50–55. http://dx.doi.org/10.1109/CBMS.2019.00020.

Pham, L., Ngo, D., Tran, K., Hoang, T., Schindler, A., McLoughlin, I., 2022. An ensemble of deep learning frameworks for predicting respiratory anomalies. In: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, pp. 4595–4598. http://dx.doi.org/10.1109/EMBC48229.2022.9871440.

Rocha, B.M.M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y.P., Jakovljevic, N., Turukalo, T.L., Vogiatzis, I.M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Maglaveras, N., Paiva, R.P., Chouvarda, I., de Carvalho, P., 2019. An open access database for the evaluation of respiratory sound classification algorithms. Physiol. Meas. 40 3, 035001. http://dx.doi.org/10.1088/1361-6579/ab03ea.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2009. The graph neural network model. IEEE Trans. Neural Netw. 20 (1), 61–80.

Sharma, N.K., Chetupalli, S.R., Bhattacharya, D., Dutta, D., Mote, P., Ganapathy, S., 2022. The second dicova challenge: Dataset and performance analysis for diagnosis of Covid-19 using acoustics. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 556–560. http://dx.doi.org/10.1109/ICASSP43922.2022.9747188.

Shirian, A., Somandepalli, K., Guha, T., 2022. Self-supervised graphs for audio representation learning with limited labeled data. IEEE J. Sel. Top. Sign. Proces. 16 (6), 1391–1401. http://dx.doi.org/10.1109/JSTSP.2022.3190083.

Suarez-Varela, J., Almasan, P., Ferriol-Galmes, M., Rusek, K., Geyer, F., Cheng, X., Shi, X., Xiao, S., Scarselli, F., Cabellos-Aparicio, A., Barlet-Ros, P., 2022. Graph neural networks for communication networks: Context, use cases and opportunities. IEEE Netw. 1–8. http://dx.doi.org/10.1109/MNET.123.2100773.

Tang, C., Luo, C., Zhao, Z., Xie, W., Zeng, W., 2021. Joint time-frequency and time domain learning for speech enhancement. In: International Joint Conferences on Artificial Intelligence. pp. 3816–3822.

Truong, H.V., Hieu, N.C., Giao, P.N., Phong, N.X., 2021. Unsupervised detection of anomalous sound for machine condition monitoring using fully connected U-net. J. ICT Res. Appl. 15 (1), 41–55. http://dx.doi.org/10.5614/itbj.ict.res.appl.2021.15.1.3.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph attention networks. In: International Conference on Learning Representations.

Vlachos, V., Siklafidis, T., Chantzi, K., 2019. GRATIS: A graph tool for information systems scientists. In: 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference. SEEDA-CECNSM, pp. 1–6. http://dx.doi.org/10.1109/SEEDA-CECNSM.2019.8908332.

Wang, N., Cao, Y., Hao, S., Shao, Z., Subbalakshmi, K., 2021. Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data. In: International Speech Communication Association. INTERSPEECH, pp. 3835–3839. http://dx.doi.org/10.21437/Interspeech.2021-2024.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32 (1), 4–24.

Yadav, K., Tiwari, S., Jain, A., Dafhalla, A.K.Y., 2021. Deep learning based cardiovascular disease diagnosis system from heartbeat sound. Int. J. Speech Technol. http://dx.doi.org/10.1007/s10772-021-09890-4.

Yang, F., Fan, K., Song, D., Lin, H., 2020. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinform. 21 (1), http://dx.doi.org/10.1186/s12859-020-03646-8.

Yang, Y., Yuan, Y., Zhang, G., Hao Wang, Y.-C.C., Liu, Y., Tarolli, C.G., Crepeau, D., Bukartyk, J., Junna, M.R., Videnovic, A., Ellis, T.D., Lipford, M.C., Dorsey, R., Katabi, D., 2022. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. In: Nature Medicine. pp. 2207–2215. http://dx.doi.org/10.1038/s41591-022-01932-x.

Ying, Y., Yang, T., Zhou, H., 2022. Multimodal fusion for alzheimer's disease recognition. Appl. Intell. http://dx.doi.org/10.1007/s10489-022-04255-z.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. AI Open (ISSN: 2666-6510) 1, 57–81.