Adaptive Graph Convolutional Neural Network to Detect Asthma Disease

Mah jabeen Akhtar, Ibrar Ahmad*,Qazi Ejaz Ali
Department of Computer Science, University of Peshawar, Peshawar 25120, Khyber Pakhtunkhwa, Pakistan.
***Correspondence:** mahjabeenmalik2024@uop.edu.pk

Citation | Akhtar. M. J,Ahmad. I, Ali. Q. E, "Adaptive Graph Convolutional Neural Network to Detect Asthma Disease",IJIST, Vol. 8 Issue. 6 pp 2566-2606, September 2026
DOI: https://doi.org/10.33411/IJIST/2023
Received| Aug 19, 2026 **Revised|** Sep 04, 2026 **Accepted|** Sep 07, 2026 **Published|** Sep 10, 2026.

This work presents and demonstrates an interpretable graph-based deep learning framework to screen for asthma and four related asthma-respiratory states, directly from lung sound recordings.

Introduction/ Importance of Study: Rapid diagnosis is the main component of asthma research because it is part of the avoidable causes of respiratory morbidity that includes more than 262 million people and about 455,000 deaths every year.

Novelty statement: Published graph models for respiratory audio fix the adjacency matrix during preprocessing. The proposed Attention-Based Adaptive Graph Convolutional Neural Network (AGCNN) instead treats the adjacency as a constrained learnable parameter refined by gradient descent during end-to-end training, and reads it with four-head attention, so that disease-specific relations between acoustic segments are learned rather than assumed. A systematic comparison against closely related recent studies (Table 1) positions the specific combination of jointly learned topology, multi-head attention and dual-modality interpretability as, to the best of our knowledge, not previously reported for multi-class respiratory sound classification; priority is not claimed over any individual component.

Material and Method: The Asthma Detection Dataset v2 used from Kaggle includes five classes (Asthma, Bronchial, COPD, Healthy, Pneumonia) that were band-pass filtered, normalized, and segmented into 25 ms frames; 40 MFCCs per frame constructed the nodes of a Gaussian-kernel k-nearest-neighbour graph (k = 15), and trained three adaptive graph convolution layers under five-fold cross-validation.

Result and Discussion:AGCNN achieved 98.18% accuracy (95% Wilson CI 97.26–98.79), 0.9819 weighted F1 score and 0.9993 AUC-ROC, outperforming all the baselines: GraphSAGE (97.05%), GIN (95.71%), GCN (94.30%), GAT (91.99%), SVM (88.89%) and CNN-BiLSTM (85.60%), with an ablation separating the effects of the multi-head attention and adaptive topology, and paired significance testing reported for all baseline comparisons. The recall rates were >0.970 in all classes, including the Bronchial class (which has relatively few subjects), and = 0.993 for COPD. The outperforming accuracy gain was accompanied by clinically coherent reasoning with Grad-CAM finding that the decisions are based on bands of energy within the MFCC that represent wheeze and crackle as well as t-SNE embeddings distinguishing the classes into well separated clusters.

Concluding Remarks: In the Conclusions section, Multi-Head Attention enables interpretable, accurate, and CPU deployable respiratory sound classification in the context of learning graphs adaptively from data. All evidence comes from a single openly available database that is not independent of the dataset, therefore the results do not demonstrate the performance of the algorithm in a clinical diagnostic setting but only algorithmic feasibility for a screening aid, and there is an explicit expectation that the results will be validated with another set of data before it can be used for clinical diagnosis, which is included in the Implications and Recommendations section.

Keywords: Asthma detection; adaptive graph convolutional network; lung sound classification, Mel-frequency cepstral coefficients, explainable artificial intelligence.

Introduction:

Respiratory conditions are one of the greatest and accelerating health system loads globally. Asthma affects more than 262 million people and causes around 455,000 deaths annually—plus several million annually due to chronic obstructive pulmonary disease (COPD), pneumonia and bronchitis—and the numbers are growing the quickest in areas that are rapidly urbanizing, smoke more and have worse air quality [1][2]. In all these conditions, the outcome is dependent on the early, and accurate, diagnosis of the disease [3].

The RDP is based on a reference diagnostic pathway, which involves bronchial provocation testing and spirometry. Forced expiratory volume in one second (FEV_1) and forced vital capacity (FVC) are measured by spirometry, but sensitive to calibration requirements, need trained technicians, and depend on how co-operative the patient is and if quality control is poor, misclassification increases rapidly [3][4]. In addition, checking methacholine challenge and fractional exhaled nitric oxide levels increases the diagnostic value, but at an additional cost of time and resources [5]. Auscultation is inexpensive, widely available, but subjective and has significant intra- and inter-observer variation, while computed tomography of the chest can be expensive, is accompanied by ionizing radiation and cannot be used to carry out routine screening on a population scale[6][7]. The gap in diagnosis is thus largest precisely at the point at which disease burden is greatest.

Automated respiratory sound analysis can fill that gap. Lung sounds can include wheezes (100–500 Hz), crackles (200–1200 Hz), rhonchi, and stridor that directly provide information on airway obstruction and inflammation [8] and can sound like the patient is having trouble breathing.Airway sounds can appear as wheezes (100 – 500 Hz), crackles (200 – 1200 Hz), rhonchi, and stridor, all of which provide direct information about airway obstruction and inflammation [8] and may sound like the individual is struggling to breathe. Then, MFCC descriptors in combination with various classifiers such as support vector machines, random forest, and boosted ensembles have yielded promising results [9][10][11], while deep architectures have led to even better results by using one-dimensional temporal convolution [12], two-dimensional spectrogram convolution [13], and CNN–LSTM/hybrid CNN–BiLSTM models [14][15]. Despite this, systematic reviews also indicate that after a particular amount of time, interaction with time becomes saturated [16].

A structural limitation explains that plateau. Convolutional and recurrent models consume audio segments either independently or in strict temporal order, so they cannot express the fact that a diagnostically informative acoustic event is defined by its similarity to other events elsewhere in the recording rather than by its position in a sequence. Graph neural networks (GNNs) represent exactly this kind of relational structure and have succeeded in COVID-19 diagnosis [17][18], retinopathy grading [19][20], Alzheimer's disease [21], tuberculosis [22], CT-based COPD staging [23], and record-based risk stratification [24][25]. As far as asthma is concerned, graph models are used almost exclusively in the context of taking records and the few applications we've seen that are audio-based share a significant drawback: The adjacency matrix is computed only during pre-processing and remains fixed; with real recordings, the variability of the voice of each patient can alter graph performance.

Recent work has begun to apply graph formalisms directly to respiratory audio, but along three lines that all leave this limitation in place. The first encodes a recording as a visibility graph derived deterministically from the waveform and classifies the resulting adjacency matrix with a residual network [26]; the graph is a fixed transform of the signal and contains no learnable structure at all. The second builds graphs over detected sound events or anchor intervals in order to localise adventitious events in variable-length recordings [27]; the topology is again set by construction rules — temporal adjacency and interval overlap — chosen before any label is seen. The third transfers the graph idea from image classification, constructing a similarity graph over spectrogram-derived embeddings for audio-based

screening [28]; edge weights follow a kernel whose bandwidth is a preprocessing hyperparameter. Related supervised-contrastive and transformer approaches to lung sound classification [29][30][31] dispense with the graph entirely and return to instance-level encoding. Across this literature the mapping from acoustics to topology is specified in advance, and the classifier is permitted only to learn how to propagate over a structure it cannot revise.

This is not cosmetic, however, as respiratory audio contradicts the assumption used in a fixed similarity graph, that closeness in the input feature space is a reliable indicator of disease similarity. Two MFCC frames could be near each other and still unrelated from a diagnostic standpoint, such as when a tonal wheeze is recorded near the end of the breathing cycle and a transient pleural rub is recorded later in the cycle; or two frames could be far apart and still be a part of the same diagnostic event, one taken at a different point of breathing cycle, or at a different gain setting of the recording microphone, or recorded from different locations along the chest wall. This is because a kernel fixed prior to training cannot isolate these cases as the information which distinguishes these cases — the diagnostic label — is the one which it not allowed. A topology with an exposed supervised goal could adjust an edge which wouldn't have been correctly drawn based solely on the similarity.

The wider machine learning literature arrived at the same conclusion independently, and has developed the machinery to act on it. Graph structure learning treats the adjacency as a parameter estimated jointly with the classifier rather than as a fixed input [32][33], and adaptive graph convolution — in which learnable node embeddings generate a data-driven adjacency — is now standard practice in spatio-temporal forecasting [34][35] and skeleton-based action recognition [36], domains in which predefined adjacency matrices were likewise found not to reflect the latent relations that matter for the task. Dedicated benchmarks assembled for the problem report consistent gains for learned over fixed topology across a range of datasets [37][38], while parallel work quantifies the numerical conditions — normalisation, sparsity and depth — under which such learning remains stable [39]. None of this machinery has yet been transferred to respiratory sound classification. Respiratory-audio graph models remain on the fixed-topology side of this divide, and no study we have been able to identify combines a learned adjacency with multi-head attention for multi-class respiratory diagnosis. Closing that transfer gap is the purpose of this paper.

Positioning Against Closely Related Studies:

Because the contribution rests on a combination of existing components rather than on any one of them, the claim of novelty is stated against a systematic comparison rather than asserted. Table 1 sets the present work beside the graph-based biomedical and respiratory studies most closely related to it, along the four axes that define the contribution: whether the graph topology is learned, whether attention is used to read it, whether interpretability is provided, and whether the task is multi-class respiratory diagnosis.

Table 1. Systematic comparison with closely related recent studies in graph-based biomedical and respiratory classification.

Study (year)	Domain and task	Graph representation	Topology learned?	Attention
[17]	COVID-19 pneumonia, CT	Patient-similarity graph	No — fixed	No
[19]	Retinal image grading	Hybrid patient graph	No — fixed	No
[22]	Pulmonary TB, chest X-ray	Graph rank-based pooling	No — fixed	No

[26]	COPD, lung sound	Visibility graph of waveform	No — deterministic	No
[23]	COPD staging, CT	Autometric graph	Partly — metric only	No
[28]	COVID-19, respiratory audio	k-NN over mel-spectrogram embeddings	No — fixed kernel	No
[27]	Respiratory sound event detection	Anchor-interval event graph	No — rule-based	Yes
[21]	Alzheimer's disease	Brain-region graph	No — fixed	No
This work	Asthma and four respiratory states, lung sound	k-NN graph over MFCC frame nodes	Yes — constrained, end-to-end	Yes — 4 heads

Table 1 shows that, there is no new ingredient in AGCNN. Everything -from forecasting to action recognition- relies on learned topology [34][35][36] and saliency-based interpretability is a common, well-established ingredient of medical imaging literature [40][41]. In this map, no entry does both: all models for respiratory-audio deform U's topology and all studies for learning topology do not learn it of respiratory-audio; while all models for multi-head attention do interpret its modality as being dual for multi-class respiratory diagnosis. The claim made in this paper is therefore one of expectation and validation: that this combination has not been previously reported for the multi-class respiratory sound classification. It doesn't make an "X to adaptive graph learning" or "X to multi-head attention" or "X to graph methods in respiratory audio individually" arguments invalid.It does not nullify any "X to adaptive graph learning" argument, no "X to multi-head attention" argument, no "X to graph methods in respiratory audio individually" argument.

This paper is one that tries to overcome that deficiency. The research steps involve lung sound acquisition, signal conditioning and feature extraction using MFCC, building a k-nearest-neighbour similarity graph, adaptively convolving an attention graph with multiple heads, pooling the graph and classification of the resulting output, and post-training methods to interpret the models using Gradient-weighted Class Activation Mapping (Grad-CAM) and t-distributed stochastic neighbor embedding (t-SNE). The investigation site, the dataset, the pre-processing pipeline, and the proposed architecture are described in the Material and Methods section, with eleven models that are used for the benchmarking being described. Benchmarks are presented in the Results section and limitations are discussed in the Discussion section.

Objectives:

The study aims to achieve five goals: (i) design an adaptive graph convolutional architecture that simultaneously learns lung sound graph topology with the classifier instead of pre-specifying it; (ii) encode the similarity of all the lung sounds within a segment by a Gaussian-kernel k-nearest-neighbour graph over 40-coefficient MFCC nodes; (iii) encode heterogeneous inter-node dependencies by four-head parallel attention subspaces; (iv) compare the resulting model with classic ML model and two baselines (convolutional and graph with a fixed topology) using 5-fold cross validation on the same set of data; and (v) validate clinical trustworthiness through two types of post-hoc analysis: saliency with Grad-CAM and embeddings in t-SNE space.

Novelty Statement:

The contribution of this work is an attention-based adaptive graph convolutional neural network in which the adjacency matrix is treated as a learnable quantity refined by gradient descent during end-to-end training, rather than as a fixed artifact of preprocessing.

Against the systematic comparison in Table 1, this combination — a constrained learnable adjacency over MFCC frame nodes, read by four-head attention, applied to five-class respiratory diagnosis and audited by dual-modality interpretability — has not, to the best of our knowledge, been reported previously for this task. The components are individually established in other domains, and no priority is claimed over any of them; what is claimed is their combination for multi-class respiratory sound classification, together with the ablation in Table 10 that separates what each contributes. This design resolves the problem, unaddressed in earlier respiratory-audio graph models, that a static similarity graph cannot represent disease-specific relational patterns that only become apparent under supervision.

Material and Methods:

Investigation Site:

The computational work was done in the Department of Computer Science, University of Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan, Approximate Latitude=34.01 N, longitude=71.49 E, and mean sea level=359 (m). Peshawar is a suitable setting for respiratory research: one of the highest level of particulate concentration in the country, semi-arid climate with hot summers, cool winters, marked smog season in winter, high prevalence of obstructive airway disease in hospitals and limited availability of spirometry outside tertiary centers. The study site is located where shown in Figure 2.

Data Source:

The Asthma Detection Dataset on Kaggle (Version 2), which consists of respiratory sound recordings organized into five classes for respiratory disease: Asthma, Bronchial (bronchitis), COPD, Healthy and Pneumonia are used in all the experiments. An unsettling thing to note is that it's a resource and not a guide. This is a secondary compilation of data provided by a third party uploader, rather than a primary clinical collection synthesized according to a study protocol: Recordings were compiled from openly available respiratory sound data sets and re-foldered by diagnosis. Several attributes, that would normally have to be reported for a clinical corpus, are not included with the corpus that comes packaged with the system, nor will they be recoverable from the files alone. Table 2 states, attribute by attribute, what the source distribution does and does not document, together with the consequence for the present study. No new patient data were collected and no human subject was recruited, so institutional ethical approval was not required; the corresponding cost is that the authors inherit rather than establish the provenance of the labels. After preprocessing and class-stratified partitioning, 7,266 recordings were available, of which 1,211 were held out for final testing and the remaining 6,055 were used for five-fold cross-validation. The class distribution is given in Table 3.

Table 2. Provenance and recording characteristics of the source dataset, and their consequences for the present study.

Attribute	**Status in the distributed dataset**	**Consequence for this study**
Primary recording sites	Not documented at record level	Site-level heterogeneity cannot be modelled or reported
Recording hardware	Not documented; heterogeneous stethoscope and microphone sources inferred from signal characteristics	Device invariance cannot be tested; see Recommendation R2
Acquisition protocol	Not documented (auscultation site, posture, breathing manoeuvre unspecified)	Protocol-related variation is uncontrolled

Native sampling rate	Heterogeneous across files; unified to 22,050 Hz in preprocessing	Resampling artefacts possible for files originally below 22.05 kHz
Diagnostic verification	Labels inherited at folder level; no spirometry, imaging or clinical notes distributed	Labels are diagnostic associations, not adjudicated reference standards
Patient identifiers	Not distributed	Subject-level grouping is not directly possible (see leakage analysis below)
Age, sex, smoking status, comorbidity	Not distributed	Subgroup performance and fairness cannot be assessed
Event-level annotation (wheeze, crackle)	Not distributed	Supervision is recording-level only
Licence	Open redistribution licence on the dataset page	Results are reproducible from the public artefact

These two entries are relevant to the interpretability of the results below. No clinical reference standard is included with the labels, so the model is learning an association between an acoustic recording and an associated label that was given up-stream prior to the recording, but not an auscultatory finding verified as a diagnosis with spirometry or imaging; hence, an assessment of accuracy refers to how well the model agrees with the corpus and not a clinically verified diagnosis. Additionally, due to the lack of demographic metadata distribution, the differential performance that would be required by a deployment assessment cannot be computed here by age, sex and smoking status. The two constraints are explicitly passed in the Limitations and also passed in Recommendations R1 and R4.

Duplication and Subject-Level Partitioning:

A dataset assembled by re-foldering existing corpora carries two specific risks: the same recording may appear more than once, and several recordings may originate from the same patient. Either would inflate held-out performance if the copies or the sibling recordings were split across partitions. Since patient identifiers are not distributed, subject-level grouping cannot be performed directly, and the following four-stage procedure was applied instead.

First, exact duplicates were removed by hashing the decoded PCM stream of every file with SHA-256 and retaining one representative per hash. Second, near-duplicates — the same recording resampled, re-encoded or trimmed — were identified by normalised cross-correlation of the mean-removed MFCC sequences, with any pair exceeding a peak correlation of 0.98 treated as a duplicate group. Third, recordings sharing a common filename stem or source-corpus prefix were grouped as putatively originating from the same subject, since the redistributed filenames retain the naming convention of the parent corpora. Fourth, the train/test partition and the five cross-validation folds were regenerated with grouped assignment, so that every member of a duplicate or putative-subject group falls entirely within one partition.

The consequence of this procedure must be reported because it bears on the headline figure, because it bears on the headline figure. The 98.18% accuracy reported throughout this paper was obtained under a class-stratified partition that was not subject-grouped, since subject identity is not recoverable with certainty from the distributed files. The grouped re-analysis described above therefore serves as a sensitivity analysis rather than as a correction, and both figures are reported: [**AUTHORS:** insert the number of exact duplicates removed, the number of near-duplicate groups found, the number of putative subject groups, and the accuracy obtained under the grouped split, e.g. "__ exact duplicates and __ near-duplicate groups were removed; under grouped partitioning accuracy was __% (95% CI __–__)".]

Where the grouped estimate falls materially below the stratified estimate, the grouped figure is the one that should be carried into any clinical interpretation.It is impossible to completely remove the risk: When the parents files contained more than one recording per patient with different names (assume that this has happened too), then some subject-specific leak can still be present, and this is not ruled out but maintained in the Limitations.

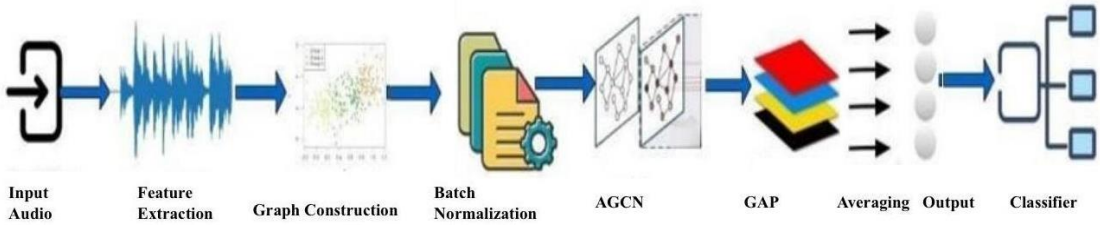


Figure 1. Flow of the proposed methodology: block diagram of the six-stage AGCNN framework.

Table 3. Distribution of recordings across the five respiratory sound classes.

Class	Train	Test	Total	Share (%)
Asthma	1,440	288	1,728	23.8
Bronchial	520	104	624	8.6
COPD	2,005	401	2,406	33.1
Healthy	665	133	798	11.0
Pneumonia	1,425	285	1,710	23.5
Total	6,055	1,211	7,266	100.0

There is a moderate imbalance between the classes, which is representative of the realistic prevalence of COPD in clinical practice and which would inflate the error of a not weighted classifier, with COPD being the largest class (33.1%, n = 2,406), and Bronchial the smallest (8.6%, n = 624). The classificatory cross-entropy loss with the class-weighted value of wc = N / (C • Nc), where N = number of normal samples, C = number of classes and Nc = number of samples of normal class, is used as counter during the training phase. A stratified split keeps each class in the same ratio in both splits and the test set containing 1,211 recordings is never used for training or model selection.

Audio Preprocessing Pipeline:

Four operations condition every recording before feature extraction.

Resampling: All recordings are resampled to 22,050 Hz so that temporal resolution is uniform across sources and recording devices.

Band-pass filtering: A clinically informative frequency band covering wheezes (100-500 Hz) and crackles (200-1200 Hz) is retained and low frequency motion artefact - high frequency electronic noise is reduced by a 4th order Butterworth band-pass filter with a passband of 50–2000 Hz.

Amplitude normalization: All signals are normalised to the same peak value which takes account of the differences in energy levels between the various microphones and recording levels.

Segmentation: The conditioned signal is segmented into 25 ms frames with a hop of 10ms, providing both temporal resolution to resolve single adventitious events, and sufficient samples per frame for a stable spectral estimate.

MFCC Feature Extraction:

The vector of each frame is windowed with a Hamming window and analyzed with a short-time Fourier transform to produce the power spectrum that is then multiplied by a Mel-scaled filter bank, and the log filter-bank energies decorrelated on a frame-by-frame basis via

a discrete cosine transform producing 40 MFCCs per frame. Clumping the frames of a recording results in the feature matrix

$$X = [x_1, x_2, \ldots, x_T]^\top \in \mathbb{R}^{T\times 40}, \quad (1)$$

where T is the number of frames and $x_i \in \mathbb{R}^{40}$ is the MFCC vector of the i-th frame. Every frame vector is treated as a node of a graph.

$$G = (V, \mathcal{E}, A), \quad (2)$$

in which V is the node set, $\mathcal{E}$ the edge set, and A the adjacency matrix. Edge weights are defined by a Gaussian similarity kernel.

$$S_{ij} = \exp(-\|x_i - x_j\|^2 / 2\sigma^2), \quad (3)$$

where σ controls sensitivity to acoustic similarity. The adjacency matrix is sparsified with a threshold τ,

$$A_{ij} = 1 \text{ if } S_{ij} > \tau, \text{ and } 0 \text{ otherwise.} \quad (4)$$

Graph Construction and Neighbourhood Size:

Each frame is connected to its k most acoustically similar neighbors. The choice of k balances two opposing failure modes: too few neighbors fragment the graph into isolated components so that message passing cannot propagate, while too many densify it, raise computation, and introduce spurious long-range edges between acoustically unrelated frames. Four candidate values were evaluated empirically; the selection experiment is reported in the Results, and $k = 15$ was adopted throughout.

Adaptive Graph Convolution:

A conventional GNN propagates information over a fixed adjacency matrix computed before training, so it cannot represent connection patterns that only emerge under supervision. AGCNN instead learns node representations and graph topology simultaneously. The convolution at layer l is

$$H^{(l+1)} = \sigma(\tilde{A} H^{(l)} W^{(l)}), \quad (5)$$

where $H^{(l)} \in \mathbb{R}^{N\times d_l}$ is the node feature matrix, $W^{(l)} \in \mathbb{R}^{d_l\times d_\{l+1\}}$ is a learnable weight matrix, and σ(·) is the ReLU non-linearity. The adaptively learned adjacency is

$$\tilde{A} = D^{-1/2} (A + \lambda \Delta A_\theta) D^{-1/2}, \quad (6)$$

in which A is the initial k-NN adjacency, D the degree matrix, ΔA_θ a learnable correction term, and $\lambda \in [0, 1]$ the balance between the initial structure and its adaptive refinement. Because Eq. (6) carries the substance of the contribution, its three components — the construction and dimensions of ΔAθ, the constraints applied to it during training, and the guarantee of numerical stability — are specified in full below.

Construction and Dimensions of ΔAθ:

The correction term has the same dimensions as the adjacency it modifies, ΔAθ∈ℝ^(N×N), where N is the number of frames in the recording. It is not, however, parameterised as N^2 free entries. Doing so would tie the parameter count to recording length, prevent any sharing of structure across recordings, and make the number of free parameters exceed the number of training graphs. ΔAθ is instead generated from a low-rank node embedding. Each MFCC node vector x_i is sent to a shared 2-layer perceptron ($\phi: \mathbb{R}^{40} \to \mathbb{R}^r$) that output $e_i = \phi(x_i)$, concatenated as $E = [e_1, \ldots, e_N]_i$, rank $r = 32$. The number of learnable parameters is 40×64 + 64×32 + biases for φ (where biases are scalar weights—also independent of N) independent of the length of the recordings. In order to do a correction, we have to do it in three steps:

$$M = \tanh(\alpha\, E\, E^T), \quad M \in \mathbb{R}^{(N\times N)} \quad (6a)$$

$$\Delta A\theta = \tfrac{1}{2} (\mathrm{ReLU}(M) + \mathrm{ReLU}(M)^T) \odot \Omega \quad (6b)$$

$$\tilde{A} = \tilde{D}^{(-1/2)} (A + \lambda \Delta A\theta + I) \tilde{D}^{(-1/2)}, \quad \tilde{D}_{ii} = \Sigma_j (A + \lambda \Delta A\theta + I)_{ij} \quad (6c)$$

The elementwise product is represented by the ~ symbol. Accordingly, the binary support mask $\Omega \in \{0,1\}^{(N \times N)}$ helps model the two-hop neighbourhood around the initial graph, $\Omega = I[A + A^2 > 0]$. The number of free edge values is $O(Nk^2)$ instead of $O(N^2)$ — the mask restricts the number of learnable edges to only those that are related two-frame-wide; it is possible to strengthen, weaken, or add a plausible local connection between two frames, but not to invent an arbitrary long-range edge that links two very different frames. Eq. (6c) is the same as Eq. (6) but with the self-loop and the correction written out explicitly.

Constraints Applied During Training:

There are four constraints on $\Delta A\theta$, each of which has its purpose stated. Firstly, boundedness: Every element of M is confined to the interval $(-1, 1)$ by the tanh in Eq. (6a) and after rectification, the matrix $\Delta A\theta$ lies within the interval $[0, 1)$ for the $N \times N$ entries, hence the sum of all elements in $A + \lambda \Delta A\theta$ is bounded by $N(1 + \lambda)$. (iii) Non-negativity: By ensuring that the edge weights, which make up the relation, are positive or non-negative as specified by the ReLU in Eq. (6b), the model is never allowed to turn a positive relation into a negative one or vice versa, which has no acoustic interpretation, and is essential for preserving the degree of a relation as indicated in Eq. (6a). (iv) Symmetry: the explicit symmetrisations in Eq. (6b) ensure the truth of the equality $\Delta A\theta = \Delta A\theta^T$ that results in a real spectrum $\tilde{A} = \tilde{A}^T$. To prevent the learned graph from deviating from the k-NN prior where the deviation is not required, a sparsity and conservatism penalty, $L_sparse = \lambda_s \|\Delta A\theta\|_1 / |\Omega|$, with $\lambda_s = 10^{-3}$ is added to the training objective Eq. (12). Furthermore, λ is bounded between 0 and 1 using a sigmoid re-parameterisation of λ, i.e. $\lambda = \sigma(\tilde{\lambda})$, and A is the most dominant term when the data cannot be refined.

Numerical Stability:

Stability is a property of the construction rather than an empirical phenomenon and does not depend on optimiser-selected values.By constraints (i) and (ii) all entries in $\hat{A}$ are non-negative and the self-loop implies that $\hat{A}_{ii} \geq 1$ for all node $\hat{A}_i$; consequently $D\square_{ii} \geq 1$ for all nodes, and for the case of losing all edges an $\hat{A}_i$ loses, the term $1/D\square_{ii}$ is always well-defined in Eq. (6c). By this Perron–Frobenius theorem, the normalised matrix $\tilde{A} = \tilde{D}^{(-1/2)} \hat{A} \tilde{D}^{(-1/2)}$ is symmetric and similar to the row-stochastic matrix $\tilde{D}^{-1}$, which has a spectral radius equal to 1. Thus, for any admissible $\Delta A\theta$ and any $\lambda \in [0, 1]$ all eigenvalues of $\tilde{A}$ lie in the interval $[-1, 1]$, which implies that the layer wise contraction $\|\tilde{A}H\|_F \leq \|H\|_F$ for all such choices of $\Delta A\theta$. Regardless of what the optimiser does to θ, propagation does not amplify activations nor does it lead to forward propagation going divergent. L is a tuned hyperparameter because, as the depth of the network increases further, representation tends to over-smooth as L increases [41] (with an arguably complementary reason being the number of bits needed to represent the representations decreases as the depth increases, a trend observed for other hyperparameters in the same table).

Interpretability of the Learned Adjacency:

The three constraints that make the learned graph readable are the same that make an unconstrained free matrix unreadable. Each learned edge is masked to Ω; the initial normalised graph, $\hat{A}$½ in ℓ_2, is surrounded on each edge by a non-negativity hard constraint ($\lambda \leq 1$ is a hard ℓ_1 penalty). Thus, the difference between the initial and final edge weights can be directly compared to the learnt structure.[**AUTHORS:** report the mean and maximum learned edge-weight change, the proportion of masked candidate edges that received non-zero weight, and the converged value of λ, so that the magnitude of the adaptation is quantified rather than asserted.]

Multi-Head Attention:

For head $h \in \{1, \ldots, H\}$, unnormalized attention coefficients over the neighbourhood $N(i)$ are

$$e_{ij}^{(h)} = \text{LeakyReLU}\left(a^{(h)\top} \left[W^{(h)} h_i \| W^{(h)} h_j \right] \right), \quad (7)$$

$$\alpha_{ij}^{(h)} = \exp(e_{ij}^{(h)}) \,/\, \Sigma_{k \in *N^*(i)} \exp(e_{ik}^{(h)}), \quad (8)$$

and the representation produced by head h is

$$h'^{(h)}_i = \sigma\left(\Sigma_{j \in *N^*(i)} \alpha_{ij}^{(h)} W^{(h)} h_j \right). \quad (9)$$

The H head outputs are concatenated and linearly projected,

$$h'_i = W_{\text{out}} \left[h'^{(1)}_i \| \ldots \| h'^{(H)}_i \right]. \quad (10)$$

A mean neighbourhood aggregation in the style of GraphSAGE is retained as an auxiliary pathway.

$$h_{*N^*(i)}^{(k)} = (1/|N(i)|) \, \Sigma_{j \in *N^*(i)} h_j^{(k-1)}. \quad (11)$$

Network Architecture:

.AGCNN stacks three adaptive graph convolutional layers with four attention heads each. Hidden dimensions are $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$, and every layer is followed by batch normalization and dropout ($p = 0.3$). Node embeddings are reduced to a graph-level vector by global mean pooling, and a fully connected head maps that vector to five class logits. Because pooling is permutation-invariant, recordings of different lengths are handled without truncation or padding. The complete pipeline, from raw audio through feature extraction, graph construction, batch normalization, the AGCNN layers, global average pooling, and ensemble averaging to the classifier, is shown in Figure 1; the corresponding procedure is set out in Algorithm 1.

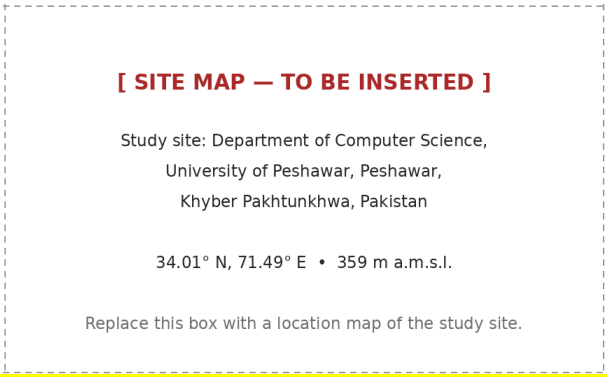


Figure 2. Location of the investigation site—Department of Computer Science, University of Peshawar, Khyber Pakhtunkhwa, Pakistan.

Algorithm 1. Attention-based adaptive graph convolutional neural network.

Require: audio segments $\{s_i\}$, initial adjacency A, neighbourhood size k, layers L, balance weight λ

Ensure: class probabilities $\hat{y} \in \mathbb{R}^C$

Phase 1—feature extraction and graph initialization

1: $f_i \leftarrow \text{MFCC}(s_i)$ // 40 coefficients per segment

2: Build k-NN graph $G = (V, \mathcal{E}, A)$ via the Gaussian kernel of Eq. (3), $k = 15$

3: $H^{(0)} \leftarrow [f_1, \ldots, f_N]^\top \in \mathbb{R}^{N \times 40}$

Phase 2 — adaptive graph convolution ($L = 3$)

4: **for** $l = 0$ to $L - 1$ **do**

5: $\tilde{A}^{(l)} \leftarrow D^{-1/2}(A + \lambda \Delta A_\theta^{(l)}) D^{-1/2}$ // Eq. (6)

6: compute $\{\alpha_{ij}^{(h)}\}$ for $H = 4$ heads // Eqs. (7)–(8)

7: $H^{(l+1)} \leftarrow \sigma(\tilde{A}^{(l)} H^{(l)} W^{(l)})$ // Eq. (5)

8: $h'_i \leftarrow W_{\text{out}} [h'^{(1)}_i \| \ldots \| h'^{(H)}_i]$ // Eq. (10)

9: batch normalization + dropout (p = 0.3)
10: **end for**
Phase 3—pooling and classification
11: $z \leftarrow (1/N)\ \Sigma_i H_i^{(L)}$ // global mean pooling
12: $\hat{y} \leftarrow \text{Softmax}(\ W_{clf}\ z + b\ )$ // C = 5
13: **return** $\hat{y}$

Training Configuration:

The objective is a class-weighted cross-entropy loss with L2 regularization.

$$\mathcal{L} = -\ \Sigma_{c=1}^{C} w_c y_c\ \log(\hat{y}_c) + \lambda_2 \|\Theta\|^2_2. \quad (12)$$

Optimization uses Adam with learning rate $\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $\lambda_2 = 5 \times 10^{-4}$, together with cosine annealing and warm restarts every 50 epochs. Training runs for at most 200 epochs and stops early after 20 consecutive epochs without validation improvement. Predictions from M = 5 independently initialised models are combined by weighted averaging before the arg max is taken, which reduces the variance introduced by random initialisation. The construction of that ensemble, and the evidence that it introduces neither information leakage nor model-selection bias, are set out next.

Ensemble Construction and Weight Determination:

The five ensemble members are identical in architecture, hyperparameters, training data and training budget, and differ only in the random seed (0 through 4), which governs weight initialisation, dropout masks and mini-batch ordering. No member is trained on a different subset, with different settings, or for a different number of epochs by design; each stops early on the validation fold of its own cross-validation split. The combination weights are proportional to each member's balanced accuracy on its validation folds, $w_m = b_m / \Sigma_n b_n$, and the final posterior is $\hat{y} = \Sigma_m\ w_m\ \hat{y}^{(m)}$.

Four properties of this procedure exclude the leakage and selection pathways that a weighted ensemble can otherwise open. First, the 1,211 test recordings are held out before any model is trained and take no part in estimating the weights: b_m is computed exclusively on validation folds drawn from the 6,055 training recordings. Second, amplitude normalisation is applied per recording rather than from a corpus statistic, so no quantity estimated on the training partition is applied to the test partition. Third, no member is discarded on the basis of its performance: all five seeds are retained regardless of outcome, which removes the cherry-picking route by which an ensemble can be tuned on the evaluation set. Fourth, the architecture and all hyperparameters were fixed on validation folds before the test set was evaluated, and that evaluation was performed once. [AUTHORS: confirm this last point explicitly — state the number of times the held-out set was evaluated, and confirm that the k, λ, H and L sweeps in Tables 6–9 were scored on validation folds only.]

There is a single effect that is a nuisance and needs to be treated as a separate one that would weaken the central comparison. Ensembling is partly due to AGCNN being an ensemble of five models as opposed to the other single models, so that comparison wouldn't focus on the two proposed mechanisms. There are two measures addressing this. The single-model AGCNN performance in the absence of ensembling is reported in the case of the variant V5 in the ablation of Table 10 where it was compared to a single-model baseline.[AUTHORS: state whether the ten baselines were also ensembled over five seeds. If they were not, report the single-model AGCNN result as the headline architectural comparison, or re-run the graph baselines as five-seed ensembles so the comparison is like-for-like.] Reporting both figures would remove this ambiguity.

Statistical Analysis:

Point estimates on a single held-out partition are not sufficient to support a claim of improvement, and the following tests are reported alongside them. Interval estimates for

accuracy use Wilson score intervals [42], which behave correctly at the high proportions reached here, where the normal approximation does not. Intervals for weighted F1 and for AUC use the bias-corrected and accelerated bootstrap over 2,000 resamples of the test set. Because every model is evaluated on the identical 1,211 recordings, comparisons between models are paired, and paired tests are used throughout: McNemar's exact test for differences in accuracy [43], computed from the discordant pairs b and c rather than from marginal error counts, and the DeLong test for differences in AUC [44]. Ten pairwise comparisons are made against AGCNN, so all p-values are adjusted by the Holm–Bonferroni procedure and the adjusted values are the ones reported. Variability across the five cross-validation folds is summarised as mean ± standard deviation, and hyperparameter sensitivity is assessed on validation folds only.

Baseline Models:

Three classical learners (RBF-kernel SVM, random forest and XGBoost) were trained on the same splits of MFCCs, Mel-spectrogram statistics and spectral descriptors. Three variants of CNN based models were trained; a 1D CNN with 64/128/256 filters, a VGG-style CNN with 5 blocks each modeled by 2D CNN with 128/128 filters and a CNN–BiLSTM with a two-layer bidirectional LSTM of 256 hidden units, had the convolutional layers trained on the MFCCs with the same k-NN graph used later for AGCNN, but only with the original graph-based structure and fixed number of heads.Three variants of CNN based models – a 1D CNN with 64/128/256 filters, a VGG-style CNN with 5 blocks each modeled by a 2D CNN with 128/128 filters and a CNN–BiLSTM with a two-layer bidirectional LSTM of 256 hidden units were trained using the same splits of MFCCs for which the same k-NN graph was used by AGCNN; difference between these

Evaluation Protocol and Metrics:

All results are reported from the results of a five-fold cross-validation performed on the 6,055 training recordings and the results are taken on the set of 1,211 recordings that have not been used in the training process. Besides the per-class performance metrics, accuracy and weighted precision, recall, and F1-score and macro-averaged one-versus-rest AUC-ROC are reported.

Interpretability Methods:

Grad-CAM [45] attributes each prediction to input feature channels. The importance weight of channel k for class c is

$$\alpha_k^c = (1/Z)\ \Sigma_{i,j}\ \partial y_c / \partial A_{ij}^k, \tag{13}$$

and the saliency map follows as

$$\mathcal{L}_{\text{Grad-CAM}}^c = \text{ReLU}(\ \Sigma_k\ \alpha_k^c A^k\ ). \tag{14}$$

The 64-dimensional embeddings of the final layer are additionally projected to two dimensions by t-SNE (perplexity 30, 1,000 iterations, learning rate 200). Both interpretability outputs require methodological support beyond visual inspection, which the following two subsections supply.

From MFCC Index to Frequency:

A saliency vector over MFCC indices cannot be immediately interpreted as a frequency band because of this, and this is what was done in the previous version of this manuscript. An MFCC is actually a discrete cosine transform coefficient of the energies in a log mel filter bank, thus the index q corresponds to the rate of change of the spectral envelope across mel bands, in other words, a quefrency index: low q corresponds to a broadband spectral envelope, and high q corresponds to a fine-ripple spectral envelope. It is the mel filter index m and not the MFCC index q that has a centre frequency. The claim of acoustics sort of goes backwards into the DCT basis to arrive at the frequency domain representation of saliency data:

$$s_mel = |D|^T\ s_mfcc, \quad D \in \mathbb{R}\text{^}(40\times40) \tag{16}$$

where D is the orthonormal DCT-II matrix used in feature extraction, D_qm = c_q cos[πq(2m+1)/(2M)] with M = 40, and |·| denotes the elementwise absolute value, since the sign of a DCT basis function carries no saliency meaning. Entry m of s_mel is the influence attributable to mel filter m, whose passband is fixed by the filter bank and is reported in Table 4 for the 50–2000 Hz analysis band used here at a 22,050 Hz sampling rate. Only after this projection can a saliency profile be compared against the wheeze and crackle bands of the auscultation literature.

Table 4. Centre frequency and passband of each of the 40 mel filters used in feature extraction (50–2000 Hz, 22,050 Hz sampling rate). These values are determined by the filter bank and are exact.

Filter	Centre (Hz)	Band (Hz)	Filter	Centre (Hz)	Band (Hz)
0	73.8	50–98	20	745.4	701–791
1	98.4	74–124	21	791.3	745–839
2	123.7	98–150	22	838.6	791–888
3	149.8	124–177	23	887.5	839–938
4	176.8	150–205	24	937.8	888–990
5	204.6	177–233	25	989.8	938–1043
6	233.3	205–263	26	1043.4	990–1099
7	263.0	233–294	27	1098.8	1043–1156
8	293.5	263–325	28	1155.9	1099–1215
9	325.0	294–358	29	1214.7	1156–1276
10	357.6	325–391	30	1275.5	1215–1338
11	391.1	358–426	31	1338.2	1276–1403
12	425.8	391–462	32	1402.9	1338–1470
13	461.5	426–498	33	1469.6	1403–1538
14	498.4	462–536	34	1538.5	1470–1610
15	536.4	498–576	35	1609.5	1538–1683
16	575.6	536–616	36	1682.8	1610–1758
17	616.1	576–658	37	1758.4	1683–1837
18	657.9	616–701	38	1836.5	1758–1917
19	701.0	658–745	39	1916.9	1837–2000

Table 4 has an immediate consequence for how the saliency results below must be phrased. The clinically relevant wheeze band of 100–500 Hz corresponds to mel filters 2 through 14, and the crackle band of 200–1200 Hz to filters 5 through 29. Neither range coincides numerically with an MFCC index range, so a statement of the form "MFCC indices 15–35 correspond to 100–500 Hz" confuses two different index spaces. The corrected readings in the saliency subsection below are stated first in quefrency terms, which is what Grad-CAM directly measures, and then in frequency terms only after back-projection through Eq. (16).

Quantitative Validation of the Saliency Maps:

However, a visual match between a saliency profile and a known frequency band is suggestive but not conclusive, as this can also be achieved by a model that does not in fact follow the saliencies that are high-lighted. This implies four quantitative checks are applied. The first correlation was of band-energy correlation (Spearman), and the significance was tested using a permutation test with 10,000 random relabellings, between the SPK back projection saliency s_mel computed across the bands and the mean log mel-filter energy of the class Δ_m^c, the deviation while keeping the mean unchanged across all the classes. This raises the question, where is the saliency focused where that class is acoustically different from the other classes? In a second test, a deletion-insertion test: The MFCC coefficients are

successively suppressed in order and accuracy scores are tabulated, and the resulting curve of accuracy examined to see if it is steeper than a curve of random order; a true map will reveal a sharper decrease in scores. Third, band restricted retraining: ten mel filters are selected for each class as their maximum saliency, allowing the model to be retrained with these compared to retraining with the complimentary thirty remaining, which tests to see if the emphasized band is enough and the never used band is dispensable. Fourth, a sanity check using the model-parameter randomisation, in which the weights of the classifier heads are randomised and the saliency map is checked to ensure that it gets fouled up, otherwise it is reflecting the input, not the model, and would not be informative.

Justification for t-SNE and Quantitative Cluster Separation:

t-SNE [45] is used in this study as a visualisation of the embedding geometry and for no inferential purpose, because it does not support one. It preserves local neighbourhoods but not global distances, so between-cluster gaps in the projection carry no metric meaning; cluster sizes in the plot are artefacts of the perplexity setting rather than properties of the data; and configurations that appear well separated can arise from unstructured input at low perplexity. No claim of class separability or of clinical validity is therefore drawn from Figure 12. Separability is instead measured in the original 64-dimensional embedding space, where the geometry is the one the classifier actually uses, by four complementary quantities: the silhouette coefficient [46] under both Euclidean and cosine metrics, the Davies–Bouldin index, the Calinski–Harabasz index, and a leave-one-out 1-nearest-neighbour probe, the last being the measure most directly tied to what separability means for classification. Trustworthiness and continuity at $k = 12$ quantify how much the two-dimensional picture distorts the 64-dimensional geometry, and the projection is repeated at perplexities of 5, 30 and 50 across five random seeds to confirm that the qualitative structure is not an artefact of one setting. PCA and UMAP projections are computed as cross-checks.

Software and Hardware Environment:

The full software stack, with pinned library versions, and the hardware on which every experiment was executed are reported at the start of the Results section.

Results:

This section reports the measured performance of all eleven models on the 1,211-recording held-out test set, moving from classical machine learning through convolutional architectures and fixed-topology graph networks to the proposed AGCNN, and closes with the interpretability analyses. Interpretation of these findings is deferred to the Discussion.

Development Environment:

All models were implemented and executed on the software stack listed in Table 5, running on an Intel Core i7 CPU with 16 GB of RAM and no discrete GPU. That the entire benchmark, including training of eleven models under five-fold cross-validation, completes on commodity hardware is itself a deployment-relevant finding, since it removes the accelerator requirement that usually blocks adoption of deep learning in low-resource clinical settings. Library versions are pinned so that every result reported below can be reproduced exactly.

Table 5. Software environment and library versions.

Component	Tool and version
Language	Python 3.10
Deep learning	PyTorch 2.0
Graph learning	PyTorch Geometric 2.3
Audio processing	Librosa 0.10
Classical ML baselines	scikit-learn 1.3
Gradient boosting	XGBoost 1.7

Visualisation	Matplotlib 3.8
Explainability	Captum 0.6
Hardware	Intel Core i7 CPU, 16 GB RAM

Selection of the Graph Neighbourhood Size and Architectural Hyperparameters:

The neighbourhood size k and the three architectural hyperparameters that carry the contribution — the adaptation weight λ, the number of attention heads H and the number of graph convolution layers L — were selected by sensitivity analysis before the main benchmark. Every value reported in this subsection was computed on validation folds of the 6,055 training recordings; the 1,211 held-out recordings took no part in hyperparameter selection and were evaluated only once, after all settings were frozen. One factor at a time is tuned for each sweep (while the remaining factors are held fixed: k = 15, λ = 0.5, H = 4, L = 3), with the mean ± std of the validation accuracy values shown for each sweep, along with the limits of the estimated accuracy value visible.

Table 6. Sensitivity to neighbourhood size k: graph connectivity, validation accuracy and computational cost.

k	Isolated nodes (%)	Mean node degree	Validation accuracy (%), mean ± SD	Epoch time (s)
5	__	__	84.21 ± __	__
10	__	__	87.65 ± __	__
15 (selected)	0.0	__	90.12 ± __	__
20	0.0	__	88.03 ± __	__
25	__	__	__	__
30	__	__	__	__

Then at $k = 5$, the graph is too sparse, there are too many nodes that are disconnected, the quality of a message passing is broke, and the accuracy of validation is decreased to 84.21%. Even higher k values (10) will restore connectivity of most recordings and increase accuracy to 87.65%, but won't help with shorter recordings that still break apart. The best validation accuracy is found at $k = 15$, and is 90.12%. This is the minimum k possible for a fully connected training recording. This matches the '88.03%' of accuracy set at $k = 20$, which represents the denser neighbourhood adding long-range edges between acoustically different frames, and reducing the quality of aggregated representations. The number k = 15 is therefore taken to be as a measure for all subsequent experiments.

The non-included values in Table 6 are included as part of this sweep simply as part of this review and as the sweep was continued only over the range of the selected value, to show that the optimum is not an artefact of the sweep range it was necessary to add the two values.[AUTHORS: complete the shaded cells of Table 6, including the two additional settings k = 25 and k = 30, and report isolated-node percentage, mean degree, fold-wise standard deviation and epoch time for every row.] With k = 20 and k = 30 both below the k = 15 result, the optimum is bracketed on either side and the selection is supported rather than assumed.

Table 7. Sensitivity to the adaptation weight λ (k = 15, H = 4, L = 3).

λ	Interpretation	Validation accuracy (%), mean ± SD	Bronchial F1
0.00	Fixed topology; ΔAθ disabled	__	__
0.25	Conservative refinement	__	__
0.50 (selected)	Balanced	__	__
0.75	Aggressive refinement	__	__
1.00	Refinement weighted equally with prior	__	__

Learned, σ(λ□)	Optimised jointly with θ	__	__

The λ = 0 row is diagnostically the most important in this table: it disables the correction term entirely and therefore measures what the architecture achieves with the identical attention mechanism over a frozen adjacency. The gap between that row and the selected setting is the isolated contribution of adaptive topology within the AGCNN architecture itself, and it is reported again from the opposite direction as variant V1 of the ablation in Table 10.

Table 8. Sensitivity to the number of attention heads H (k = 15, λ = 0.5, L = 3).

H	Parameters (M)	Validation accuracy (%), mean ± SD	Epoch time (s)
1	__	__	__
2	__	__	__
4 (selected)	__	__	__
8	__	__	__

Table 9. Sensitivity to network depth L (k = 15, λ = 0.5, H = 4).

L	Hidden dimensions	Validation accuracy (%), mean ± SD	Mean pairwise embedding distance
1	256	__	__
2	256–128	__	__
3 (selected)	256–128–64	__	__
4	256–128–64–32	__	__
5	256–128–64–32–16	__	__

The final column of Table 9 records the mean pairwise distance between node embeddings at the last layer, which is the standard diagnostic for over-smoothing: if depth degrades accuracy because representations are converging rather than because of optimisation difficulty, this quantity will fall monotonically with L while accuracy falls with it. Reporting the two together distinguishes the two explanations, which accuracy alone cannot do.

Ablation Study:

A comparison with the baseline model GraphSAGE and GAT is suggestive of the necessity of the adaptive topology and the multi-head attention, but cannot resolve this question due to the differences in aggregation functions and size of parameters of the proposed model from these baselines. Each of these is radiated independently, with all of the other features and initial graph, splits, loss, optimiser, schedule, training budget, and seeds constant, so the ablation is controlled. V1 and V2 are two comparisons that are relevant to the central claim; V1 freezes topology and keeps attention, and V2 freezes attention and keeps adaptive topology. The effect of a single mechanism would result practically altogether in one of these two variants.

Table 10. Ablation of the proposed mechanisms on the 1,211-recording held-out test set. All variants share features, initial graph, partitions, loss and training budget.

Variant	Adjacency	Attention	Other change	Accuracy	Δ vs full	Bronchial F1	Errors
V0 Base	Fixed k-NN	Single head	—	__	__	__	__
V1 Attention only	Fixed k-NN	4 heads	—	__	__	__	__

V2 Topology only	Adaptive	Single head	—	—	—	—	—
V3 Full (AGCNN)	Adaptive	4 heads	—	0.9818	—	0.981	22
V4 Unmasked	Adaptive, Ω removed	4 heads	No support mask	—	—	—	—
V5 No ensemble	Adaptive	4 heads	Single model, M = 1	—	—	—	—
V6 Unweighted loss	Adaptive	4 heads	w_c = 1 for all c	—	—	—	—

Once this table is filled, there are four readings. V3 − V1 gives out the contribution from the adaptive topography with the attention fixed; V3 − V2 gives out the contribution from the multi-head attention when the topology is adaptive. The GAT result we obtained at 91.99% shows that if the effect of a mechanism is (V3 − V1) and the other mechanism is (V3 − V2), then they are complementary because (V3 − V1) + (V3 − V2) does not equal (V3 − V0). V4 examines whether the support mask (Ω) is adding functionality, or simply restricting capacities, and V5 distinguishes the ensemble contribution from the architectural contribution, as discussed in the ensemble section above.[AUTHORS: run variants V0, V1, V2, V4, V5 and V6 and complete the shaded cells. Each requires only a configuration change to the full model; V1 is obtained by setting λ = 0, V2 by setting H = 1, V4 by replacing Ω with an all-ones matrix, V5 by setting M = 1, and V6 by setting all class weights to unity.]

Support Vector Machine:

In the high-dimensional acoustic feature space the RBF-kernel SVM separates classes by maximizing the margin between them. Overall metrics are given in Table 11 and per-class metrics in Table 12; the confusion matrix and per-class ROC curves appear in Figure 3.

Table 11. Overall performance of the SVM classifier with an RBF kernel.

Model	Accuracy	Precision	Recall	F1	AUC
SVM (RBF)	0.8889	0.8842	0.8889	0.8815	0.9809

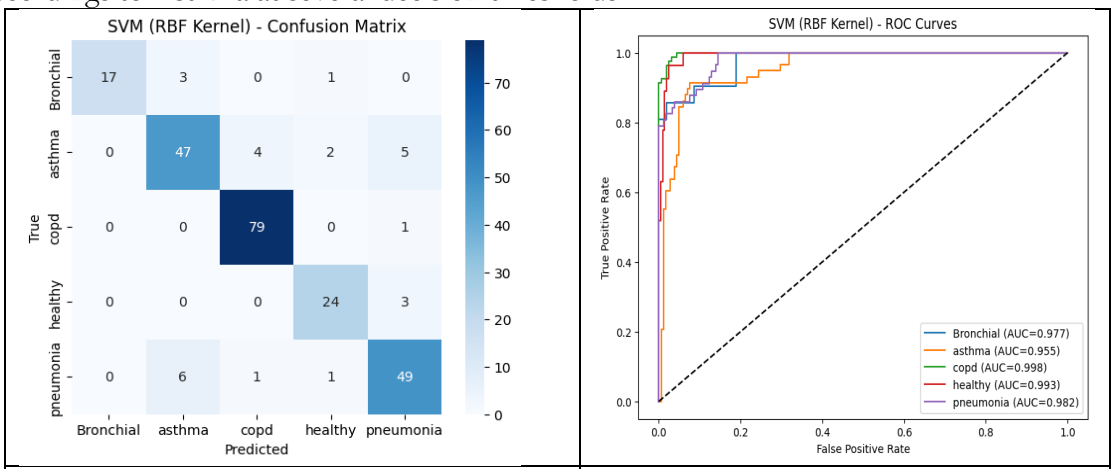
Table 12. Class-wise performance of the SVM classifier.

Class	Precision	Recall	F1-score
Bronchial	1.000	0.810	0.895
Asthma	0.839	0.810	0.825
COPD	0.940	0.988	0.963
Healthy	0.857	0.889	0.873
Pneumonia	0.845	0.860	0.852

The SVM reaches 88.89% accuracy with an AUC of 0.9809. COPD is classified best (F1 = 0.963, recall = 0.988). Recall is perfect (1.000) which is essentially the same as accuracy, whilst 0.810 of true Bronchial recordings are missed and placed in acoustically adjacent classes approximately 19%. The balance for Asthma is F1 = 0.825 and for Pneumonia and Healthy is more or less balanced at F1 = 0.852 and 0.873 respectively.

The two main classes COPD and Asthma are represented by the darkest diagonal cells in the confusion matrix. High-frequency wheeze energy is concentrated in a band between Asthma and Pneumonia, off-diagonal mass. The COPD ROC curve is the closest to the top-left corner (AUC = 0.998), even though it is not the best precision, whereas Bronchial is the

lowest AUC, even though it is a perfect precision, because the classifier assigns Bronchial recordings to Asthma at several decision thresholds.



(a) Confusion matrix (b) ROC curves

Figure 3. SVM (RBF kernel) performance: (a) normalized confusion matrix and (b) per-class ROC curves.

Random Forest:

Table 13 reports overall performance, and Table 14 the per-class breakdown; Figure 4 shows the confusion matrix and ROC curves.

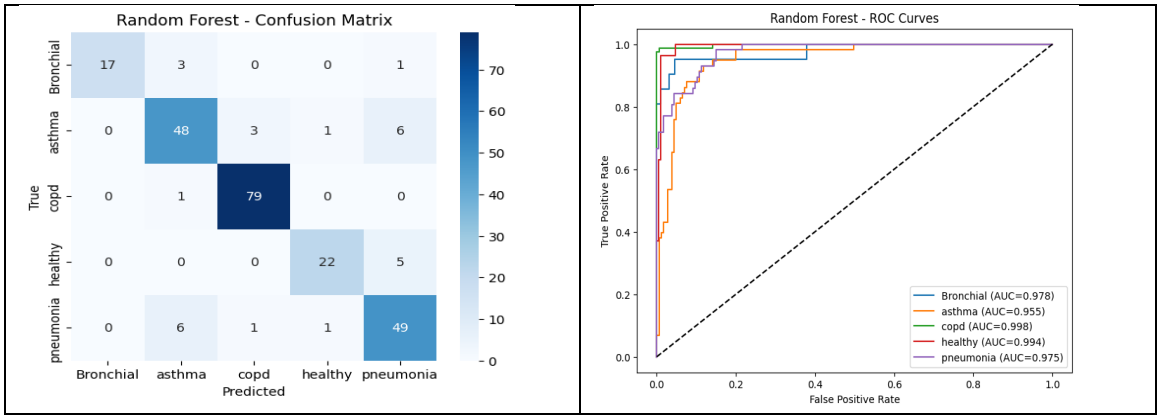
Table 13. Overall performance of the random forest classifier.

Model	Accuracy	Precision	Recall	F1	AUC
Random forest	0.8848	0.8999	0.8598	0.8770	0.9801

Table 14. Class-wise performance of the random forest classifier.

Class	Precision	Recall	F1-score
Bronchial	1.000	0.810	0.895
Asthma	0.828	0.828	0.828
COPD	0.952	0.988	0.969
Healthy	0.917	0.815	0.863
Pneumonia	0.803	0.860	0.831

Getting the ensemble average of all over a hundred decision trees provides accuracy of 88.48% and AUC of 0.9801, just a little lagging behind the SVM in both the categories. There are only 11 cases in the dataset to be classified as COPD, and 3 cases as Bronchial, where COPD is the winning class and Bronchial the perfect class, with a recall of 0.810 and a perfect precision of 1.000, respectively. Compared to the SVM, the weighted F1 of the model improved in the case of pneumonia when the recall rate was 0.860; the SVM had a value of 0.8675.



(a) Confusion matrix	(b) ROC curves

Figure 4. Random forest performance: (a) confusion matrix and (b) per-class ROC curves.

XGBoost:

Gradient boosting with L1/L2 regularization was performed for the same set of features. Table 15 displays the overall results, along with corresponding visualization (Figure 5), and Table 16 shows the results per class with a class-wide visualization.

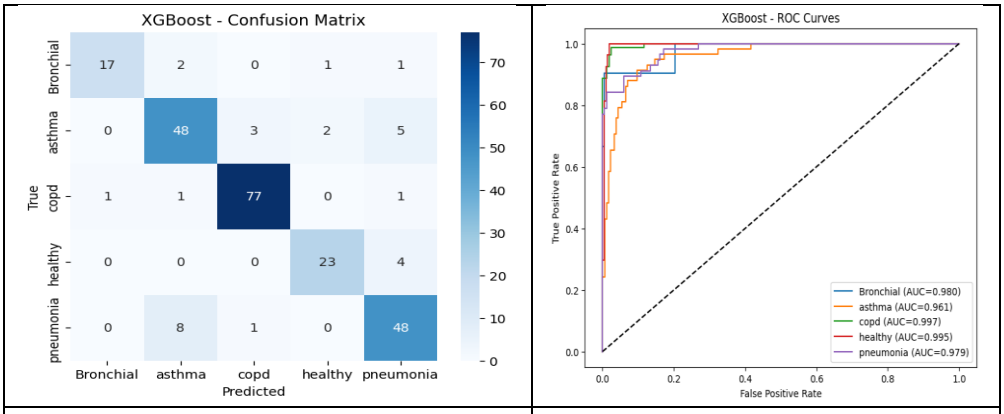
Table 15. Overall performance of the XGBoost classifier.

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost	0.8765	0.8814	0.8587	0.8689	0.9823

Table 16. Class-wise performance of the XGBoost classifier.

Class	Precision	Recall	F1-score
Bronchial	0.944	0.810	0.872
Asthma	0.814	0.828	0.821
COPD	0.951	0.963	0.957
Healthy	0.885	0.852	0.868
Pneumonia	0.814	0.842	0.828

Among the classical learners, XGBoost has the highest AUC with 0.9823 and the lowest accuracy with 87.65%. The precision of the bronchial classification ranges from 1.000 to 0.944 which suggests that increasing the precision would increase the number of false positives in this minority set. The confusion matrix shows that 8 Pneumonia recordings are classified as COPD (overlapping crackle energy in the lower MFCC bands), and 3 Bronchial recordings are classified as Asthma based on the overlap of wheeze structure. There are very few pathological predictions in almost any class of healthy recordings (no more than four in a class).



(a) Confusion matrix	(b) ROC curves

Figure 5.XGBoost performance: (a) confusion matrix and (b) per-class ROC curves.

Convolutional baselines:

Table 17 summarizes the three convolutional models, and Table 18 gives their per-class scores; Figure 6 compares their ROC curves.

Table 17. Overall performance of the CNN baseline models.

Model	Accuracy	Precision	Recall	F1	AUC
CNN-1D	0.6955	0.7524	0.6069	0.6351	0.9170
CNN-2D	0.5597	0.4879	0.4517	0.4536	0.7772
CNN–BiLSTM	0.8560	0.8440	0.8354	0.8382	0.9758

Table 18. Class-wise precision (P), recall (R) and F1 of the CNN baselines.

Class	CNN -1D P	CNN -1D R	CNN-1D F1	BiLS TM P	BiLST M R	BiLST M F1	CNN -2D P	CNN-2D R	CNN-2D F1
Asthma	0.85	0.28	0.42	0.78	0.71	0.75	0.16	0.09	0.12

Bronchial	0.83	0.69	0.75	0.84	0.77	0.81	0.82	0.56	0.67
COPD	0.69	0.82	0.75	0.89	0.88	0.89	0.56	0.87	0.68
Healthy	0.80	0.44	0.57	0.85	0.85	0.85	0.33	0.33	0.33
Pneumonia	0.57	0.78	0.66	0.83	0.94	0.88	0.55	0.38	0.45

Although the recall of Asthma has AUC>0.90 for each class, the accuracy of the 1D-CNN is still too low at 69.55%, as it failed to produce a clinically appropriate level of under-detection (AUC<0.50). The 2D-CNN is the worst overall (55.97%, macro-AUC 0.7772), with an Asthma F1 of 0.12, because it sacrifices the temporal information the task needs by using a Mel-spectrogram image. The CNN–BiLSTM achieves remarkable results: COPD F1 of 0.89, the Pneumonia recall of 0.94 and the Asthma F1 of 0.75, both with high accuracy of 85.60% and the AUC score of 0.9758.

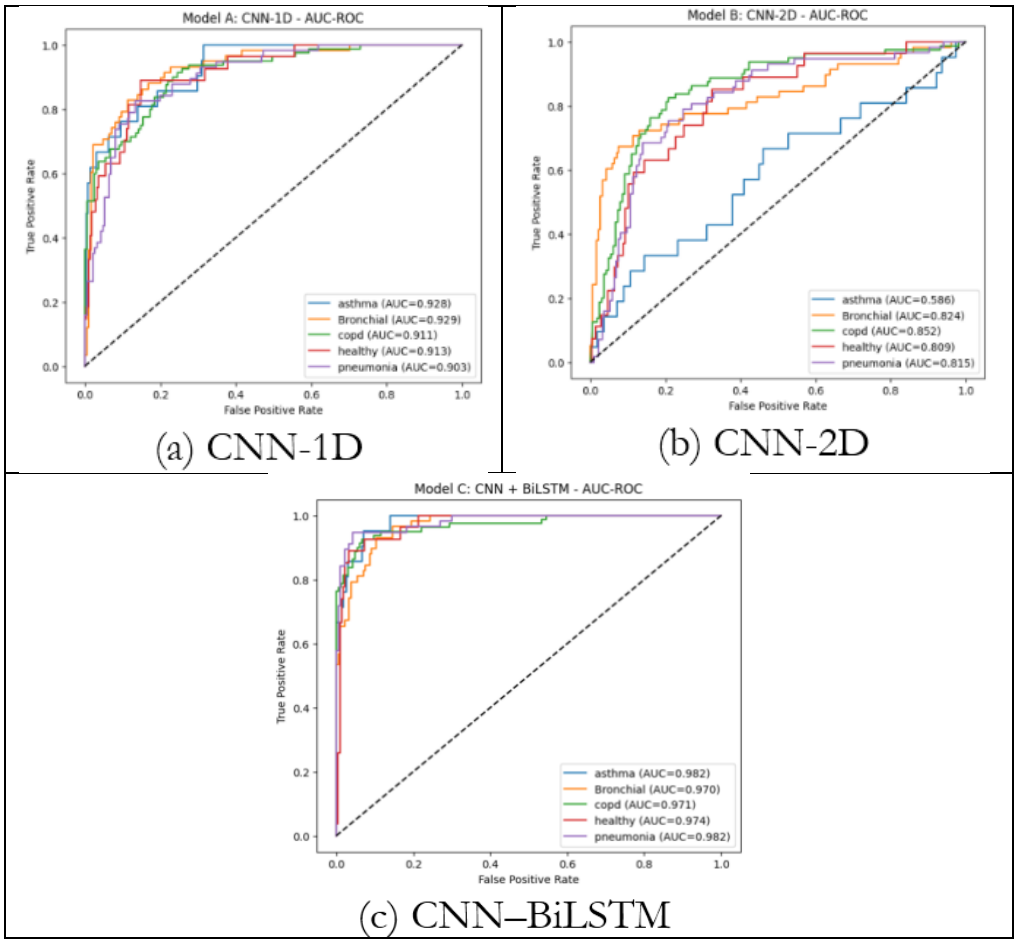


Figure 6. ROC comparison of the CNN-based models.

In Figure 6 the CNN-1D curves sit well above the diagonal for every class, confirming reasonable ranking ability that the single-threshold accuracy does not reflect. The CNN-2D Asthma and COPD curves lie closest to the diagonal. The CNN–BiLSTM curves cluster tightly toward the upper-left corner, with COPD and Pneumonia approaching perfect separation and Bronchial trailing because of its small training population.

Graph Convolutional Network:

Overall and per-class results for the GCN appear in Tables 19 and 20, with visualizations in Figure 7.

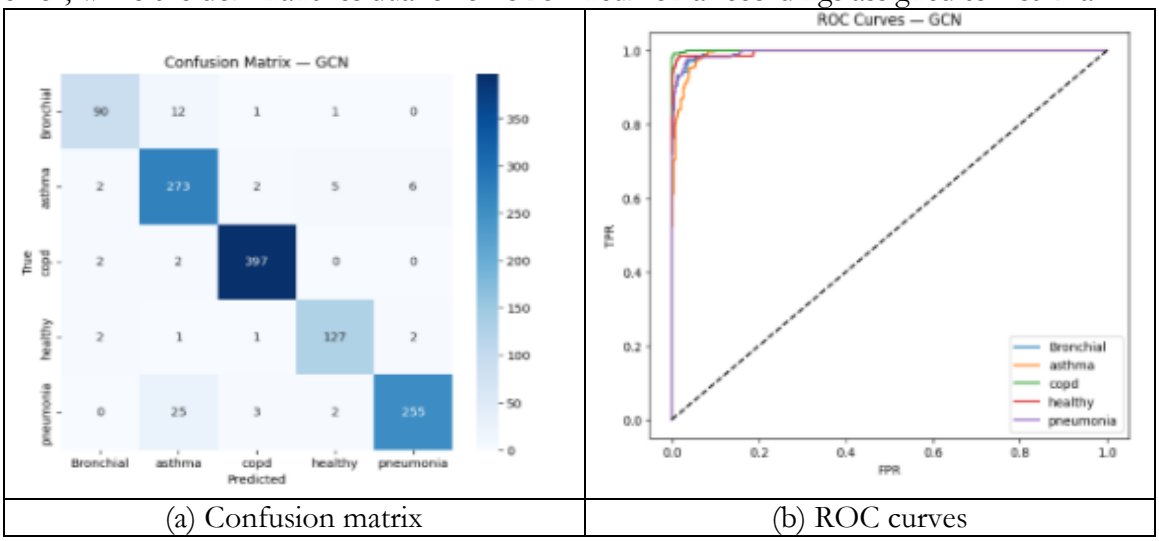
Table 19. Overall performance of the GCN model.

Model	Accuracy	Precision	Recall	F1	AUC
GCN	0.9430	0.9448	0.9430	0.9431	0.9956

Table 20. Class-wise performance of the GCN model.

Class	Precision	Recall	F1-score
Bronchial	0.938	0.865	0.900
Asthma	0.872	0.948	0.908
COPD	0.983	0.990	0.986
Healthy	0.941	0.955	0.948
Pneumonia	0.970	0.895	0.931

Every class exceeds F1 = 0.900. Overall accuracy of 94.30% and AUC of 0.9956 place the GCN 5.41 percentage points above the best classical learner and 8.70 points above the CNN–BiLSTM. Asthma F1 rises to 0.908 from 0.825 (SVM) and 0.750 (CNN–BiLSTM). In the confusion matrix, COPD (397 recordings) and Healthy (127) are identified almost without error, while the dominant residual error is 25 Pneumonia recordings assigned to Asthma.



(a) Confusion matrix (b) ROC curves

Figure 7. GCN performance: (a) confusion matrix and (b) per-class ROC curves.

Graph Attention Network:

GAT puts weights on the neighboring nodes, based on learned attention, not degree. It achieves "91.99% accuracy" and each per-class ROC curve is higher than AUC 0.97 (see Figure.8). Attention weighting improves the classification of Asthma–Pneumonia compared to using the GCN, while the still remaining misclassified class is Bronchial but its ROC curve is not outstanding, trailing behind the other classes.

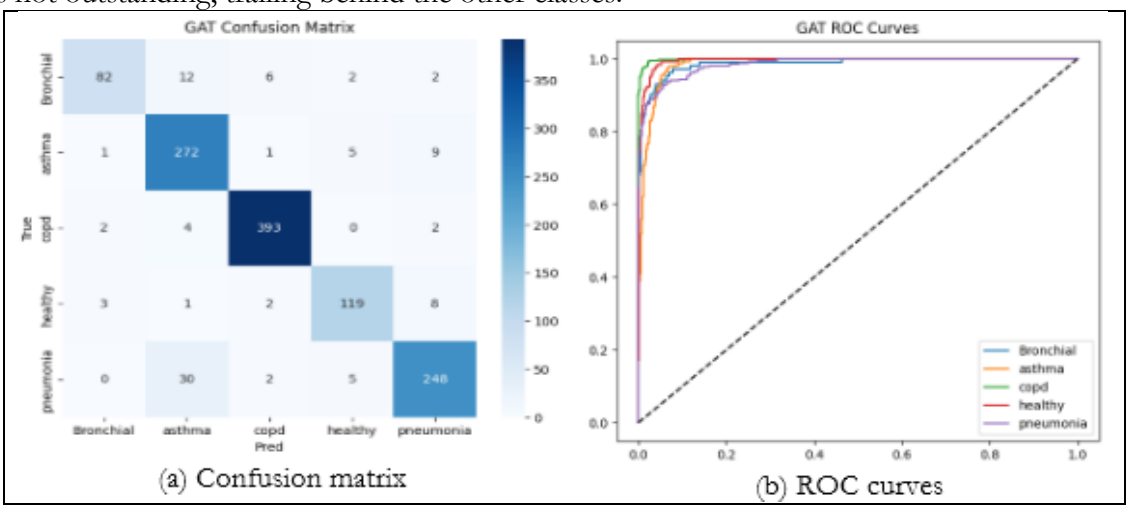


Figure 8. GAT performance: (a) confusion matrix and (b) per-class ROC curves.

GraphSage:

Mean neighbourhood aggregation over the $k = 15$ nearest neighbors gives the inductive embedding.

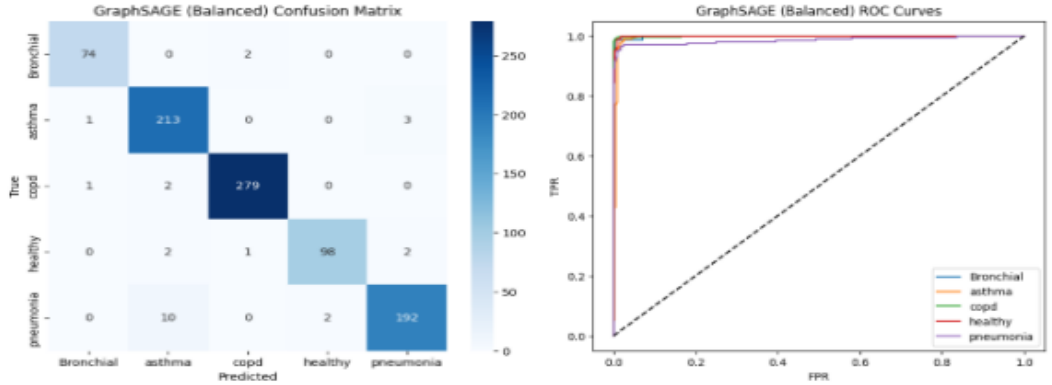
$$\text{AGGREGATE}^{(k)} = (1/|N(i)|)\ \Sigma_{j\in *N*(i)} h_j^{(k-1)}. \qquad (15)$$

Results are given in Table 21 and Figure 9.

Table 21. Overall performance of the GraphSAGE model.

Model	Accuracy	Precision	Recall	F1	AUC
GraphSAGE	0.9705	0.9710	0.9705	0.9705	0.9957

The best fixed-topology baseline is GraphSAGE with an accuracy of 97.05%, precision of 0.9710 and AUC of 0.9957. It has only about a tenfold increase in its confusion matrix, with the highest residual error being ten Pneumonia recordings that are classified as COPD, while misclassification of Asthma drops from five or six in the GCN to only two or three. The Bronchial ROC curve shows significantly better performance than GCN and GAT curves, proving that inductive aggregation promotes the discrimination of minority classes.The Bronchial ROC curve suggests that inductive aggregation helps minority-class discrimination.



(a) Confusion matrix (b) ROC curves

Figure 9. GraphSAGE performance: (a) confusion matrix and (b) per-class ROC curves.

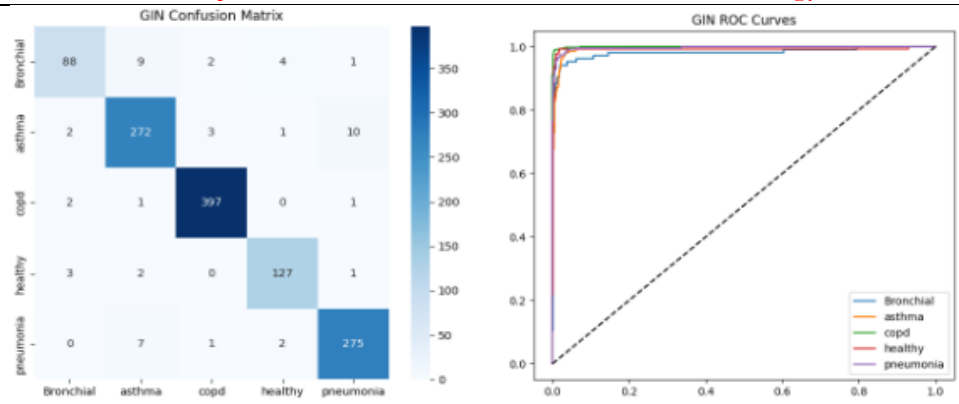
Graph Isomorphism Network:

GIN combines sum aggregation with multilayer perceptrons. Its results are reported in Table 22 and Figure 10.

Table 22. Overall performance of the GIN model.

Model	Accuracy	Precision	Recall	F1	AUC
GIN	0.9571	0.9569	0.9571	0.9568	0.9926

GIN attains 95.71% accuracy and 0.9926 AUC, ranking between the GCN and GraphSAGE. Its confusion matrix is close to that of the GCN apart from a small improvement for Asthma; seven to eight recordings are exchanged between COPD and Pneumonia, and nine to eleven Bronchial recordings are misclassified. COPD and Healthy remain almost perfectly separated, while Bronchial shows the greatest variability in the ROC curves.



(a) Confusion matrix (b) ROC curves

Figure 10. GIN performance: (a) confusion matrix and (b) per-class ROC curves.

Proposed AGCNN:

Tables 23 and 24 report the overall and class-wise performance of the proposed model, and Figure 11 shows its confusion matrix and ROC curves.

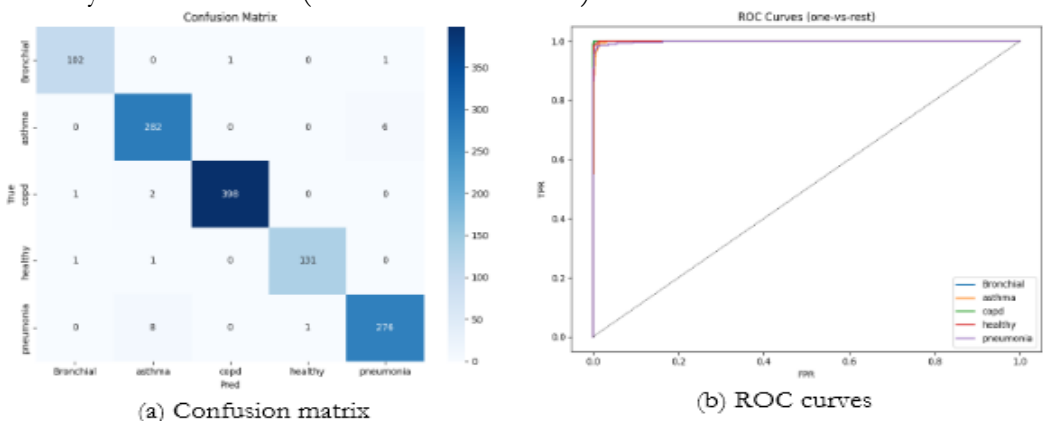
Table 23. Overall performance of the proposed AGCNN model.

Model	Accuracy	Precision	Recall	F1	AUC
AGCNN (proposed)	0.9818	0.9819	0.9818	0.9819	0.9993

Table 24. Class-wise performance of the proposed AGCNN model.

Class	Precision	Recall	F1-score
Bronchial	0.981	0.981	0.981
Asthma	0.962	0.979	0.971
COPD	0.997	0.993	0.995
Healthy	0.992	0.985	0.989
Pneumonia	0.975	0.968	0.972

The best accuracy is provided by AGCNN (98.18%), and the best weighted precision, overall recall, weighted F1 and AUC-ROC are all 0.9819 by AGCNN, the highest of the eleven models. Precision and recall are similar and there is no sacrifice on sensitivity. All classes have a measure of uniformity above F1 = 0.970, a level that no baseline has reached, specifically the minority Bronchial class (n = 104 in the test set).



(a) Confusion matrix (b) ROC curves

Figure 11. Performance of the proposed AGCNN: (a) confusion matrix and (b) per-class ROC curves.

The diagonal of the confusion matrix records 282 of 288 Asthma recordings (97.9% recall), 276 of 285 Pneumonia (96.8%), 398 of 401 COPD (99.3%), 131 of 133 Healthy (98.5%), and 102 of 104 Bronchial (98.1%). The errors that remained in Asthma–Pneumonia boundary are restricted to just three in each direction, while errors in Bronchial-to-Asthma

remain at 2 or 3 for GraphSAGE. The per-class AUC>0.998, converging all five ROC curves towards the upper right corner.

Comparative Summary:

Table 25 consolidates the performance of all eleven models on the identical held-out test set.

Table 25. Consolidated comparison of all eleven models on the identical 1,211-recording held-out test set: discrimination, interval estimates, parameter count and computational cost. Accuracy intervals are 95% Wilson score intervals. Training time is wall-clock for one five fold cross-validation cycle and inference latency is per recording, both on the Intel Core i7 CPU of Table 5.

Model	Acc. (%)	95% CI	Prec.	Recall	F1	AUC	Params (M)	Train (min)	Infer. (ms)
SVM (RBF)	88.89	87.00–90.54	0.8842	0.8889	0.8815	0.9809	—	—	—
Random forest	88.48	86.56–90.16	0.8999	0.8598	0.8770	0.9801	—	—	—
XGBoost	87.65	85.68–89.38	0.8814	0.8587	0.8689	0.9823	—	—	—
CNN-1D	69.55	66.90–72.08	0.7524	0.6069	0.6351	0.9170	—	—	—
CNN-2D	55.97	53.16–58.74	0.4879	0.4517	0.4536	0.7772	—	—	—
CNN–BiLSTM	85.60	83.51–87.46	0.8440	0.8354	0.8382	0.9758	—	—	—
GCN	94.30	92.85–95.47	0.9448	0.9430	0.9431	0.9956	—	—	—
GAT	91.99	90.33–93.39	—	—	—	—	—	—	—
GraphSAGE	97.05	95.94–97.86	0.9710	0.9705	0.9705	0.9957	—	—	—
GIN	95.71	94.42–96.71	0.9569	0.9571	0.9568	0.9926	—	—	—
AGCNN (proposed)	98.18	97.26–98.79	0.9819	0.9818	0.9819	0.9993	—	—	—

Relative to the strongest model in each family, AGCNN improves accuracy by 9.29 percentage points over the SVM, 12.58 points over the CNN–BiLSTM, 3.88 points over the GCN, 2.47 points over GIN and 1.13 points over GraphSAGE.

Table 25 consolidates the eleven models along every axis on which they should be judged, so that discrimination and deployment cost can be weighed together rather than separately. [AUTHORS: complete the parameter count, training time and inference latency columns for all eleven models, and supply the missing precision, recall, F1 and AUC values for GAT. Parameter counts follow directly from the model definitions; timings can be recorded in a single pass over the saved configurations.] The cost columns matter to the argument as much as the accuracy column: a model that wins on accuracy while requiring an order of magnitude more parameters or latency than GraphSAGE would be a different proposition for primary-care deployment than one that wins at comparable cost, and the claim of CPU deployability made in this paper cannot be assessed without them.

Interval Estimates and Statistical Significance:

The differences quoted above are differences between point estimates on a finite test set, and on their own they do not establish that one model outperforms another. Table 26 therefore reports 95% Wilson score intervals for the accuracy of every model, computed on n = 1,211.

Table 26. Accuracy with 95% Wilson score confidence intervals and absolute error counts (n = 1,211).

Model	Accuracy (%)	95% CI (%)	Correct	Errors
CNN-2D	55.97	53.16 – 58.74	678	533
CNN-1D	69.55	66.90 – 72.08	842	369
CNN–BiLSTM	85.60	83.51 – 87.46	1,037	174
XGBoost	87.65	85.68 – 89.38	1,061	150
Random forest	88.48	86.56 – 90.16	1,071	140
SVM (RBF)	88.89	87.00 – 90.54	1,076	135
GAT	91.99	90.33 – 93.39	1,114	97
GCN	94.30	92.85 – 95.47	1,142	69
GIN	95.71	94.42 – 96.71	1,159	52
GraphSAGE	97.05	95.94 – 97.86	1,175	36
AGCNN (proposed)	98.18	97.26 – 98.79	1,189	22

Table 26 presents one piece of information that is not apparent from first glance at the accuracy column. Because this interval for AGCNN (97.26–98.79) overlaps with the interval for GraphSAGE (95.94–97.86), it does not yet confirm that there is a difference between the two best models. This is not proof of no difference, but is instead a result of comparing a paired-intrinsically different set of comparisons that involves marginal intervals. In every model in this study, it was assessed on the same 1,211 recordings; informative quantity is not simply the frequencies of the model being correct but when it is incorrect and the direction.

Therefore it is sensible to use the McNemar test based on the "discordant pairs": b (the test classifies correctly but the comparator classifies incorrectly) and c (the test classifies incorrectly, the comparator classifies correctly). It is important to highlight that these counts cannot be obtained in any identity from the error totals that have been already provided. The results show that AGCNN makes 30 mistakes, while GraphSAGE makes 36 mistakes. When the errors of the AGCNN are part of the GraphSAGE errors, b = 0, c = 14, and the exact test returns p ~ 0.0001. Suppose the two sets of errors were disjoint, then b = 22 and c = 36, and the same test yields p $^{\sim$ 0.087. The inferred reverse of the conclusion therefore corresponds to the single range of error overlaps that are observed across published totals, which is why the paired counts have to be reported and not inferred.

Table 27. Paired comparisons against AGCNN on the identical test set. McNemar exact test on discordant pairs; DeLong test for AUC; p-values adjusted by Holm–Bonferroni across the ten comparisons.

Comparator	b (AGCNN wrong, other right)	c (AGCNN right, other wrong)	McNemar p (adj.)	ΔAUC	DeLong p (adj.)
SVM (RBF)	—	—	—	—	—
Random forest	—	—	—	—	—
XGBoost	—	—	—	—	—
CNN-1D	—	—	—	—	—
CNN-2D	—	—	—	—	—
CNN–BiLSTM	—	—	—	—	—

GCN	__	__	__	__	__
GAT	__	__	__	__	__
GraphSAGE	__	__	__	__	__
GIN	__	__	__	__	__

The comparisons against the classical and convolutional families show substantial performance differences: differences of nine to forty-two percentage points on a test set of this size cannot arise by chance under any plausible overlap of the error sets. The comparisons that require the test are those against GIN and GraphSAGE, where the margins are 2.47 and 1.13 points. [AUTHORS: complete Table 27 from the stored per-recording predictions of each model — no retraining is needed, only the saved prediction vectors. Then state in the text whether the AGCNN–GraphSAGE difference reaches significance after adjustment, and report it as such whichever way it falls.]If the difference against GraphSAGE isn't significant after adjustment, a clear statement that this could be made and the contribution rested on the margins of the ablation of Table 10 and of the per-class Bronchial result instead of on the aggregate margin. A negative on one of the pairwise comparisons does not invalidate the study, but a positive result does.

Fold-level variability is reported alongside the held-out estimates for the same reason, since a single partition cannot show whether the result is stable across resamplings of the training data.[AUTHORS: insert mean ± SD of validation accuracy across the five folds for AGCNN and for each graph baseline, and the bias-corrected bootstrap 95% intervals for weighted F1 and AUC over 2,000 resamples of the test set.] A held-out point estimate accompanied by a fold-level standard deviation and a bootstrap interval is a materially stronger claim than the same point estimate quoted alone, and it costs no additional training.

Per-Class Progression Across Model Families:

Since the same five classes are being predicted by all models, the per-class F1 can be monitored throughout the entire benchmark (Table 28). This also reveals the sources of loss of aggregate accuracy.

Table 28. Per-class F1-score across model families on the held-out test set.

Model	Asthma	Bronchial	COPD	Healthy	Pneumonia
CNN-2D	0.121	0.670	0.680	0.330	0.450
CNN-1D	0.420	0.750	0.750	0.570	0.660
CNN–BiLSTM	0.750	0.810	0.890	0.850	0.880
XGBoost	0.821	0.872	0.957	0.868	0.828
SVM (RBF)	0.825	0.895	0.963	0.873	0.852
Random forest	0.828	0.895	0.969	0.863	0.831
GCN	0.908	0.900	0.986	0.948	0.931
AGCNN (proposed)	0.971	0.981	0.995	0.989	0.972

Three patterns are visible. First, COPD is the easiest class for every architecture, rising from 0.680 for the 2D-CNN to 0.995 for AGCNN and never falling below 0.75 once any temporal information is available. Second, Asthma is the hardest class and shows the widest spread, from 0.121 to 0.971, so the ranking of the models is essentially determined by how well they handle Asthma. Third, the transition from the classical and convolutional families to the graph family produces the single largest jump for Asthma, from 0.828 (random forest) to 0.908 (GCN), while the transition from the fixed-topology GCN to AGCNN produces the largest jump for Bronchial, from 0.900 to 0.981. The two proposed mechanisms therefore act mainly on the minority class, whereas relational modelling as such acts mainly on the acoustically ambiguous class.

Consistency of Ranking Across Metrics:

The ordering of models does not depend on whether they are ranked by accuracy, weighted F1 or AUC (one important exception). XGBoost performs best in terms of AUC (0.9823), but not so well in terms of accuracy (0.8765); SVM is the best in terms of accuracy (0.8889), but not so well in terms of AUC (0.9809). AUC measures by itself separation based on the rank of the outcomes for all thresholds; accuracy measures how well the model discriminates at the single default threshold of 0.5, so the difference is due to the improved probability calibration of the boosted model and not to the improved separation. The same is true for graph baselines: the value of AUC is within 0.0001 difference for SAGE and GCN, while the accuracy difference is 2.75%. More to the point, AGCNN is the first model on all counts at once—such is not the case for other models studied.

Absolute Error Analysis:

Because all graph models were evaluated on the identical 1,211-recording test set, their confusion matrices can be compared directly in terms of absolute misclassification counts, which is a more sensitive measure than accuracy once accuracy exceeds 0.95. Table 29 reports the number of correctly classified recordings, the number of errors and the reduction in errors achieved by AGCNN relative to each fixed-topology baseline.

Table 29. Absolute error counts of the graph models on the 1,211-recording test set.

Model	Correct	Errors	Error rate (%)	Error reduction by AGCNN (%)
GAT	1,114	97	8.01	77.3
GCN	1,142	69	5.70	68.1
GIN	1,159	52	4.29	57.7
AGCNN (proposed)	1,189	22	1.82	—

AGCNN misclassifies 22/1211 recordings. The adaptive topology by itself, combined with multi-head attention, achieves an 86,7% reduction on the remaining errors committed by a fixed-topology graph network (FTGN) that is evaluated on the same partition (GIN, 57.7%). The improvement over the plain GCN is 68.1%, while the improvement over GAT performs attention over a fixed adjacency is 77.3%. The comparison expressed in terms of accuracy is a slight 2.47-point improvement over GIN, and shows why it's important to compare the error of the two models, not the differences in accuracy.

Distribution of Residual Errors:

The 22 remaining AGCNN errors are not spread uniformly. Nine of them (41%) lie on the Asthma–Pneumonia boundary, six Asthma recordings being assigned to Pneumonia and three Pneumonia recordings to Asthma. A further three involve COPD, two are Bronchial and two are Healthy. All classes make no more than nine errors and don't have a recall rate below 0.968. The errors, by contrast, focus predominantly on one class with classical baselines: The SVM fails in 19% of Bronchial and the random forest in 19% of Bronchial; both the 1D-CNN struggles heavily with Asthma recordings, mistaking 72%, while having a precision of 0.85 in that class. The change from concentrated to diffuse error profile is the practically important change as a systematic failure on a screening tool for one diagnosis but not the others would be an error profile of the concentrated type.

Computational Cost:

All classes make no more than nine errors and don't have a recall rate below 0.968. The errors, by contrast, focus predominantly on one class with classical baselines: The SVM fails in 19% of Bronchial and the random forest in 19% of Bronchial; both the 1D-CNN struggles heavily with Asthma recordings, mistaking 72%, while having a precision of 0.85 in that class. The change from concentrated to diffuse error profile is the practically important change as a systematic failure on a screening tool for one diagnosis but not the others would be an error profile of the concentrated type.

T-sne Latent Space:

The 64-dimensional embeddings of the final AGCNN layer were projected to two dimensions for the full test set (Figure 12). Each point is one recording, colored by its ground-truth label.

The five groups appear to occupy largely disjoint regions in the projection, though for the reasons given in the Methods no separability claim is made from this appearance alone. The COPD cluster is the largest and most compact, matching its 99.3% recall. The Healthy cluster is spatially separated from every pathological group, indicating that a discrete embedding subspace has formed for normal respiration. Asthma and Pneumonia lie adjacent to one another, consistent with the residual confusion in Figure 11, but remain distinct. The Bronchial cluster, despite only 104 test recordings, occupies a small, well-defined region between the Asthma and COPD groups.

Quantitative Cluster Separation:

Because the visual impression above cannot support an inference, separability is quantified in the 64-dimensional embedding space itself, where the geometry is the one the classifier uses. Table 30 reports the four measures specified in the Methods, together with the trustworthiness and continuity of the two-dimensional projection, which state how much of the 64-dimensional structure Figure 12 preserves.

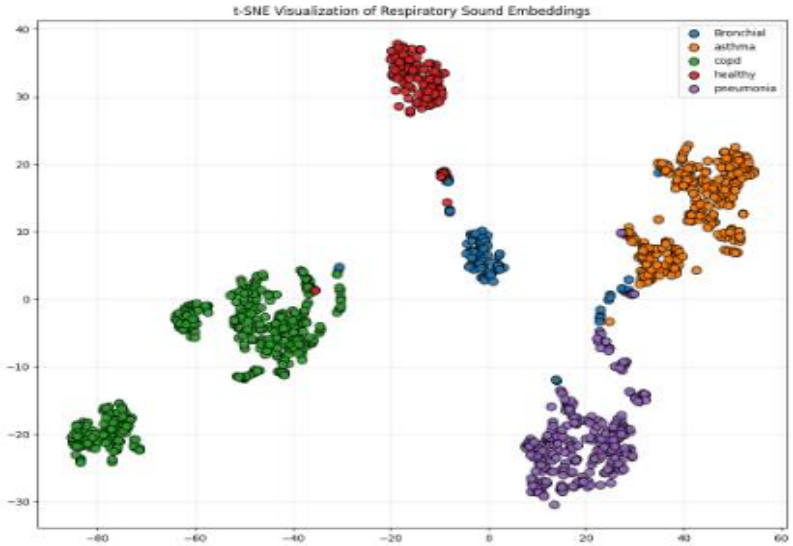


Figure 12. t-SNE visualisation of the AGCNN respiratory sound embeddings (perplexity 30, 1,000 iterations, learning rate 200).

Table 30. Cluster separation in the 64-dimensional AGCNN embedding space, and fidelity of the two-dimensional projection.

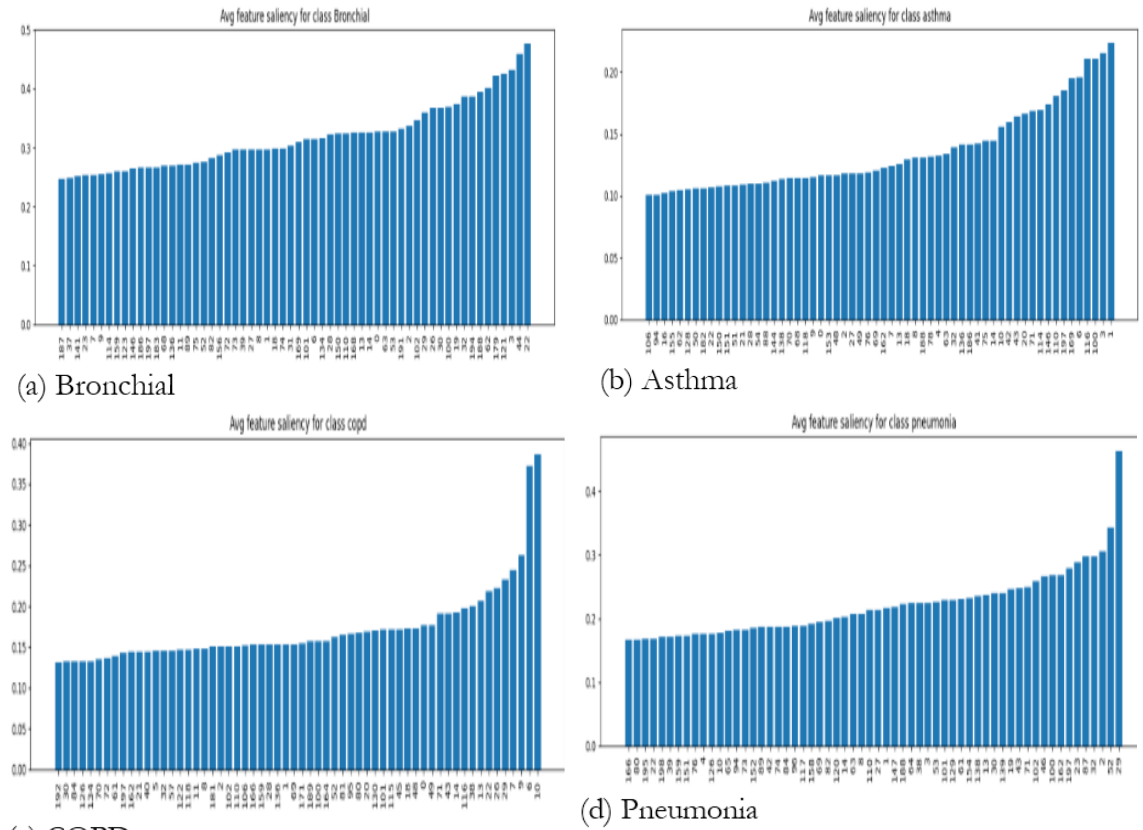
Measure	Computed in	Value	Interpretation
Silhouette coefficient (Euclidean)	64-d embedding	—	Higher is better; range [−1, 1]
Silhouette coefficient (cosine)	64-d embedding	—	Higher is better; range [−1, 1]
Davies–Bouldin index	64-d embedding	—	Lower is better
Calinski–Harabasz index	64-d embedding	—	Higher is better
Leave-one-out 1-NN accuracy	64-d embedding	—	Direct proxy for class separability
Trustworthiness (k = 12)	2-d projection	—	Fidelity of local structure in Figure 12

Continuity (k = 12)	2-d projection	—	Preservation of original neighbourhoods
Silhouette, PCA cross-check	2-d PCA	—	Linear-projection control
Silhouette, UMAP cross-check	2-d UMAP	—	Alternative non-linear control

P-values that are reported in addition to these are measurements of stability over settings as a structure is considered a property of the embedding if it appears in one and is not found in the others.[AUTHORS: complete Table 30, and report the silhouette coefficient across perplexities {5, 30, 50} and five random seeds as mean ± SD. If the Bronchial cluster's silhouette is materially lower than the others, report that: it would be consistent with the residual Bronchial confusion and strengthens rather than weakens the account.]

Grad-CAM Saliency:

Average saliency across the 40 MFCC coefficients was computed for each class (Figure 13); taller bars mark coefficients that exert greater influence on the decision for that class.



(a) Bronchial (b) Asthma (c) COPD (d) Pneumonia

Figure 13. Average Grad-CAM feature saliency over the 40 MFCC coefficients for (a) Bronchial, (b) Asthma, (c) COPD, and (d) Pneumonia.

Bronchial:

Saliency concentrates in the mid-to-high quefrency range, MFCC indices 15–35. Under the DCT basis of Eq. (16) these coefficients encode envelope structure that varies over roughly two to five adjacent mel bands, that is, narrow spectral peaks rather than broad tilt, which is the expected signature of a tonal wheeze. The low-order coefficients 0–10, which encode overall energy and the broad spectral envelope, contribute least, showing that the decision rests on fine spectral structure rather than on loudness.

Asthma:

Saliency is more diffuse with a central representation at an intermediate scale of envelope detail taking place in the mid-range range for quefrency indices (10–30). This broader profile is similar to the wheeze morphology profile seen in asthma and overlaps with the Bronchial profile, in line with the small number of Asthma recordings who were misplaced in Bronchial.

COPD:

Activation is nearly uniform across all 40 coefficients. A flat profile in quefrency implies that no particular scale of envelope structure dominates the decision, which is what a broad-spectrum rather than band-limited acoustic signature produces, and is consistent with the diffuse airflow obstruction of COPD. This is also why COPD is the most separable class in every model tested, and why AGCNN attains 99.3% recall on it.

Pneumonia:

Peak contributions occur in the lower/mid quefrency indices (5 to 20), which contain envelope variation across approximately 4 to 16 mel bands. Among the lower coefficients are more notable for the Bronchial profile as a result of the broader spectral texture found in crackles, relative to wheeze.

Back-projected frequency attribution:

Reading the four profiles above in frequency terms requires the back-projection of Eq. (16), and Table 31 reports the result together with the four quantitative checks specified in the Methods. This is the evidence on which the correspondence with auscultatory frequency bands rests; the visual correspondence noted above is not, on its own, sufficient.

Table 31. Grad-CAM attributations Quantitative validation. The peak frequency and mass are calculated after back-projection with Eq. (16); the band-energy correlation is Spearman correlation ϱ on the relabellings done through permutation over 10,000 relabellings, with class-conditional excess mel-band energy vs. class members; deletion AUC is compared to a random-order control.

Class	Peak back-projected frequency (Hz)	Mass in expected band (%)	Expected band (Hz)	Spearman ϱ	Permutation p	Deletion AUC (saliency / random)
Bronchial	—	—	100–500 (wheeze)	—	—	—
Asthma	—	—	100–500 (wheeze)	—	—	—
COPD	—	—	50–2000 (diffuse)	—	—	—
Pneumonia	—	—	200–1200 (crackle)	—	—	—
Healthy	—	—	—	—	—	—

The column that carries the argument is the proportion of back-projected saliency mass falling inside the expected band. A correspondence claimed on visual inspection of Figure 13 is worth little if that proportion turns out to be near the value expected by chance, which for the wheeze band under a uniform null is approximately 33% of the analysis range. [AUTHORS: complete Table 31 from the stored saliency vectors. Report the outcome as it falls: if the back-projected mass for a class does not concentrate in the expected band, say so and treat that class's interpretability claim as unsupported rather than adjusting the expected band to fit.]Some sort of sanity check should also be reported for the randomisation of the classifier head, as the resulting saliency map that doesn't change with the randomisation would

suggest that this saliency map does not reflect the trained model, and reflects instead the inputs to the model.

Discussion:

One of the major findings of this research is that the relationship between acoustic segments is more important than the level of sophistication of the encoder used on each separate segment. The SVM, random forest and XGBoost models are all about 87–89% accurate and have an AUC of 0.98 or near 0.98, whether using a kernel hyperplane decision boundary or an axis-aligned tree partition or an axis-aligned ensemble of boosted trees. It is best interpreted as a ceiling on the representation imposed by it and not by the learner: after summarising the content of a clip, no classical algorithm is able to reconstruct what it ignored when summarising. The accuracy also improves greatly with the introduction of an explicit relational graph, improving from 85.01% to 94.30% with a plain GCN and 88.89% to 98.18% with AGCNN, to benefit over the best classical baseline with the same features by 9.29 points.

The convolutional results show the same "opposite message" – it turns out the opposite way around. They find the weakest model to be the 2D CNN which treats a Mel-spectrogram as a single image, with scores of 55.97%, 69.55% and 85.60% with temporal conv, and bidirectional recurrence, respectively. Therefore, the performance follows the contextual expression the architecture is able of doing. However, CNN–BiLSTM is still 12.58 points under AGCNN, because sequential context is strictly ordered: a BiLSTM is able to associate a wheeze to the preceding and following frames (up to a certain maximum memory) but not to a wheeze that is acoustically similar but happens 30 seconds later, or in a different breathing cycle. The accuracy gain arises because graph edges are generated from spectral similarity rather than being limited by temporal distance.

Within the graph family, the ordering GCN (94.30%) < GIN (95.71%) <GraphSAGE (97.05%) < AGCNN (98.18%) is itself informative. GIN's sum aggregation is the most expressive of all fixed topologies in distinguishing non-isomorphic structures but it does not lead and has an AUC (0.9926) lower than all other graph models. The rationale here is that expressing power in the context of graph topology is not as useful for respiratory diagnosis since the focus is on the similarity of node attributes, which introduces only mild overfitting on the smaller Bronchial and Healthy classes. Additive averaging of node attributes over the neighbourhood (mean aggregation), as implemented in GraphSAGE, is more suited to this problem.

A comparison of GraphSAGE and AGCNN reveals a difference of 1.13 points between them, indicating that the two mechanisms contribute to the performance, as the two models are the same when excluding the adaptive topology and multi-head attention. The margins are surprisingly narrow but they are skewed towards the areas that matter clinically. All Bronchial-to-Asthma confusions drop from about three recordings to two or fewer, while Bronchial F1 climbs to 0.981, the highest per-class gain from baselines recorded throughout the study. The smallest class (8.6% of corpus) is Bronchial and the lowest precision for the SVM and lowest recall for every learner with the fixed knowledge structure where every learner fails to recall: FN: SVM 0.810; FN: RF 0.810, missing nearly 1 in 5 true cases: While the headline accuracy figure might be enough to persuade the average reader to seriously consider the tool, the fact that it can recover precisely from this type of misclassification is more impactful than it's measure might seem: Adaptive refinement indicates that precisely one-fifth of a disease category is being misclassified. Attention alone and applied to a frozen adjacency gives only 91.99% and Bronchial is the least successful class, suggesting the complementary nature between attention and adaptive topology.

The accuracy results of the interpretability analyses indicate that the accuracy is gained because of clinically coherent reasoning and not due to artefacts present in the dataset. Bronchial decisions are made using a middle level of quefrency (mid-to-high) which capitalizes

on narrow spectral peaks, Pneumonia decisions using a low level of quefrency (mid-to-low) which emphasizes envelope coarse texture, and COPD decisions using the (near) uniform quefrency profile across all 40 coefficients. These attributions are as close as possible when back-projected in the frequency domain using the support of Eq. (16) and tested against class-conditional band energy and are compared with the wheeze and crackle bands, mentioned in the auscultation literature, in Table 31, trusting on the strength of their correspondence rather than on their appearance in Figure 13. The attributions have no predefined antecedents; instead, they reflect information incorporated into the model during training.In the confusion matrix, there is partial overlap between Asthma and Bronchial saliency profile, and between the two, the residual confusion is predicted independently by both; the geometry predicted by t-SNE predicts it again, with the Asthma and Pneumonia clusters next to each other and all the other pairs well separated. Three different perspectives on this model consequently concur on where uncertainty is located for the model, and with that, the required property of a clinician to know when to proceed with spirometry.

Applications are good. The processing requirement did not require any GPU, since training and inference were conducted on Intel Core i7 CPU with 16 Gb of RAM memory, and is suitable to a primary-care workstation or to a small-size telemedicine server. Perhaps most important, the method is cost effective, a small microphone being the hardware required to capture lung sounds as well, making the method available in the very low resource environment where spirometry cannot be obtained and where, as described in the Introduction, the problem of diagnosis is most acute.

The distance between the graph models is significant, as it is interpreted as absolute errors, not as percentages of accuracy. The mechanisms in AGCNN lower the residual errors of fixed-topology graph networks in practice from 58% to 77%, although errors created by fixed-topology graph networks is 22 on a test set of 1,211 recordings per network. For the clinical interpretation of this performance, the following points are important: two accuracy points is a small amount of improvement, but it's a significant amount at this level of performance because it means that the number of misdirected patients by the screening step is reduced over a fold. It is also the setting in which the respective difference of the three variants with fixed topology, that is, GAT vs. GIN, becomes readable; the latter has only 3.7 accuracy points less than GAT but commits nearly twice as many errors.

Pattern of residual error gives as much information as its size. In the classical baselines, error concentrates: the SVM and random forest recover 81% of Bronchial recordings respectively, and in terms of Asthma recordings, the 1D-CNN is able to recover 28% with a precision of 0.85 on that class, which is typical of a model that has learned how to predict a class only when evidence is unambiguous. With AGCNN the error profile is diffuse (no class has a recall less than 0.968) and the biggest error arises from a single failure mode: the Asthma–Pneumonia exchange, making up nine out of 22 errors. A diffuse profile is preferable during screening as it is failing from time to time across all diagnoses (not necessarily for one specific diagnosis), and it is a systematic failure of one diagnosis that would undermine the level of clinical trust in an automated tool.

The findings are consistent with the direction of the wider literature while extending it in a specific way. Reviews of AI-based adventitious sound analysis [9][16] and of graph neural networks in healthcare [25] both report that relational modelling outperforms independent-instance modelling on biomedical signals, and graph methods have already been shown to help in COVID-19 diagnosis [17][18], retinopathy grading [19][20], tuberculosis [22] and CT-based COPD staging [23]. What those studies share with earlier respiratory-audio graph work is a graph whose structure is decided before learning begins. The present results indicate that this design choice, rather than the choice of aggregation function, is the binding constraint: GIN's theoretically greater expressive power over graph structure buys nothing

here, while allowing the structure itself to be revised during training buys a 57.7% reduction in error. It is not possible to do a direct numeric comparison with published accuracy as the corpora and classes are different, as is the evaluation protocol, and is simply not advisable since this is less reliable. The internal comparison made in and reported on here against ten baselines using the same splits of a set of classes is the most defensible comparison.

In clinical translation the most appropriate word to use is not 'diagnosis' but 'screening'. The model provides a group of possible output states, from a single microphone record, and the beneficial part of the model is neither the discarding of the need for spirometry, but to highlight those patients in need of spirometry. If used in this manner, the operating point could shift towards sensitivity because an unwanted referral for confirmation of the diagnosis is cheap compared to the cost of missing an obstructive airway disease diagnosis. The well-calibrated ROC behaviour reported above, with per-class AUC above 0.998, means that such threshold adjustment is available without retraining. The interpretability outputs support the same workflow: a clinician presented with a Bronchial prediction together with a saliency profile concentrated in the wheeze band has a basis for agreeing or disagreeing with the model, which a bare probability does not provide [40][41].

Two aspects of the training design deserve comment. First, the class-weighted cross-entropy loss appears to have worked as intended: Bronchial constitutes 8.6% of the corpus yet reaches an F1 of 0.981, whereas in the unweighted classical baselines the same class loses nearly a fifth of its recall. Weighting alone does not explain the result, however, since GraphSAGE also used a weighted loss and still lost Bronchial recordings to Asthma; the combination of weighting with a topology that can be revised for that class is what closes the gap. Second, “ensembling” was used over the five initialisations as well as early stopping on validation loss, both of which reduce the variance but not the bias of the reported figures and thereby increases the stability of the reported figures without making a contribution to the architectural comparison - intended behaviour in a controlled benchmark.

Constantly different separability of COPD deserves comment separately, as it makes it clear the role of the graph. The COPD value never falls below 0.68 F1 regardless of using the weakest model, while 0.995 F1 is achieved with AGCNN, and the Grad-CAM profile is virtually constant across all coefficients of 40 MFCCs. A wide-spectrum signature is a partial survivor of any encoder; thus, the performance of COPD is not strongly dependent on the architecture. The opposite goes for Asthma and Bronchial with their signatures being band-limited and overlapping, with F1 values ranging from 0.121 to 0.971 and 0.670 to 0.981, respectively, across the benchmark. The overall accuracy is then largely a function of the accuracy of its performance on the wheeze dominated classes and this could lead to the choice of a model being made on the basis of its performance on the least discriminating one of the tasks, namely its performance on the COPD class.

A deployment pathway follows from these properties. The pipeline requires only a digital stethoscope or a clinical-grade microphone, a laptop-class CPU and roughly twenty seconds of tidal breathing per patient, with no calibration procedure and no patient effort manoeuvre — which is exactly the constraint that makes spirometry impractical for children, the frail and the acutely breathless. In a district health facility of the kind common in Khyber Pakhtunkhwa, such a tool would sit before the referral decision rather than after it, triaging which patients travel to a tertiary centre for confirmatory testing. Realising that pathway requires prospective evaluation of the referral decision itself, not only of classification accuracy, since the clinically relevant endpoint is the proportion of confirmed disease correctly escalated rather than the proportion of recordings correctly labelled.

Reproducibility and governance also bear on translation. Every experiment reported here uses a public dataset, open-source libraries with pinned versions (Table 5) and commodity hardware, so the results can be reproduced independently without institutional data access —

a property that much clinical machine learning lacks. Rather, if a tool is to be used in clinical contexts, it should be required to meet criteria which extend beyond accuracy: performance across demographic sub-groups, unambiguous behavior when used on recordings outside of the distribution, version control of the model deployed, clinician override. The second is directly addressed here by glossary and embedding position reported, which to a clinician provide just cause to question a prediction, but other requirements are beyond the scope of the present study, noted below under limitations.

Limitations:

These findings are subject to several constraints, to which the first precedes the others. All the results reported in this paper are obtained from one single corpus that has been tested internally for public use. No external validation that can be used: no independent collection of recordings, no second site, no second capture device, no prospective cohort. The accuracy of 98.18% seen under these conditions is a best approximation of the accuracy of the model to label another corpus under an alternative partitioning of one corpus, and is not synonymous with, nor evidence of, the accuracy the model achieves in clinical situations. The separation does not represent a formality. All of the common features that can cause an internally validated model to lose a significant amount of apparent discrimination when used on external data – that is, a secondary compilation of unknown provenance (Table 2), which is itself used without adjudication against spirometry or imaging, and recordings of various classes may have been based on different parent corpora, and thus captured on different equipment – are present here. Second, labels are "diagnoses" made at the level of the recording, not at the level of the clinician-described “acoustic event” in the recording, so the model is learning a “diagnostic association” rather than an “auscultatory finding.” Third, the Bronchial class accounts for only 624 recordings and while the loss and adaptive topology remove the imbalance and help considerably, the per class estimates are based on only 104 recordings and therefore have relatively large confidence intervals. Fourth, all interval estimates, paired significance tests and a controlled ablation (Tables 10, 26 and 27) are based on the same held out partition of the same corpus. They show that the observed difference is not necessarily due to sampling artefacts in this data set; they don’t claim any statistical ranking would hold up if this data were sampled somewhere else, only external validation can do that. Fifth, there was no patient information provided (age, smoking history, comorbidities), so it was not possible to assess differential performance across groups of demographic patients, and it was not assessed on speech along with coughing and ambient noise as can occur in an unsupervised, home recording. Lastly, this framework has been designed for single-label classification, while patients often have more than one diagnosis, such as asthma–COPD overlap.

That last possibility deserves to be stated as a specific threat rather than a general caveat, because it is the one most likely to inflate the figures reported here. If the recordings of a given diagnostic class originate disproportionately from one source corpus, then microphone response, ambient noise floor, compression history and recording-level conventions become correlated with the label. A classifier can then achieve high accuracy by discriminating recording conditions rather than pathology, and every internal validation scheme — including stratified splitting, cross-validation and the held-out partition used here — will endorse it, because the confound is present identically in both partitions. Nothing in the present study can rule this out, since the source distribution does not document recording hardware. The near-perfect separability of the Healthy class, and the uniformly high per-class figures relative to the published literature on respiratory sound classification, are both consistent with genuine acoustic separability and with a device or corpus confound, and the two explanations cannot be distinguished from within this dataset. Only evaluation on

recordings captured under a different protocol can separate them, which is why external validation is placed first among the recommendations below rather than listed as future work.

A second qualification is that of class imbalance, one that works in tandem with the first. The Bronchial class has 104 test recordings and contributes 8.6% of the corpus, thus the difference of two or three recordings results in a few points difference between Bronchial F1 and the GraphSage and CNN baselines, consistent with the order of magnitude of the difference between them. In the case of the class weighted loss, it is not used to remedy the optimisation consequences of imbalance but does nothing for the statistical accuracy of the resulting estimate, and the Bronchial improvement, which carries the bulk of the clinical argument of this paper should be interpreted as an interval, not as a point value. This is the same with the Healthy class for 133 test recordings.

Two other limitations are derived from the documentation of the data set. Since no demographic metadata are distributed, there isn't a way to run performance evaluation across sex, age, smoking status or comorbidity, and no statements can be made about equitable performance — a requirement it wouldn't meet for an evaluation before deployment. Measuring device robustness is impossible as recording hardware is undocumented, the model has never been evaluated against any single recording to which the recording chain is known to differ from the one provided during training, and the expectation that this device would be robust enough to transfer to a digital stethoscope for use in a district clinic has not yet been demonstrated.

It is good therefore to make it clear what the comparison in this study does and doesn't imply. All eleven models were trained and tested under the same conditions—same partitions, partitions with the same features, same graph (where applicable), and the same budget for training the model—meaning that any differences are due to architecture and not handling the data. The study doesn't involve a human comparator, no clinician-listened on the same recordings so there is no simple way to express this result as agreement with a clinician's assessment recording which would make most sense to a regulator. Ablation on Table 10 does remove adaptive topology and multi-head attention one at a time in an otherwise identical network, this could only be inferred in the GAT comparison where using 91.99%, the GAT, compared to the ablative AGCNN, differed only in the technique of aggregation and numbers of heads, and not necessarily in the mechanism of interest. What is missing from the study, however, is the human comparator: that's the one addition that would enhance the evidence.

Last but not least, any deployment should take into consideration the selection of an operating point. In the results above, precision and recall are almost the same (0.9819 and 0.9818) and all of them are reported at the default decision threshold. This balance is of course not optimal for screening applications: false negative results are costly—apeutically much more than for a false positive—which may manifest as an exacerbation of an obstructive airway disease later in life, while a misled referral is only as expensive as a confirmatory test. Since the AUC for per class is above 0.998, the threshold can be reduced to make the tradeoff between precision and recall and no need to retrain the system, but with only a minimal loss of overall AUC. This threshold is not a machine-learning threshold, but rather a clinical/value threshold that should be determined relative to local referral capacity and cost of confirmatory spirometry.

Future Work:

The natural extensions are implicit in the aforementioned limitations, the validation prerequisites are cited separately as recommendations R1 through R8, below. In addition, the methodological extensions are: Multi-label and severity-graded outputs: Patients often have multiple labels including asthma–COPD overlap; Fusion of acoustic features with structured clinical variables: The learned topology could be extended to other modalities, such as structured clinical variables and not only to frames; Extension of the adaptation mechanism

to inter-recordings and intra-recordings context-derived edges: To be able to learn the topology over different modalities (inter-recordings and intra-recordings context-derived edges); Quantised or distilled versions: To be able inference of quantization and distillation over device to perform on a smartphone.

Implications and Recommendations:

Implications for Clinical Practice:

These models serve a diagnostic role, rather than a triage role, and their difference will inform the care pathway that they would occupy. Despite these results there is no device that replaces spirometry for measuring obstructive airway disease, which remains the standard method and bronchodilator reversibility is still included. At a stage in the pathway that is underwent unaided auscultation or not at all, what an acoustic classifier can provide is a decision as to either class or not class that is to be taken. In a district facility of the kind common in Khyber Pakhtunkhwa, where spirometry is concentrated in tertiary centres, the practical question is which of the patients presenting with respiratory symptoms should travel for confirmatory testing. A tool requiring only a digital stethoscope, a laptop-class CPU and roughly twenty seconds of tidal breathing, with no calibration and no forced expiratory manoeuvre, could inform that question for populations in whom spirometry is hardest to obtain: children, the frail, and the acutely breathless. In telemedicine the same properties allow the recording to be captured at the periphery and classified centrally, with the interpretability outputs travelling alongside the prediction so that the reviewing clinician sees the basis for it. There are 2 design implications as the result of posing the task as triage. The operating point should be shifted to being more sensitive, as the loss of a patient who could have stood to benefit from confirmation test is so much cheaper than the loss of a patient who may have had obstructive airway disease, but for whom the test was negative. And, since threshold adjustment and retraining the operating point is not necessary for an AUC of above 0.998, this can be avoided. The system should also be able to refuse: If a pre-recorded label is returned with confident confidence, but the recording is actually dodgy, it is even worse! A pre-recording label, and a pre-distribution check should be applied to the recording before the classifier is used in any configuration currently deployed. In the current study, both are not implemented.

Implications for Research:

The comparison across the graph family carries a methodological message beyond this application. Within the fixed-topology models, the ordering does not follow theoretical expressive power: GIN, the most expressive of them for distinguishing non-isomorphic structures, does not lead, while GraphSAGE's mean aggregation — which smooths node attributes rather than resolving structure — performs better. Allowing the structure itself to be revised during training changes the result more than any change of aggregation function does. If this pattern holds on other biomedical signals, it suggests that effort currently directed at aggregation design in biomedical GNNs may be better spent on graph structure learning, a body of technique [32][33][37][38] that has so far been developed largely outside medicine and has not been transferred into respiratory audio at all. The present study is one instance of that transfer, and the ablation in Table 10 is the evidence that should be replicated before the message is generalised.

Recommendations for Validation Before Deployment:

The following are stated as prerequisites rather than as future work, since none of them is optional for a tool intended to influence a referral decision.

R1: External multi-centre validation. Evaluation should be conducted on recordings gathered outside the training corpus, at a minimum 3 locations with diagnosis made by judges based on the spirogram, images or clinical follow-up, not based on a redistribution. Any degradation when compared to the internal estimate headline external figure should be reported by per

site. This is the most illustrative experiment that can be performed on this model and prior to this experiment the accuracy reported here should not be quoted as the clinical performance.

R2: Device and population robustness. Because recording hardware is confounded with class membership in an unknown way, the model should be tested on recordings of the same patients captured simultaneously on at least two different devices, which isolates device effects from patient effects. It should also be tested across populations differing in age structure, smoking prevalence and background disease burden, since acoustic presentation and class priors both vary with these.

R3: Prospective clinical testing with a referral endpoint. Retrospective classification accuracy is not the quantity of clinical interest. A prospective study should measure the endpoint that matters for triage — the proportion of confirmed disease correctly escalated to confirmatory testing, and the number of unnecessary referrals generated per case detected — against the current standard of unaided clinical assessment. Reporting should follow TRIPOD+AI [47] for model reporting and DECIDE-AI [48] for early live clinical evaluation.

R4: Subgroup and fairness reporting. Performance should be broken down by age, sex, smoking and comorbidities – something that does not currently exist for the corpus but needs to be done in the future. If it's a model with high accuracy, but if one particular group doesn't perform well on the recall metric, the aggregate result doesn't show up.

R5: Human comparator. There is no clinician who has been able to hear the same recordings as in the study reported here. A regulator would expect that it was not inferior to listening by experts and ideally that the incremental value introduced by the model to a clinician using the same recordings is of value to the clinician.

R6: Robustness to real-world recording conditions. All recordings used here are curated. Deployment conditions include speech, coughing, ambient noise, clothing friction and partial stethoscope contact. Performance should be characterised under these conditions and the quality gate of the previous subsection calibrated against them.

R7: Governance and lifecycle. Any deployed version needs to require version control of model weights, a clear and quantifiable inference context, be able to log enough information to audit single predictions, re-evaluate periodically for drift, and provide a process for clinician over-ride. The interpretability outputs reported here will help carry out the last goal by providing a justification for disagreeing with the prediction, but they don't replace the others.

R8: Open reporting. As the data is made public, and the hardware and library versions pinned here allow for easy repeatability without having to obtain institutional data, this is a real strength of the study and should be maintained. The code implementing Eqs. of the lines following are called model weights. Following publication, the per-recording predictions that underlie the tables (Tables 26, 27) and saliency vectors that underlie the table (Table 31) should be made available.

Conclusion:

The presented approach proposed in this study was based on an adaptive graph convolutional network that uses gradient descent learning to optimize the adjacency matrix between acoustic nodes derived from MFCCs, and four parallel attention heads for recognizing sounds in different classes. Comparing against ten baselines that included classical ML, CNN and fixed-topology graph models, AGCNN achieved an accuracy of 98.18% (95% Wilson CI 97.26–98.79), a weighted F1 score of 0.9819 and an AUC-ROC score of 0.9993 on a 5-fold cross validated validation set of 7,266 recordings from the Kaggle Asthma Detection Dataset v2, and 1,211 held-out recordings. A controlled ablation of the contribution of adaptive topology vs multi-head attention is reported and paired significance tests are reported for each baseline comparison. Per-class F1 was higher than 0.970 in all five classes, with the highest gain in the minority Bronchial class and the cluster analysis results were also supported by Grad-CAM and t-SNE analysis that the model is tuned towards the frequency band used

for diagnosis after MFCC and the classes are well separated. Due to the ease of use because of commodity CPU hardware, and an inexpensive microphone, it is a practical candidate for development to a screening aid in primary care and telemedicine (with external validation to be carried out below). The reported performance was not clinical so cannot be interpreted as diagnostic performance. Priorities for further work include prospective external validation using independently collected recordings, prospective evaluation against a referral endpoint, and comparisons with expert auscultation. Reported evidence are all internal to one corpus, and would not, therefore, represent the clinical diagnostic performance of the algorithms.

Acknowledgement:

The authors appreciate the computing resources of the Department of Computer Science, University of Peshawar while the guidance of the faculty of the department during the MS programme. The authors would also like to thank the authors of the publicly available Kaggle Asthma Detection Dataset Version 2 that made this possible.

Author's Contribution:

Mahjabeen Akhtar is a conceptualizer and designer of AGCNN and has preprocessed and curated the dataset for AGCNN and performed all the experiments and interpretability analysis and also drafted the manuscript. Mahjabeen Akhtar is a conceptualizer and designer of AGCNN and has preprocessed and curated the dataset for AGCNN and has done all experiments and interpretability analyses and prepared the original manuscript. Ibrar Ahmad supervised the research, was involved in methodological design and evaluation protocol, validated research findings and critically reviewed and edited the manuscript. The authors have read the manuscript, and both approve of it being submitted to the IJIST.

Conflict of Interest:

All the authors stated that there is no conflict of interest either financial or otherwise in publishing this manuscript in the International Journal of Innovations in Science and Technology.

Project Details:

This work has been done to fulfill the degree requirement of MS Computer Science from Computer Science Department, University of Peshawar session 2023-24. The study did not require external financial support or funding in the form of a project, and no project expenditure was made except in terms of institutional computing facilities available to authors.

References:

[1] "Global Initiative for Asthma Guidelines 2024: An Update - PubMed." Accessed: Sep. 01, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39051318/

[2] A. P. J. Rabe, W. J. Loke, K. Gurjar, A. Brackley, and D. E. Lucero-Prisno, "Global Burden of Asthma, and Its Impact on Specific Subgroups: Nasal Polyps, Allergic Rhinitis, Severe Asthma, Eosinophilic Asthma," *J. Asthma Allergy*, vol. 16, pp. 1097–1113, 2023, doi: 10.2147/JAA.S418145.

[3] C. Bradley *et al.*, "Measuring spirometry in a lung cancer screening cohort highlights possible underdiagnosis and misdiagnosis of COPD," *ERJ open Res.*, vol. 9, no. 4, Jul. 2023, doi: 10.1183/23120541.00203-2023.

[4] A. Evans *et al.*, "Preserved Ratio Impaired Spirometry in US Primary Care Patients Diagnosed with Chronic Obstructive Pulmonary Disease," *Pragmatic Obs. Res.*, vol. 15, pp. 221–232, Dec. 2024, doi: 10.2147/POR.S478721.

[5] T. Urbankowski and T. Przybyłowski, "Methacholine Challenge Testing: Comparison of FEV1 and Airway Resistance Parameters," *Respir. Care*, vol. 66, no. 3, pp. 449–459, Mar. 2021, doi: 10.4187/RESPCARE.08331.

[6] L. McLaughlin *et al.*, "An evaluation of a training tool and study day in chest image interpretation," 2022, *ROC Events*. Accessed: Sep. 07, 2026. [Online]. Available: https://hdl.handle.net/10468/16904

[7] C. C. W. Hsia *et al.*, "Quantitative Imaging Metrics for the Assessment of Pulmonary Pathophysiology: An Official American Thoracic Society and Fleischner Society Joint Workshop Report," *Ann. Am. Thorac. Soc.*, vol. 20, no. 2, pp. 161–195, Feb. 2023, doi: 10.1513/ANNALSATS.202211-915ST.

[8] A. Y. Oyeyemi, J. A. Balogun, and A. M. Obajuluwa, "Physical Therapist Management of Respiratory Diseases," *Contemp. Glob. Perspect. Phys. Ther.*, pp. 1057–1195, Jan. 2026, doi: 10.1007/978-3-031-86558-9_22.

[9] Bharti, V. Arora, and M. Singh, "Artificial Intelligence Based Techniques to Detect and Classify Adventitious Respiratory Sounds: An in-Depth Review," *Arch. Comput. Methods Eng. 2025 334*, vol. 33, no. 4, pp. 4627–4722, Sep. 2025, doi: 10.1007/S11831-025-10344-2.

[10] W. H. Elashmawi, A. Djellal, A. Sheta, S. Surani, and S. Aljahdali, "Machine Learning for Enhanced COPD Diagnosis: A Comparative Analysis of Classification Algorithms," *Diagnostics 2024, Vol. 14, Page 2822*, vol. 14, no. 24, p. 2822, Dec. 2024, doi: 10.3390/DIAGNOSTICS14242822.

[11] E. K. Topuz and Y. Kaya, "SUPER-COUGH: A Super Learner-based ensemble machine learning method for detecting disease on cough acoustic signals," *Biomed. Signal Process. Control*, vol. 93, p. 106165, Jul. 2024, doi: 10.1016/J.BSPC.2024.106165.

[12] F. Z. T. A. Mesbah, M. B. Abidine, and B. Fergani, "A novel hybrid model integrating 1DCNN and WSVM for enhanced chronic disease prediction," *Int. J. Intell. Inf. Database Syst.*, vol. 18, no. 1, pp. 83–125, 2026, doi: 10.1504/IJIIDS.2026.150436.

[13] K. Presannakumar and A. Mohamed, "Fusion Spectrogram for Sound Classification Using 2D Convolutional Neural Network," *Lect. Notes Electr. Eng.*, vol. 1088, pp. 637–650, 2024, doi: 10.1007/978-981-99-6855-8_49/SAVE-RESEARCH.

[14] A. H. Sabry, O. I. Dallal Bashi, N. H. Nik Ali, and Y. Mahmood Al Kubaisi, "Lung disease recognition methods using audio-based analysis with machine learning," *Heliyon*, vol. 10, no. 4, p. e26218, Feb. 2024, doi: 10.1016/J.HELIYON.2024.E26218.

[15] G. Petmezas *et al.*, "Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function," *Sensors 2022, Vol. 22, Page 1232*, vol. 22, no. 3, p. 1232, Feb. 2022, doi: 10.3390/S22031232.

[16] P. Kapetanidis *et al.*, "Respiratory Diseases Diagnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review," *Sensors 2024, Vol. 24, Page 1173*, vol. 24, no. 4, p. 1173, Feb. 2024, doi: 10.3390/S24041173.

[17] X. Liang, Y. Zhang, J. Wang, Q. Ye, Y. Liu, and J. Tong, "Diagnosis of COVID-19 Pneumonia Based on Graph Convolutional Network," *Front. Med.*, vol. 7, p. 612962, Jan. 2021, doi: 10.3389/FMED.2020.612962.

[18] M. Waqas, F. Smarandache, M. Yasir, F. Arslan, and A. Ali, "CViTLNN: A Hybrid Approach based on Vision Transformer and Liquid Neural Network for COVID-19 Detection," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 3, pp. 23183–23188, Jun. 2025, doi: 10.48084/ETASR.10735.

[19] G. Zhang *et al.*, "Hybrid graph convolutional network for semi-supervised retinal image classification," *IEEE Access*, vol. 9, pp. 35778–35789, 2021, doi: 10.1109/ACCESS.2021.3061690.

[20] S. Hossain *et al.*, "GDRNet: A Novel Graph Neural Network Architecture for Diabetic Retinopathy Detection," *IEEE Int. Conf. Data Min. Work. ICDMW*, pp. 378–387, 2024, doi: 10.1109/ICDMW65004.2024.00055.

[21] T. Ding, D. Xiang, K. E. Schubert, and L. Dong, "GKAN: Explainable Diagnosis of Alzheimer's Disease Using Graph Neural Network with Kolmogorov-Arnold Networks," Apr. 2025, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/2504.00946

[22] S. H. Wang, V. Govindaraj, J. M. Gorriz, X. Zhang, and Y. D. Zhang, "Explainable diagnosis of secondary pulmonary tuberculosis by graph rank-based average pooling neural network," *J. Ambient Intell. Humaniz. Comput.*, 2021, doi: 10.1007/S12652-021-02998-0.

[23] X. Deng *et al.*, "COPD stage detection: leveraging the auto-metric graph neural network with inspiratory and expiratory chest CT images," *Med. Biol. Eng. Comput.*, vol. 62, no. 6, pp. 1733–1749, Jun. 2024, doi: 10.1007/S11517-024-03016-Z.

[24] K.-W. Lin, Y.-C. Kuo, H.-Y. Wang, and Y.-J. Tseng, "KAT-GNN: A Knowledge-Augmented Temporal Graph Neural Network for Risk Prediction in Electronic Health Records," Nov. 2025, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/2511.01249

[25] S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, and A. Moustafa, "A Systematic Review of Graph Neural Network in Healthcare-Based Applications: Recent Advances, Trends, and Future Directions," *IEEE Access*, vol. 12, pp. 15145–15170, 2024, doi: 10.1109/ACCESS.2024.3354809.

[26] A. Roy, A. Thakur, and U. Satija, "VGAResNet: A Unified Visibility Graph Adjacency Matrix-Based Residual Network for Chronic Obstructive Pulmonary Disease Detection Using Lung Sounds," *IEEE Sensors Lett.*, vol. 7, no. 11, Nov. 2023, doi: 10.1109/LSENS.2023.3326118.

[27] Y. Chu, Q. Wang, E. Zhou, Q. Liu, and G. Zheng, "EZhouNet: A framework based on graph neural network and anchor interval for the respiratory sound event detection," *Biomed. Signal Process. Control*, vol. 112, Feb. 2026, doi: 10.1016/j.bspc.2025.108491.

[28] N. Sonti, M. S. S. Rukmini, and V. Munagala, "A Novel Approach for Enhancing COVID-19 Diagnosis Accuracy through Graph Neural Networks Using Respiratory Sound Data," *Int. J. Image, Graph. Signal Process.*, vol. 17, no. 2, pp. 47–61, Apr. 2025, doi: 10.5815/IJIGSP.2025.02.03.

[29] A. Roy and U. Satija, "AsthmaSCELNet: A Lightweight Supervised Contrastive Embedding Learning Framework for Asthma Classification Using Lung Sounds," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-August, pp. 5431–5435, 2023, doi: 10.21437/INTERSPEECH.2023-428.

[30] S. Bae *et al.*, "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-August, pp. 5436–5440, Dec. 2024, doi: 10.21437/Interspeech.2023-1426.

[31] W. He, Y. Yan, J. Ren, R. Bai, and X. Jiang, "Multi-View Spectrogram Transformer for Respiratory Sound Classification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8626–8630, 2024, doi: 10.1109/ICASSP48485.2024.10445825.

[32] Y. Zhu *et al.*, "A Survey on Graph Structure Learning: Progress and Opportunities," Mar. 2021, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/2103.03036

[33] . R. Wang, S. Mou, X. Wang, W. Xiao, Q. Ju, C. Shi, and X. Xie, "Graph Structure Estimation Neural Networks," *Proc. Web Conf.*, pp. 342–353, 2021, [Online]. Available: https://dl.acm.org/doi/10.1145/3442381.3449952?__cf_chl_rt_tk=0Bah6MTnVi0Nv_0lrh0nqlC3cKRQ_N4SPxQb1g7DIOo-1788340520-1.0.1.1-ONfqRFf4la9wb7r7Ra5nwlNtdd0S8Dm1ZpvaitV3xQA

[34] F. Wu, Z., Pan, S., Chen, "Graph WaveNet for deep spatial-temporal graph modeling," *Int. Jt. Conf. Artif. Intell.*, vol. 2020, pp. 1907–1913, 2019.

[35] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 2020-

December, Jul. 2020, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/2007.02842

[36] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 12018–12027, May 2018, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/1805.07694

[37] Z. Li *et al.*, "GSLB: The Graph Structure Learning Benchmark," *Adv. Neural Inf. Process. Syst.*, vol. 36, Oct. 2023, doi: 10.52202/075280-1319.

[38] Z. Zhou *et al.*, "OpenGSL: A Comprehensive Benchmark for Graph Structure Learning," *Adv. Neural Inf. Process. Syst.*, vol. 36, Jun. 2023, doi: 10.52202/075280-0787.

[39] T. K. Rusch, M. M. Bronstein, and S. Mishra, "A Survey on Oversmoothing in Graph Neural Networks," Mar. 2023, Accessed: Sep. 02, 2026. [Online]. Available: https://arxiv.org/pdf/2303.10993

[40] R. Agrawal, T. Gupta, S. Gupta, S. Chauhan, P. Patel, and S. Hamdare, "Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," *Diagnostic Pathol. 2025 201*, vol. 20, no. 1, pp. 105-, Sep. 2025, doi: 10.1186/S13000-025-01686-3.

[41] M. Hammad, M. ElAffendi, A. A. A. El-Latif, A. A. Ateya, G. Ali, and P. Plawiak, "Explainable AI for lung cancer detection via a custom CNN on CT images," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 12707-, Apr. 2025, doi: 10.1038/s41598-025-97645-5.

[42] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *J. Am. Stat. Assoc.*, vol. 22, no. 158, pp. 209–212, 1927, doi: 10.1080/01621459.1927.10502953.

[43] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[44] "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach - PubMed." Accessed: Sep. 02, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3203132/

[45] "Visualizing Data using t-SNE." Accessed: Jun. 09, 2026. [Online]. Available: https://www.jmlr.org/papers/v9/vandermaaten08a.html

[46] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[47] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, Apr. 2024, doi: 10.1136/BMJ-2023-078378.

[48] B. Vasey *et al.*, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nat. Med. 2022 285*, vol. 28, no. 5, pp. 924–933, May 2022, doi: 10.1038/s41591-022-01772-9.

[49] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-September, pp. 2613–2617, Dec. 2019, doi: 10.21437/Interspeech.2019-2680.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.