Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transformer Architectures for Respiratory Sound Analysis and Multimodal Diagnosis

Theodore Aptekarev

aptekath@bgu.ac.il

Ben-Gurion University of the Negev

Vladimir Sokolovsky

Ben-Gurion University of the Negev

Gregory Furman

Ben-Gurion University of the Negev

Evgeny Furman

Perm State Medical University

Research Article

Keywords: artificial neural networks, deep learning, transformer architectures, computer-assisted diagnostics, respiratory sound classification, bronchial asthma

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10848586/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Operations Research Forum on October 15th, 2026. See the published version at https://doi.org/10.1007/s43069-026-00698-z.

Transformer Architectures for Respiratory Sound Analysis and Multimodal Diagnosis

Theodore Aptekarev[1,3*], Vladimir Sokolovsky[1], Gregory Furman[1], Evgeny Furman[2]

[1*]Ben Gurion University of the Negev, Beer Sheva, Israel.
[2]E.A.Vagner Perm State Medical University, Perm, Russia.
[3]HSE University in Nizhny Novgorod, Nizhny Novgorod, Russia.

*Corresponding author(s). E-mail(s): aptekath@bgu.ac.il;
Contributing authors: sokolovv@bgu.ac.il; gregoryf@bgu.ac.il;
furmaneg@gmail.com;

Abstract

Breath sound analysis is one of the diagnostic approaches for various lung diseases. However, traditional auscultation is subjective and dependent on the physician's experience. Recent studies have demonstrated that computer-assisted methods analyzing these sounds can significantly improve the accuracy of diagnosis of the patient's condition. However, the absence of a common basis for fair comparison, limited use of patient metadata, and lack of interpretability make these approaches difficult to adopt in clinical practice.

In the paper, we propose to utilize the benefits of the recent state-of-the-art sound classification models, and a multimodal approach for respiratory disease screening and diagnostics. To achieve this goal we use the Audio Spectrogram Transformer (AST) model and a vision-language model (VLM) Moondream2. We compare them with our previous work on DenseNet201 Convolutional Neural Network (CNN). Proposed methodology allows a fair comparison of different computer assisted methods for respiratory disease screening and diagnostics. The training and evaluation of the models has been conducted using the same dataset containing respiratory sound recordings and patient information.

The obtained results demonstrate that these transformer-based models can serve as a basis for the development of computer-assisted methods for asthma diagnosis, as well as for continuous and remote monitoring of patient condition. In addition our approach enables interpretability of the model performance.

The achieved performance is much better than the best results of the physician's auscultation.

Keywords: artificial neural networks, deep learning, transformer architectures, computer-assisted diagnostics, respiratory sound classification, bronchial asthma

MSC Classification: 92C55 , 68T07 , 68T10

1 Introduction

Respiratory diseases represent a significant portion of global morbidity and are a common cause of mortality. Bronchial asthma, in particular, affects up to 29% of the population in various countries [6, 30], amounting to approximately 262 million people worldwide and causing over 460,000 deaths annually [50]. It consistently ranks among the top 30 conditions affecting all age groups [50]. Traditional diagnosis is based on a comprehensive examination that includes anamnesis, spirometry, auscultation, X-ray, and other tests. However, manual auscultation is subjective and strongly depends on the physician's experience [20, 31]. In addition, functional tests are difficult to perform in young children, and access to specialized care is often limited in remote areas. Studies suggest that up to 30% of asthma patients may receive an incorrect diagnosis when it is based solely on symptoms [47]. Continuous monitoring of the patient's condition, early prediction, and prevention of disease exacerbations (attacks) are important parts of the treatment and management. Computer-assisted methods that utilize recordings of respiratory sounds have already been applied to bronchial asthma diagnostics and have demonstrated high diagnostic performance [4, 16]. These methods can be used for both clinical diagnosis and remote monitoring of patients with asthma (telemedicine), as well as for near-real-time assessment of treatment effectiveness.

In early computer-based diagnostic methods, respiratory sounds were analyzed using signal-processing approaches, including spectral analysis, wavelet-based processing [8, 14, 32, 48], and other.

Subsequently, machine learning methods based on Convolutional Neural Networks (CNNs) have demonstrated considerable promise. Studies comparing various CNN architectures (e.g., VGG, ResNet, DenseNet) for respiratory pathology diagnostics reported accuracies ranging from 80% to over 90%, depending on the dataset and task [4, 16, 25, 36]. The clinical diagnosis of asthma is based on a comprehensive medical examination that includes the patient's anamnesis and other contextual information [18]. In our prior CNN pipeline, only the respiratory sound signal was used, and structured patient metadata were not incorporated into the classifier [4].

However, comparing the results of these studies is difficult due to the lack of standardization in datasets and recording protocols. As noted in the recent review [43], public datasets vary significantly in size, annotation quality (e.g., matching of respiratory sounds to clinical diagnosis), and heterogeneity in the sound recording systems. For example, the performance of models trained on cough-centric datasets, such as COUGHVID [35], cannot be directly compared to performance of models trained on records of breath sounds.

Recent reviews highlight a rapidly expanding set of respiratory sound datasets and machine learning (ML) approaches [43]. Xia et al. (2022) [52] analyzed 11 datasets

and reported that traditional ML methods typically achieve the diagnostic accuracies of 75-82%, while deep learning models trained on larger datasets and spectrogram representations of respiratory sound records reach the accuracy of 85-90%. During the COVID-19 pandemic, attention shifted toward cough-centered analyses, which increased the number of published datasets (e.g., COUGHVID and Coswara) but reduced the emphasis on creating breathing sound datasets [9, 35, 44]. Existing public datasets, such as ICBHI 2017 [42] and HF_Lung [22], often focus on detecting sound events such as crackles and wheezes rather than nosological diagnosis, or they are limited by specific recording equipment.

Recent studies have begun to apply transformer-based models to the analysis of respiratory sounds. Aljaddouh et al. (2024) [1] used the Vision Transformer (ViT) model for multi-class classification of respiratory spectrograms of patients with respiratory conditions and reported an accuracy of 91%. For multimodal integration, Tang et al. (2022) [46] combined cough audio with patient metadata using a hierarchical multimodal transformer for COVID-19 detection.

Despite these advances, research using transformers specifically focused on asthma diagnosis based on breath sound analysis remains limited, and the role of structured clinical context has not been systematically assessed.

In the paper, we propose to utilize the benefits of the recent state-of-the-art sound classification model and a multimodal approach for respiratory disease screening and diagnostics. To achieve this goal we use the Audio Spectrogram Transformer (AST) model[19], and a vision-language model (VLM) Moondream2[26]. The AST model is sound-only. The Moondream2 is a multimodal vision-language model that combines spectrograms with structured patient metadata. To ensure a fair comparison, we use a controlled internal baseline. In our previous work on the dataset [4], DenseNet201 [24] showed high sensitivity for respiratory sound classification. In the present study, we use that DenseNet201 result as a reference point and compare it with two transformer-based models on the same dataset.

We address two objectives: (i) adapt the Audio Spectrogram Transformer for respiratory sound classification in the "Asthma vs Not Asthma" task domain; and (ii) evaluate Moondream2 as a multimodal model that combines spectrograms with structured patient metadata (age, sex, recording site). The second approach is intended to mimic the clinical decision-making process in which diagnosis depends on analysis of both the breath sound signal and the patient's anamnesis.

For the "Asthma vs Not Asthma" task, Moondream2 demonstrates quantitative performance comparable to that of the DenseNet201 CNN baseline, with 86.5% accuracy, 96.0% sensitivity, 77.0% specificity, an F1-score of 87.7%, and a Youden index of 0.730. It was shown that the incorporation of structured patient metadata directly into the inference process using structured JavaScript Object Notation (JSON) outputs increases the Moondream2 quantitative performance. By training and evaluating the models on a supervised dataset, we demonstrated that AST clearly outperforms both Moondream2 and the established DenseNet201 CNN baseline. For the "Asthma vs Not Asthma" task, AST achieved 98.7% across accuracy, sensitivity, specificity, and F1-score, and a Youden Index of 0.974.

The obtained results demonstrate that these transformer-based models can serve as a basis for the development of computer-assisted methods for asthma diagnosis, as well as for continuous and remote monitoring of patient condition. In addition our approach enables interpretability of the model performance. The achieved performance is much better than the best results of the physician's auscultation.

2 Dataset

In this study, we used an anonymized dataset collected at the Regional Children's Clinical Hospital of Perm Krai (Perm, Russia), as described in previous work [4, 13, 15, 16]. The protocol was approved by the Ethics Committee of Perm State Medical University and followed the Declaration of Helsinki [7]. The full dataset includes records and information on 1371 participants (aged 0–47 years), including healthy volunteers and patients with confirmed respiratory diseases. Asthma was diagnosed according to GINA guidelines [18]. The dataset is the institutional property of Perm State Medical University and is available upon request from the designated institutional contact. Inclusion criteria for patients required a confirmed diagnosis of respiratory disease; for volunteers - the absence of acute respiratory infection at the time of recording. Informed consent was required for both patients and volunteers. Each audio recording is accompanied by clinical annotations and structured metadata (sex, age, recording location, diagnosis, recording date, and quality flags), which allows us to perform the multimodal experiments (diagnosis) described below. A more detailed description of the data collection protocol and metadata schema is available in [4].

To maintain compliance with regulatory requirements, access to the dataset is managed by Perm State Medical University. The dataset is provided to institutions upon request. Requests for access should be directed to Sergey Malinin at info@mstek.ru.

Respiratory sound recordings were performed during quiet breathing at four anatomical points (mouth, trachea, right second intercostal space, and right paravertebral area). Each recording lasted several breathing cycles, averaging about 25 seconds (minimum 16 seconds). Data were collected using various hardware ranging from specialized computerized systems with external microphones and electronic stethoscopes to mobile phones. These systems ensure amplitude-frequency linearity in the 100–3000 Hz range and have been validated in prior technical and clinical studies [13, 15, 16, 41]. Sampling rates varied from 22 kHz to 96 kHz. Original files were stored as WAV, MP3, or M4A and were harmonized to WAV during preprocessing. The dataset includes clinician-assigned quality labels and automated technical defect flags, which were used to filter out defective recordings during preprocessing [4].

During preprocessing, recordings were trimmed, normalized, and segmented into clips of fixed length [4]. Both 10-second and 5-second clips were evaluated. The 5-second setting did not reduce classification quality while increasing the number of training examples and improving the temporal localization of short respiratory events. It also aligned the experiments with the 5-second clip format used in the reference implementations and datasets that the reference implementations were trained on,

Table 1 Dataset summary

Metric	Details
Total number of recordings	1,613
Sex distribution	F (542), M (1,071)
Age distribution	0–47 years
Conditions	Asthma (1,113), Healthy (133), Other pathologies (367)
Recording points	Trachea (˜60%), Chest, Back, Mouth

such as ESC-50 [39]. Therefore, 5-second clips that were sequentially extracted from the central part of each recording were used in the final experiments [4, 39].

The 1613 recordings include multiple recordings and multiple recording points per patient. Tracheal recordings dominate (about 60%), followed by chest, back, and mouth. Recording sources include specialized systems (878 recordings), mobile phones (345), remote web collection (85), and standard computers (63), with 242 recordings lacking source metadata [4]. The dataset is single-center, which reduces device variability but limits external validity.

Table 1 summarizes the dataset characteristics. Other pathologies in the dataset include cystic fibrosis, recurrent obstructive bronchitis, pneumonia, and bronchopulmonary dysplasia [4].

3 Models and fine-tuning

3.1 Audio spectrogram transformer model

The possibility of applying CNN models for asthma diagnostics was studied in our previous work using the DenseNet201 CNN architecture [4]. In this pipeline, respiratory sounds were converted into two-dimensional mel-spectrograms, where time and frequency information are represented as an image.

Within this framework, here we evaluate all models using spectrograms of the same breath sound recordings.

The source code of the implementation accompanies the paper and is available at https://github.com/Ben-Gurion-University-Lab/transformers-paper [3].

The AST model [19] is a transformer-based model designed for spectrogram analysis. It is largely based on the ViT architecture [12]. The input two-dimensional spectrogram is split into small patches, which are then linearized and passed through a transformer encoder with positional embeddings. A special class token is added, and its output is used for classification, following the same general principle as Bidirectional Encoder Representations from Transformers (BERT) [11].

AST is based on transfer learning from a ViT pretrained on ImageNet [10]. In the original AST implementation, these weights are further adapted on AudioSet [17], a

large sound dataset. In this study, we use publicly available AST weights obtained through this pretraining pipeline. The model was selected because it showed strong reported results on ESC-50, reaching 95.6% accuracy [39].

The AST model is larger than DenseNet201 but still moderate in size for a transformer model, with about 87 million parameters in the reference implementation. Since the medical dataset is limited, to reduce overfitting, during training we used increased weight decay, gradient clipping, low learning rate with warmup, and early stopping when metrics stagnated.

To maintain clinical validity, extensive audio augmentation, such as noise injection, time shifts, or pitch shifts, was avoided.

3.2 Moondream2 vision-language model

For the multimodal experiment, we selected Moondream2 [26], a compact vision-language model (VLM). The Moondream2 architecture combines a vision encoder based on the ViT with a language model based on Phi [27]. Its size of about 1.9B parameters makes it substantially smaller than many VLMs available at the time of the research, which made it a practical candidate for experiments that may later target clinical deployment on standard hardware. As input the model expects a 3-channel image, usually a red, green, blue (RGB) image, together with a text prompt.

Respiratory sounds were converted to mel-spectrograms using the approach described in [4]. Unlike standard spectrograms with a linear frequency scale, mel-spectrograms use a non-linear scale that better imitates human hearing. The spectra were computed in the 0-8 kHz frequency range using 128 mel coefficients. To build a three-channel input for the vision encoder, each audio clip was converted into a pseudo-RGB image in which the three color channels correspond to mel-spectrograms computed with different analysis window sizes: short (25 ms), medium (100 ms), and long (175 ms). Unlike the DenseNet201 pipeline, which used three-channel mel-spectrograms normalized to [0,1] value range, the Moondream2 pipeline uses Tagged Image File Format (TIFF) RGB images in the standard [0,255] value range to describe pixel intensity. Figure 1 shows an example input image.

In addition to the visual input, the model receives structured text metadata (sex, age, recording point) and a short instruction (task text) that explicitly asks the model to produce a diagnosis from the spectrogram. The structured prompt used in the experiments is shown in Figure 2.

Two prompt formats were tested: natural language and JSON-like structured text. Pilot runs on a balanced subset showed no measurable difference in classification accuracy, so the structured format was chosen because it simplifies automatic parsing of model outputs during evaluation (Figure 2).

The Moondream2 experiments used a balanced sample containing 200 persons: 100 patients with asthma and 100 without asthma (both healthy and with other respiratory diagnoses). The split into training and test sets was stratified to preserve class balance and to avoid clip overlap between training and evaluation.

The model was fine-tuned for the target classification task. Because pre-trained weights on large-scale image and text datasets are openly available, transfer learning

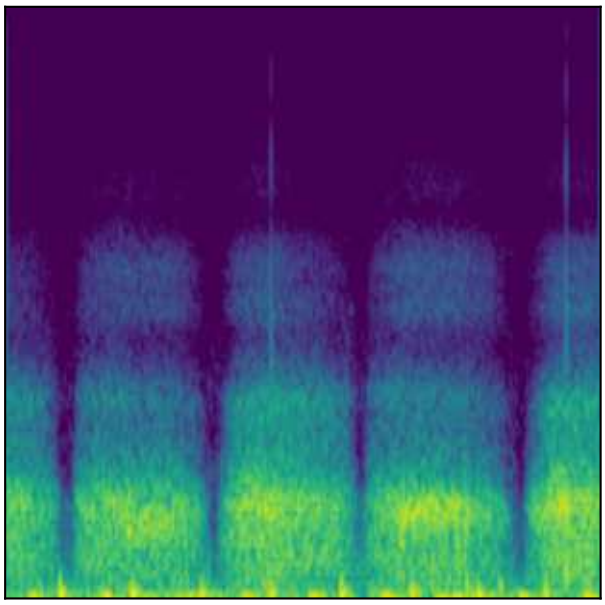


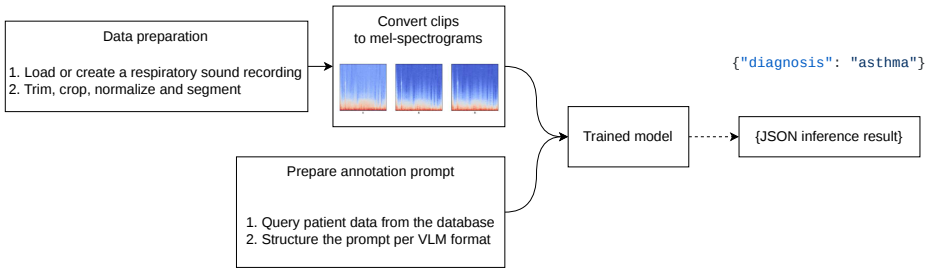
Fig. 1 Example spectrogram input in TIFF format

was used to reduce computational cost. The fine-tuning was performed using the Low-Rank Adaptation (LoRA) [23] method.

The LoRA method keeps the original model weights frozen and adds a small set of trainable low-rank matrices (called adapters) to selected layers. This reduces memory requirements while preserving most of the pretrained model. In our setup, these adapters were introduced into the query/value projection layers (`Wqkv` and `out_proj`) in the language model, the intermediate multi-layer perceptron (MLP) blocks (`fc1` and `fc2`), and the `proj` layer that maps visual features into the language embedding space.

Adapter parameters accounted for about 3.8% of the total parameter count, keeping the trainable footprint small relative to full fine-tuning. The final classifier head and embeddings were unfrozen and trained along with the adapters.

Training ran for 3 epochs (125 batch iterations) with a batch size of 30 and the Adam optimizer in 8-bit mode to reduce memory usage. Training was performed on an Nvidia V100 graphics processing unit (GPU). A cosine learning rate schedule with a short warmup was used. During evaluation, the model produced a JSON-formatted diagnosis (e.g., {`"diagnosis": "asthma"`}); outputs were normalized for JSON validity and compared with reference labels.




```
{
    "task": "Write a diagnosis for this patient by analyzing the respiratory sound spectrogram.",
    "patient_sex": "M",
    "patient_age": 9,
    "record_point": "posterior thoracic rib",
    "spectrogram_bins_width": "[0.025, 0.1, 0.175]"
}
```

Fig. 2 Moondream2 inference pipeline from spectrogram and prompt to JSON output

3.3 Evaluation method and metrics

Computer-assisted analysis of breath sounds was performed at Ben-Gurion University of the Negev (Israel). The software was implemented in Python using NumPy [21], librosa [29], SciPy [49], scikit-learn [38], PyTorch [37], the Hugging Face Transformers library [51], and the Python standard library. All tools used in the study are open-source software distributed under permissive licenses. AI assistants used during research code maintenance included OpenAI GPT-4o [33], OpenAI GPT-4.1 [34], Qwen3-235B-A22B-Instruct-2507 [40], and Claude Sonnet 4 [2].

All models were evaluated as binary classifiers for the "Asthma vs Not Asthma" task. Predictions were compared with reference diagnostic labels. The same metric set was used to compare the AST, Moondream2, and DenseNet201 models.

Given the TP (True Positive) is the number of asthma recordings correctly classified as asthma, FN (False Negative) is the number of asthma recordings classified as non-asthma, TN (True Negative) is the number of non-asthma recordings correctly classified as non-asthma, and FP (False Positive) is the number of non-asthma recordings classified as asthma. Then sensitivity was calculated as TP / (TP + FN), specificity as TN / (TN + FP), and accuracy as (TP + TN) / (TP + FP + TN + FN). F1-score was calculated as the harmonic mean of precision and sensitivity, where precision is TP / (TP + FP). The Youden index was calculated as sensitivity + specificity - 1.

Sensitivity and specificity were reported because the task is diagnostic: the first metric measures asthma cases missed by the classifier, and the second metric measures non-asthma cases incorrectly classified as asthma. F1-score combines precision and sensitivity for the asthma class. The Youden index summarizes the sensitivity-specificity balance in a single value.

Table 2 Diagnostic metrics for the "Asthma vs Not Asthma" task (diagnostics)

Model	Accuracy, %	Sensitivity, %	Specificity, %	F1-score, %	Youden Index
DenseNet201	87.0	88.0	86.0	87.1	0.740
AST	98.7	98.7	98.7	98.7	0.974
Moondream2	86.5	96.0	77.0	87.7	0.730

4 Results

4.1 Result comparison

All results reported below are for the "Asthma vs Not Asthma" task. For this task, the AST model achieved the best overall performance. In terms of accuracy, sensitivity, specificity and F1-measure, the model achieved an accuracy of 98.7%, and the Youden index was 0.974 (test split contained 153 recordings of patients and healthy volunteers). These values indicate a clear improvement over the DenseNet201 CNN baseline evaluated on the same dataset (Table 2).

After fine-tuning, the Moondream2 model achieved lower performance than the AST model but remained competitive with the DenseNet201 CNN baseline (Table 2). For the best checkpoint, the "Asthma vs Not Asthma" task yielded an accuracy of 86.5%, sensitivity of 96.0%, specificity of 77.0%, F1-score of 87.7%, and a Youden Index of 0.730.

The DenseNet201 checkpoint developed in our previous work [4] serves as the internal baseline for the same dataset and task. Its metrics were 87.0% accuracy, 88.0% sensitivity, 86.0% specificity, 87.1% F1-score, and a Youden Index of 0.740. Table 2 and Figure 3 summarize the results for DenseNet201, AST, and Moondream2.

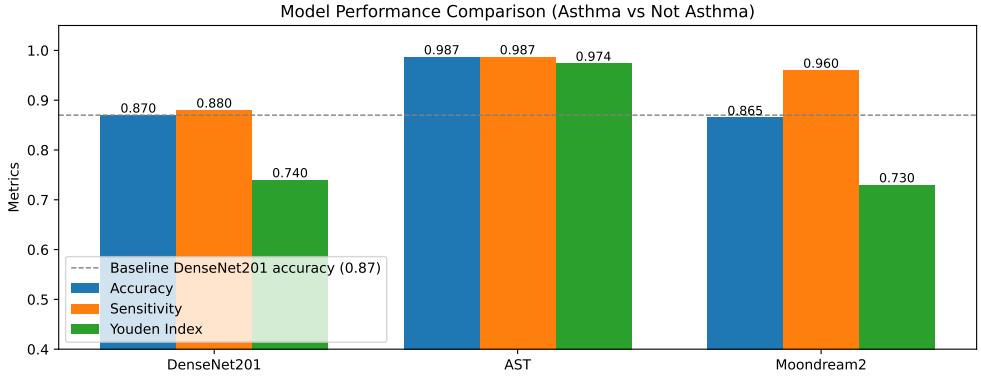


Fig. 3 Model performance comparison for the "Asthma vs Not Asthma" task (diagnostics)

The AST model outperformed both alternative models on this task. DenseNet201 and Moondream2 were close in accuracy, but their error profiles differed: Moondream2

showed higher sensitivity and lower specificity, whereas DenseNet201 remained more balanced.

4.2 The role of metadata

Selective removal of certain textual information demonstrated that the text input is highly important for stable inference of the Moondream2 model in our setup.

Removing the technical text block defining spectrogram parameters (e.g., bin widths) resulted in a decrease in accuracy to 67%, though the model retained partial predictive capacity.

Removing the patient information block (patient age, sex, recording point) caused a collapse of the classifier, producing 100% false negatives for asthma. In this configuration, the model relied on the text conditioning to reach a stable decision boundary.

4.3 Computational efficiency and clinical use

DenseNet201 (about 20M parameters) required fractions of a second on the GPU and roughly 2-5 seconds on the central processing unit (CPU). In our measurements, AST (about 87M parameters) also produced results within a few seconds on the CPU, with sub-second inference on the GPU. Moondream2 required about 1.8 seconds per 5-second clip on the CPU and less than 0.5 seconds on the GPU.

In a real clinical workflow where interaction with the patient is involved, collecting patient metadata, recording a respiratory sample, checking signal quality, and repeating the recording if needed take substantially longer than the model inference. In this context, a processing time below 5 seconds per clip is compatible with practical use during a clinical visit.

5 Discussion

5.1 Interpretability and analysis model inference

Interpretability is an important consideration for clinical applications. For DenseNet and other CNNs, Grad-CAM [45] can produce heatmaps that highlight spectrogram regions most responsible for the predicted class. These visualizations can indicate frequency bands and time intervals that drive a diagnosis. For AST, Grad-CAM is not directly applicable due to the absence of convolutional layers. Instead, attention map analysis can be used. AST splits the spectrogram into patches. By averaging attention matrices across heads, a two dimensional attention map can be reconstructed. This can show which time-frequency regions contribute most to the class token. Figure 4 shows two processing paths for the same spectrogram sample: one processed with the DenseNet201 CNN and visualized with Grad-CAM, and the other processed with AST and visualized with an attention map. The horizontal lines mark the frequency band where both spectrogram-only models show increased activation, approximately in the 800-1800 Hz range around the third breath cycle. For VLMs, interpretability can be enhanced by generating not only a diagnosis but also a textual rationale (e.g., a description of detected wheezes). Implementing this requires additional training data

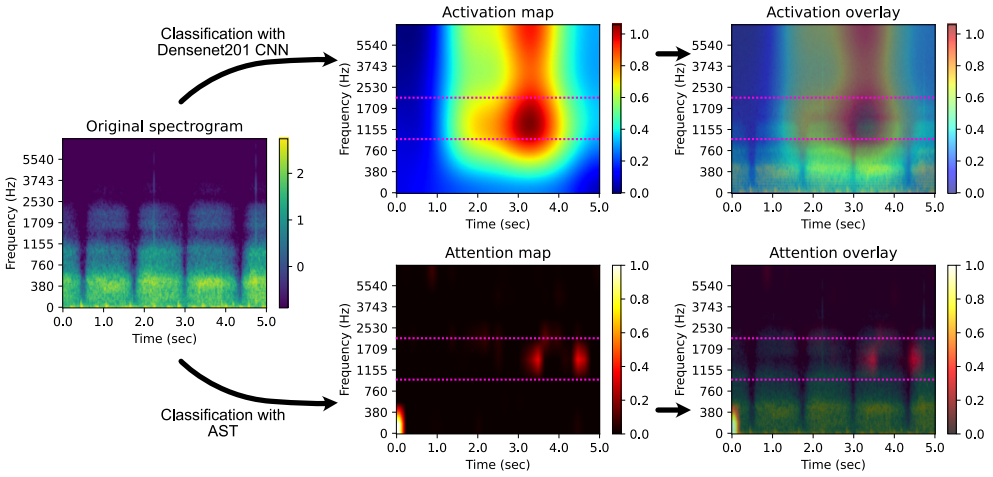


Fig. 4 Interpretability via visualization: Grad-CAM heatmap and activation map overlays on a spectrogram

with detailed expert annotations (e.g., time-aligned descriptions of specific acoustic events or data derived from other clinical modalities). This shifts the task toward data annotation.

5.2 AST versus Moondream2

The comparison between AST and Moondream2 is an important aspect of the interpretation of the study. AST uses only respiratory sound information encoded as a spectrogram, whereas Moondream2 uses a similar spectrogram together with structured patient metadata. Despite the additional clinical context, the multimodal formulation did not outperform the specialized sound-only transformer.

This result does not imply that multimodal modeling is unhelpful for respiratory diagnostics. One plausible explanation is that the image-based encoder in a compact VLM is less well matched to respiratory spectrograms than a model designed specifically for sound classification. A related bottleneck in the visual part of a VLM has been discussed in another application domain by Liu et al. [28], so it is reasonable to treat the same mechanism here as a working hypothesis rather than as an established explanation. Under this interpretation, the main limitation is not the use of metadata itself but the mismatch between an image-based encoder and respiratory sound data.

This observation suggests a clear direction for further work: combine structured patient metadata with an audio-native encoder rather than with a generic vision encoder. A hybrid architecture built around the AST model or a similar sound-specialized transformer may preserve the gains of multimodal inference without sacrificing the quality of spectrogram analysis.

6 Conclusion

To develop computer-assisted methods for asthma diagnostics, we evaluated Moondream2, a vision-language model, and the Audio Spectrogram Transformer model and compared them with a DenseNet201 CNN baseline. For training and evaluating models, the same dataset containing respiratory sound recordings and patient information was used. Moondream2 demonstrates quantitative performance comparable to that of the DenseNet201 CNN baseline, with 86.5% accuracy, 96.0% sensitivity, 77.0% specificity, an F1-score of 87.7%, and a Youden index of 0.730 (Table 2 and Figure 3).

It was shown that structured patient metadata can be incorporated directly into the inference process while producing structured JSON outputs. This incorporation increases the quantitative performance of Moondream2. By training and evaluating the models on a supervised dataset, we demonstrated that AST clearly outperforms both Moondream2 and the established DenseNet201 CNN baseline. For the "Asthma vs Not Asthma" task, AST achieved 98.7% across accuracy, sensitivity, specificity, and F1-score, and a Youden Index of 0.974.

The achieved performance exceeds the sensitivity of physician lung auscultation reported in the literature, where sensitivity for obstructive lung disease is around 30% and specificity is around 90% [5].

The multimodal formulation remains valuable as a proof of concept and points toward a natural next step: combining structured patient metadata with an audio-native encoder.

The limitations of this study are mainly related to the scope of the multimodal experiment. Moondream2 was only given the structured clinical context available in the dataset: sex, age, and recording point. Other data that can be important for asthma diagnosis, such as spirometry, allergy status, blood test results, symptom history or smoking status for adults, were not available. Also the multimodal experiment used Moondream2 which uses a generic vision encoder, while AST is designed for spectrogram analysis.

The results demonstrate that transformer-based models can serve as a basis for the development of computer-assisted methods for asthma diagnosis, as well as for continuous and remote monitoring of patient condition.

Statements and Declarations

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. This study used anonymized recordings and clinical metadata for secondary analysis, with no new participant recruitment or intervention, therefore, no ethics approval was required for this analysis. The original dataset was collected under a protocol approved by the Ethics Committee of Perm State Medical University in accordance with the Declaration of Helsinki, with informed consent obtained from participants or their legal representatives.

Consent for publication. Not applicable.

Data availability. The dataset used in this study is the institutional property of Perm State Medical University, has no public DOI, URL, or accession number. To maintain compliance with regulatory requirements the dataset is available to institutions upon request. Requests for access should be directed to Sergey Malinin at info@mstek.ru.

Materials availability. Not applicable.

Code availability. The source code of the implementation accompanies the paper [3].

Author contributions. Conceptualization contributed by TA and VS; data collection and annotation by EF and TA; data curation by TA and EF; formal analysis contributed by TA, VS, GF; investigation contributed by TA, VS, GF, EF; methodology contributed by TA, VS; software contributed by TA; visualization contributed by TA; writing the original draft contributed by TA, VS; editing contributed by TA, VS, GF, EF. All authors reviewed and approved the final manuscript.

References

[1] Aljaddouh B, D M, Alaswad F (2024) Multimodal disease detection and classification using breath sounds and vision transformer for improved diagnosis. Procedia Computer Science 235:1436–1444. https://doi.org/10.1016/j.procs.2024.04.135, URL https://www.sciencedirect.com/science/article/pii/S1877050924008111, international Conference on Machine Learning and Data Engineering (ICMLDE 2023)

[2] Anthropic (2025) Claude 4 system card: Claude opus 4 and claude sonnet 4. URL https://www.anthropic.com/claude-4-system-card

[3] Aptekarev T (2026) Transformer architectures for respiratory sound analysis and multimodal diagnosis. URL https://github.com/Ben-Gurion-University-Lab/transformers-paper

[4] Aptekarev T, Sokolovsky V, Furman E, et al (2023) Application of deep learning for bronchial asthma diagnostics using respiratory sound recordings. PeerJ Computer Science 9:e1173. https://doi.org/10.7717/peerj-cs.1173

[5] Arts L, Lim E, Ven P, et al (2020) The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. Scientific Reports 10. https://doi.org/10.1038/s41598-020-64405-6

[6] Asher MI, Rutter CE, Bissell K, et al (2021) Worldwide trends in the burden of asthma symptoms in school-aged children: Global asthma network phase I cross-sectional study. Lancet 398:1569–1580. https://doi.org/10.1016/S0140-6736(21)01450-1

[7] Association WM (2013) World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. JAMA 310(20):2191–2194. https://doi.org/10.1001/jama.2013.281053

[8] Bahoura M, Lu X (2006) Separation of crackles from vesicular sounds using wavelet packet transform. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, pp II–II, https://doi.org/10.1109/ICASSP.2006.1660533

[9] Bhattacharya D, Sharma NK, Dutta D, et al (2023) Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. Scientific Data 10(1):397. https://doi.org/10.1038/s41597-023-02266-0, URL https://doi.org/10.1038/s41597-023-02266-0

[10] Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255, https://doi.org/10.1109/CVPR.2009.5206848

[11] Devlin J, Chang M, Lee K, et al (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL http://arxiv.org/abs/1810.04805, 1810.04805

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. CoRR abs/2010.11929. URL https://arxiv.org/abs/2010.11929, 2010.11929

[13] Furman E, Malinin S, Furman G, et al (2014) Computer-assisted assay of respiratory sounds of children suffering from bronchial asthma. Sovremennye Tehnologii v Medicine Volume 6. URL https://med-click.ru/uploads/files/docs/computer-assisted-assay-of-respiratory-sounds-of-children-suffering-from-bronchial-asthma.pdf

[14] Furman E, Yakovleva E, Malinin S, et al (2014) P50 - a new modality using breath sound analysis in pediatric asthma. Clinical and Translational Allergy 4(Suppl 1):P105. https://doi.org/10.1186/2045-7022-4-S1-P105, URL https://doi.org/10.1186/2045-7022-4-S1-P105

[15] Furman E, Malinin S, Furman G, et al (2020) Respiratory sound analysis for bronchial asthma diagnostics. IOSR Journal of Engineering 10(1):53–59. Received 17 January 2020; Accepted 05 February 2020

[16] Gelman A, Furman EG, Kalinina NM, et al (2022) Computer-aided detection of respiratory sounds in bronchial asthma patients based on machine learning method. Sovremennye Tehnologii v Medicine 14(5):45–51. https://doi.org/10.17691/stm2022.14.5.05

[17] Gemmeke JF, Ellis DPW, Freedman D, et al (2017) Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 776–780, https://doi.org/10.1109/ICASSP.2017.7952261

[18] Global Initiative for Asthma (2021) Global strategy for asthma management and prevention. Tech. rep., Global Initiative for Asthma, URL https://ginasthma.org/wp-content/uploads/2023/04/GINA-Main-Report-2021-V2-WMSA.pdf

[19] Gong Y, Chung YA, Glass J (2021) Ast: Audio spectrogram transformer. In: Interspeech 2021, pp 571–575, https://doi.org/10.21437/Interspeech.2021-698, URL https://www.isca-archive.org/interspeech_2021/gong21b_interspeech.html

[20] Gurung A, Scrafford CG, Tielsch JM, et al (2011) Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis. Respir Med 105(9):1396–1403. https://doi.org/10.1016/j.rmed.2011.05.007

[21] Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2, URL https://doi.org/10.1038/s41586-020-2649-2

[22] Hsu FS, Huang SR, Huang CW, et al (2021) Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database-hf_lung_v1. PLOS ONE 16(7):1–26. https://doi.org/10.1371/journal.pone.0254134, URL https://doi.org/10.1371/journal.pone.0254134

[23] Hu EJ, Shen Y, Wallis P, et al (2022) Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations, URL https://arxiv.org/abs/2106.09685

[24] Huang G, Liu Z, Weinberger KQ (2016) Densely connected convolutional networks. CoRR abs/1608.06993. URL http://arxiv.org/abs/1608.06993, 1608.06993

[25] Kim Y, Hyon Y, Jung S, et al (2021) Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. Scientific Reports 11(1):17186. https://doi.org/10.1038/s41598-021-96724-7, URL https://doi.org/10.1038/s41598-021-96724-7

[26] Korrapati V (2025) moondream2 (revision 6b714b2). https://doi.org/10.57967/hf/6762, URL https://huggingface.co/vikhyatk/moondream2

[27] Li Y, Bubeck S, Eldan R, et al (2023) Textbooks are all you need ii: phi-1.5 technical report. CoRR abs/2309.05463. https://doi.org/10.48550/arXiv.2309.05463, URL https://arxiv.org/abs/2309.05463

[28] Liu J, Zeng W, Zhang X, et al (2025) On the perception bottleneck of VLMs for chart understanding. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics, Suzhou, China, pp 10829–10841, https://doi.org/10.18653/v1/2025.findings-emnlp.573, URL https://aclanthology.org/2025.findings-emnlp.573/

[29] McFee B, Metsai A, McVicar M, et al (2022) librosa/librosa: 0.9.2. https://doi.org/10.5281/zenodo.6759664, URL https://doi.org/10.5281/zenodo.6759664

[30] Mortimer K, Lesosky M, García-Marcos L, et al (2022) The burden of asthma, hay fever and eczema in adults in 17 countries: Gan phase I study. Eur Respir J 60:2102865. https://doi.org/10.1183/13993003.02865-2021

[31] Murphy RL, Vyshedskiy A, Power-Charnitsky VA, et al (2004) Automated lung sound analysis in patients with pneumonia. Respir Care 49(12):1490–1497

[32] Nissan(Nisselevitch) M, Gavriely N (1993) A microcomputer based lung sounds analysis. Computer Methods and Programs in Biomedicine 40(1):7–13. https://doi.org/10.1016/0169-2607(93)90045-M, URL https://www.sciencedirect.com/science/article/pii/016926079390045M

[33] OpenAI (2024) GPT-4o system card. URL https://openai.com/index/gpt-4o-system-card/

[34] OpenAI (2025) Introducing GPT-4.1 in the API. URL https://openai.com/index/gpt-4-1/

[35] Orlandic L, Teijeiro T, Atienza D (2021) The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Scientific Data 8(1):156. https://doi.org/10.1038/s41597-021-00937-4, URL https://doi.org/10.1038/s41597-021-00937-4

[36] Palanisamy K, Singhania D, Yao A (2020) Rethinking CNN models for audio classification. CoRR abs/2007.11154. URL https://arxiv.org/abs/2007.11154, 2007.11154

[37] Paszke A, Gross S, Massa F, et al (2019) Pytorch: An imperative style, high-performance deep learning library. In: Wallach HM, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems 32, pp 8024–8035, URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[38] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12:2825–2830

[39] Piczak KJ (2015) Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '15, p 1015–1018, https://doi.org/10.1145/2733373.2806390, URL https://doi.org/10.1145/2733373.2806390

[40] Qwen Team (2025) Qwen3-235B-A22B-Instruct-2507. URL https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507

[41] Reichert S, Gass R, Hajjam A, et al (2009) The asap project: A first step to an auscultation's school creation. Respiratory Medicine CME 2(1):7–14. https://doi.org/10.1016/j.rmedc.2009.01.001, URL https://www.sciencedirect.com/science/article/pii/S1755001709000025

[42] Rocha B, Filos D, Mendes L, et al (2018) A respiratory sound database for the development of automated classification. In: Maglaveras N, Chouvarda I, Carvalho P (eds) Precision Medicine Powered by pHealth and Connected Health, IFMBE Proceedings, vol 66. Springer, Singapore, pp 33–37, https://doi.org/10.1007/978-981-10-7419-6_6, URL https://doi.org/10.1007/978-981-10-7419-6_6

[43] Ruchonnet-Métrailler I, Siebert JN, Hartley MA, et al (2024) Automated interpretation of lung sounds by deep learning in children with asthma: Scoping review and strengths, weaknesses, opportunities, and threats analysis. Journal of Medical Internet Research 26:e53662. https://doi.org/10.2196/53662

[44] Santosh KC, Rasmussen N, Mamun M, et al (2022) A systematic review on cough sound analysis for covid-19 diagnosis and screening: is my cough sound covid-19? PeerJ Computer Science 8:e958. https://doi.org/10.7717/peerj-cs.958, URL https://doi.org/10.7717/peerj-cs.958

[45] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 618–626, https://doi.org/10.1109/ICCV.2017.74

[46] Tang S, Hu X, Atlas L, et al (2022) Hierarchical multi-modal transformer for automatic detection of covid-19. In: Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning. Association for Computing Machinery, New York, NY, USA, SPML 22, p 197–202, https://doi.org/10.1145/3556384.3556414, URL https://doi.org/10.1145/3556384.3556414

[47] Tomita K, Nagao R, Touge H, et al (2019) Deep learning facilitates the diagnosis of adult asthma. Allergol Int 68(4):456–461. https://doi.org/10.1016/j.alit.2019.04.010

[48] Urquhart R, McGhee J, Macleod J, et al (1981) The diagnostic value of pulmonary sounds: A preliminary study by computer-aided analysis. Computers in Biology and Medicine 11(3):129–139. https://doi.org/10.1016/0010-4825(81)90002-0,

URL https://www.sciencedirect.com/science/article/pii/0010482581900020

[49] Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

[50] Vos T, et al (2019) Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet 396(10258):1204–1222

[51] Wolf T, Debut L, Sanh V, et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6, URL https://aclanthology.org/2020.emnlp-demos.6

[52] Xia T, Han J, Mascolo C (2022) Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. Experimental Biology and Medicine 247(22):2053–2061. https://doi.org/10.1177/15353702221115428, URL https://doi.org/10.1177/15353702221115428, pMID: 35974706, https://doi.org/10.1177/15353702221115428