Pretrained Neural Audio Models for Asthma Detection from Voice and Speech

Leticia Puttlitz Boll
Universidade de São Paulo (USP), Brazil
leticia.puttlitz@usp.br

Antonio Oss Boll
Universidade de São Paulo (USP), Brazil
aoboll@usp.br

Yan Anderson Pires de Oliveira
Universidade de São Paulo (USP), Brazil
yananderson@usp.br

Victor dos Santos Silva
Universidade de São Paulo (USP), Brazil
victorsantos@usp.br

Mariana Lopes Pestana
Universidade de São Paulo (USP), Brazil
marylopestana@usp.br

Celso Ricardo Fernandes de Carvalho
Universidade de São Paulo (USP), Brazil
cscarval@usp.br

Marcelo Matheus Gauy
Universidade Estadual Paulista (UNESP), Brazil
marcelo.gauy@unesp.br

Marcelo Finger
Universidade de São Paulo (USP), Brazil
mfinger@ime.usp.br

Abstract

Asthma is a chronic respiratory disease that affects breathing and may also influence speech and voice production. In this paper, we examine whether short mobile-recorded Brazilian Portuguese voice and speech audio contain cues that can be used to distinguish individuals with asthma from those without asthma. We approach this problem using transfer learning with pretrained neural audio models based on convolutional architectures trained on large-scale audio datasets (PANNs). We evaluate two recording types: sustained vowel phonation and read speech. Models are trained for a binary classification task and evaluated at both the segment level and the patient level. Read speech performs better than sustained vowels. The best configuration (CNN14 on speech) achieves 0.85 patient-level balanced accuracy (accuracy 0.85) with ROC-AUC 0.93 and PR-AUC 0.98, performing comparably to CNN10. Training from scratch performs worse than fine-tuning a pretrained model, showing that pretraining helps when data is limited. Performance also varies across age groups, suggesting demographic sensitivity. These findings support the feasibility of audio-based asthma classification from voice and speech and motivate further investigation of pretrained audio models in biomedical applications.

1 Introduction

Asthma is a chronic airway disease characterized by airflow limitation and symptoms such as wheezing, dyspnea, chest tightness, and cough, which may lead to exacerbations requiring hospitalization (Global Initiative for Asthma, 2024; Chipps et al., 2023; Fuhlbrigge et al., 2012). Because asthma affects breathing patterns, it can also introduce acoustic changes in voice and speech.

Advances in deep learning have enabled powerful models for audio processing, such as PANNs (Kong et al., 2020), which are pretrained on thousands of hours of audio and can achieve strong performance even with limited labeled data. This makes pretrained audio models a promising approach for asthma detection from short recordings. In this context, voice and speech data collected via mobile devices provide a non-invasive and low-cost data source that can be analyzed with machine learning to investigate clinically relevant acoustic patterns for respiratory disease classification.

In this work, we focus on the binary task of distinguishing *asthma* vs. *non-asthma* using short recordings of sustained vowels and read speech. We evaluate pretrained convolutional audio networks adapted to this task. We compare shallow and deeper architectures, pretrained versus scratch models, and analyze demographic subgroups to assess robustness. On read speech, the pretrained CNN14 achieves the best patient-level balanced accuracy (0.85) and accuracy (0.85), with ROC-AUC 0.93 and PR-AUC 0.98, although its performance is comparable to CNN10. The pretrained CNN10 shows slightly higher accuracy (0.86) but lower

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 58–67
April 13-16, 2026

balanced accuracy, and both outperform sustained vowels and scratch models.

2 Related Work

2.1 Pretrained Neural Audio Models

Pretrained neural audio models have increasingly been adopted in healthcare and related audio applications due to their ability to learn robust acoustic representations. Such models have demonstrated strong performance across diverse tasks, including sound event detection (Xu et al., 2023), emotion recognition (Gauy and Finger, 2022), and respiratory disease identification (Gauy et al., 2023; Matheus Gauy et al., 2026).

In the context of emotion recognition, Gauy and Finger (2022) showed that large-scale audio pretraining improves performance, allowing models to outperform baselines even with limited labeled data. Similar gains have been reported in voice-based neurological disorder detection, where pretrained convolutional neural networks operating on spectrograms of sustained vowel recordings outperformed models based on handcrafted acoustic features for Parkinson's disease classification (Rahmatallah et al., 2025).

Furthermore, results on the OPERA respiratory audio benchmark indicate that models pretrained on large and diverse datasets such as AudioSet consistently surpass both models trained from scratch and those pretrained exclusively on respiratory sounds, reinforcing the value of large-scale general-domain pretraining for medical audio analysis (Niizumi et al., 2025).

2.2 Machine Learning for Asthma Detection

Artificial intelligence and machine learning have been increasingly applied to asthma screening, phenotyping, and disease monitoring across a wide range of clinical and biomedical data modalities (Exarchos et al., 2020). Prior work includes asthma classification using machine-learning models trained on pulmonary function test results combined with clinical variables (Topalovic et al., 2017), as well as leveraging cough acoustics as a complementary audio biomarker (Alqudaihi et al., 2021). In addition to acoustic signals, other respiratory measurements have also been explored, including quantitative features derived from exhaled CO_2 waveforms (Singh et al., 2018) and breath-based biomarkers such as exhaled nitric oxide for diagnosis and severity monitoring (Yin et al., 2025). Beyond respiratory signals, asthma classification has further been investigated using molecular biomarkers, such as nasal gene-expression signatures (Pandey et al., 2018), and routine blood biomarkers modeled with machine-learning techniques (Zhan et al., 2020).

In a related line of work focusing on voice-based biomarkers for asthma, several studies have relied on sustained vowel recordings. An XGBoost-based classifier was proposed using handcrafted acoustic features extracted from sustained vowel /a:/ recordings (Lyu et al., 2025). Similarly, MeLoDicA (Looi et al., 2024) introduced a framework based on handcrafted spectral and temporal voice features, showing that sustained vowels achieved the highest performance among the evaluated audio types.

Beyond sustained vowels, asthma detection has also been explored using speech signals. Real-time asthma classification using speech and respiratory sounds has been investigated (Iqbal et al., 2022), and conventional classifiers such as GMMs and CNNs have been applied to MFCC features extracted from speech (Iqbal et al., 2024). In addition, machine-learning models have been trained on short Turkish phonetic utterances (Gezer et al., 2025).

Overall, these approaches rely predominantly on manually designed acoustic features and conventional machine-learning models, without leveraging large-scale pretrained audio representations or end-to-end learned embeddings.

2.3 Voice and Speech as Biomarkers in Respiratory Diseases

Recent reviews have highlighted the use of audio-based biomarkers for respiratory disease detection, including cough, lung sounds, and voice or speech signals (Kapetanidis et al., 2024). Beyond asthma, similar approaches have been applied to the identification of respiratory diseases. Pretrained audio models have been used to analyze speech and voice recordings from patients with respiratory conditions (Gauy et al., 2023; Matheus Gauy et al., 2026). Voice and speech have also been investigated as biomarkers for COVID-19 detection, where machine-learning models based on acoustic features have shown strong discriminative performance in identifying infected individuals (Verde et al., 2023; Dash et al., 2022). In addition, voice-based methods have been applied to chronic obstructive pulmonary disease (COPD), using embeddings from a wav2vec 2.0 model to classify disease

presence and severity from short voice recordings (Lee et al., 2025).

3 Data

We use a dataset of short voice recordings collected from adult speakers of Brazilian Portuguese performing two speaking tasks: *Speech* and *Vowel*. In the Speech task, participants read a short predefined sentence, while in the Vowel task they sustained the vowel /a:/ for as long as they could. These tasks were chosen to capture different aspects of speech production and respiratory control.

The recordings were collected using mobile devices and processed at a sampling rate of 16 kHz. Data were recorded both in hospital settings and in more uncontrolled environments using participants' personal devices. Because of this, the dataset includes variation in background noise, microphone quality, and recording conditions.

The dataset includes 549 Speech and 538 Vowel recordings, with asthma representing 79% of the samples. The average recording length is 7.5 seconds.

3.1 Clinical and demographic metadata

In addition to the audio recordings, clinical and demographic metadata are collected for each participant. These include age, sex, anthropometric measures (weight and height), vital signs, as well as information on comorbidities and smoking history. These variables provide important contextual information for the analysis and allow for the assessment of potential demographic biases in the data.

3.2 Demographic distribution

Table 1 summarizes the distribution of recordings by sex for both speaking tasks. The dataset is predominantly female in both cases.

Table 1: Sex distribution by speaking task (unique patients).

Sex	Speech	Vowel
Female	434	426
Male	115	112
Total	549	538

Table 2 reports the age distribution using four age bins: under 30 years, 30–45 years, 45–60 years, and over 60 years.

Table 2: Age distribution by speaking task (unique patients).

Age range (years)	Speech	Vowel
< 30	81	80
30–45	140	137
45–60	244	239
> 60	84	82
Total	549	538

4 Preprocessing

Using an energy-based trimming technique that eliminates low-energy regions in relation to the signal's peak, we eliminate silence at the start and finish of each audio file. The resulting waveform is then resampled to 16 kHz and peak-normalized by scaling each waveform to a fixed maximum absolute amplitude (Labied et al., 2022).

4.1 Dataset splitting

The dataset is divided into training, validation, and test sets using a stratified splitting strategy at the patient level, with proportions of 60%, 20%, and 20%, respectively. In order to guarantee that these demographic characteristics are evenly distributed throughout splits, stratification is carried out according to age group and sex (Xu and Goodacre, 2018).

4.2 Class balancing and data augmentation

The dataset is imbalanced with respect to the target classes. To mitigate this effect during model learning, we apply class balancing on the training set only. In our experiments, we use random oversampling, duplicating minority-class recordings until the class counts match those of the majority class.

In addition, to increase robustness to recording variability, we apply waveform perturbations during preprocessing for training examples selected as augmented duplicates. Specifically, one of the following transformations is sampled uniformly at random: additive Gaussian noise, random gain perturbation, pitch shifting, or time stretching (Wei et al., 2020). These operations are applied at the waveform level prior to feature extraction.

4.3 Temporal windowing

We use a sliding-window approach to create fixed-length segments because recordings vary in length. Following previous literature, each waveform is split into 4.0 s windows with a 2.0 s hop (Casanova

et al., 2021), producing a variable number of segments depending on the length of the recording. Recordings are zero-padded to 4.0 s if they are shorter than the window length. Although segments from the same recording may overlap, windowing is performed after splitting the dataset at the patient level. This ensures that segments from the same speaker are only present in one split, preventing similar audio samples from appearing in both training and test sets.

4.4 Noise injection

To further reduce sensitivity to background conditions, we inject environmental noise by mixing each window with a randomly selected noise sample drawn from a separate noise pool. This pool consists of recordings collected in the same hospital environments as the patient data. The noise is added to the signal with a randomly scaled amplitude after being trimmed to the same length as the window. The noise amplitude is randomly scaled to simulate different noise levels while ensuring that the added noise does not dominate or distort the speech signal.

By using this technique, the model is prevented from learning hospital background noise as a cue for asthma, for example, or from linking background noise with the target labels. The model is encouraged to concentrate on vocal and speech-related information rather than environmental artifacts by changing background conditions independently of the labels.

4.5 Time-frequency representations

For each window, the audio signal is converted into a log-Mel spectrogram, a time-frequency representation that summarizes how the signal energy is distributed over time and frequency. This representation is used as input to the neural models.

Figure 1 summarizes the full preprocessing pipeline used in our experiments, from waveform normalization and participant-level splitting to windowing, augmentation, noise mixing, and feature extraction.

5 Models

We use pre-trained convolutional neural networks for audio classification, specifically the CNN10 and CNN14 architectures, as proposed in the PANNs framework (Kong et al., 2020). These models were pre-trained on the AudioSet database (Gemmeke et al., 2017), which contains over two million labeled audio clips corresponding to more than 5,000 hours of audio, and are widely used in prior work as general-purpose audio feature extractors.

The networks take log-Mel spectrograms as input and consist of a sequence of convolutional layers with pooling applied along the time and frequency dimensions. A global pooling layer then summarizes the features into a fixed-length vector for classification.

For the downstream task, the resulting representation is passed to a task-specific classification layer. Instead of keeping the pre-trained backbone fixed, we fine-tune all model layers during training. This enables the learned representations to adapt to the target dataset while still benefiting from the information captured during pre-training.

We evaluate both CNN10 and CNN14 in order to examine the effect of model complexity on classification performance in our experiments.

6 Experimental Setup

6.1 Task definition

We consider a binary classification task between *asthma* and *non-asthma*. Each model takes a fixed-length audio segment, converts it into a log-Mel spectrogram, and outputs a prediction for one of the two classes.

Performance is evaluated at two levels: (i) **segment level**, where each audio segment is classified independently, and (ii) **patient level**, where predictions from multiple segments belonging to the same participant are aggregated to produce a single label per patient. We aggregate segment predictions by averaging the model outputs (logits) across all segments from the same participant and then taking the argmax to obtain a single patient label.

6.2 Models and fine-tuning strategy

Using the CNN10 and CNN14 variants, we assess pretrained convolutional audio models from the PANN family. For classification, a MLP head with two linear layers, dropout (Srivastava et al., 2014), and ReLU activation is added.

All models are fine-tuned end-to-end on the asthma classification task.

6.3 Optimization details

Training is implemented in PyTorch using Adam with weight decay. Using a single learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, we

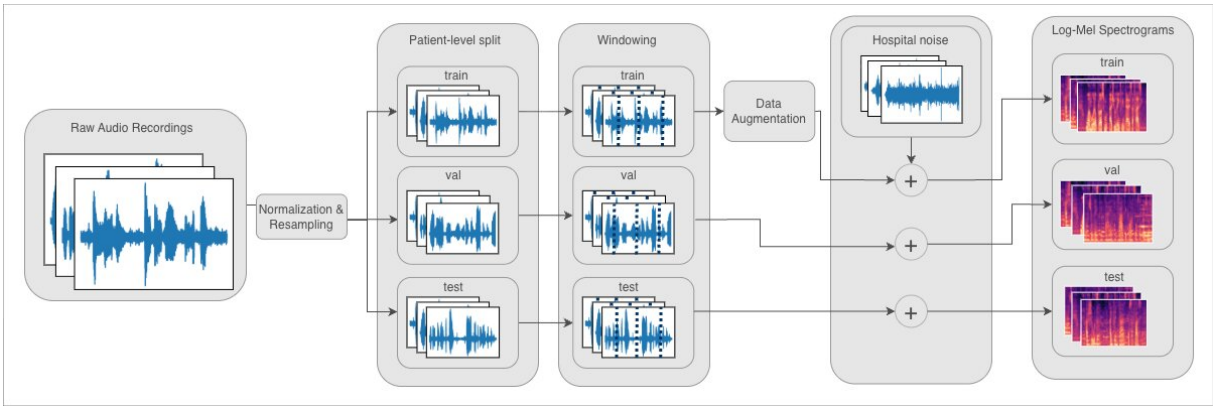

Figure 1: Overview of the audio preprocessing pipeline. Raw recordings are normalized and resampled, split at the participant level into training, validation, and test sets, segmented into fixed-length windows, augmented and mixed with environmental noise, and finally converted into time-frequency representations used as model inputs.

train for up to 50 epochs with batch size 16 and a fixed train/validation/test split of 60/20/20. Early stopping monitors validation, balancing accuracy with ten epochs of patience. Dropout ($p = 0.3$) is included in the MLP head.

6.4 Reproducibility

We fix random seeds and enable deterministic execution. All experiments are repeated with 10 seeds and reported as mean $\pm$ standard deviation.

7 Results

Table 3 reports accuracy and balanced accuracy for asthma vs. non-asthma classification at both segment and patient levels, reported as mean $\pm$ sample standard deviation across random seeds. Read speech yields higher performance than sustained vowels across models, with the strongest results obtained by CNN14 on speech.

Table 4 provides clinical metrics (sensitivity, specificity, and MCC) for all configurations, while Table 5 reports ROC-AUC and PR-AUC. Consistent with Table 3, speech-based models generally show stronger discrimination than vowel-based models.

Effect of pretraining (scratch baseline). We trained a scratch baseline with the same CNN10 Speech configuration but randomly initialized weights in order to separate the impact of extensive audio pretraining. The pretrained CNN10 Speech configuration outperforms the scratch model across Tables 3–5, suggesting that pretraining enhances generalization in this setting.

Patient-level ROC curves. Figure 2 visualizes patient-level ROC curves averaged across 10 random seeds. Pretrained models achieve higher ROC-AUC than the scratch baseline (Table 5), confirming the benefit of large-scale audio pretraining. CNN14 Speech achieves the highest ROC-AUC among pretrained models, closely followed by CNN10 Speech.

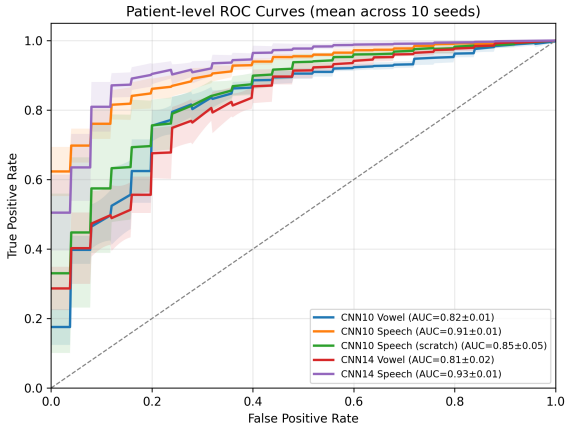

Figure 2: Patient-level ROC curves for the main speech-based configurations.

Statistical comparison (McNemar test). We use McNemar's exact test on paired *patient-level* predictions (Table 6). Speech vs. Vowel is significant for both CNN10 and CNN14 ($p < 0.001$) and pretrained vs. scratch, while CNN10 Speech vs. CNN14 Speech is not ($p = 0.454$).

Bootstrap confidence intervals (patient level). We estimate 95% bootstrap confidence intervals (1,000 patient resamples) for patient-level accuracy and balanced accuracy (Table 7). **CNN14 Speech** achieves the highest mean balanced accuracy, but its confidence interval overlaps with

Table 3: Sample mean ± standard deviation across random seeds for accuracy and balanced accuracy in asthma vs. non-asthma classification.

Model	Input	Segment		Patient	
		Acc	Bal	Acc	Bal
CNN10	Vowel	0.77 ± 0.03	0.77 ± 0.02	0.76 ± 0.03	0.76 ± 0.02
	Speech	**0.84 ± 0.02**	**0.80 ± 0.02**	**0.86 ± 0.01**	**0.81 ± 0.04**
	Speech (scratch)	0.73 ± 0.09	0.75 ± 0.04	0.73 ± 0.11	0.75 ± 0.05
CNN14	Vowel	0.73 ± 0.04	0.72 ± 0.03	0.73 ± 0.05	0.74 ± 0.03
	Speech	**0.84 ± 0.05**	**0.84 ± 0.02**	**0.85 ± 0.05**	**0.85 ± 0.03**

Table 4: Sample mean ± standard deviation across random seeds for sensitivity, specificity, and MCC in asthma vs. non-asthma classification.

Model	Input	Sens	Spec	MCC
CNN10	Vowel (Seg.)	0.77 ± 0.03	0.76 ± 0.03	0.49 ± 0.05
	Vowel (Pat.)	0.75 ± 0.04	0.78 ± 0.03	0.46 ± 0.03
	Speech (Seg.)	**0.86 ± 0.03**	**0.73 ± 0.07**	**0.55 ± 0.03**
	Speech (Pat.)	**0.90 ± 0.02**	**0.73 ± 0.09**	**0.61 ± 0.05**
	Speech (scratch, Seg.)	0.71 ± 0.14	0.78 ± 0.12	0.43 ± 0.07
	Speech (scratch, Pat.)	0.72 ± 0.17	0.77 ± 0.13	0.45 ± 0.08
CNN14	Vowel (Seg.)	0.74 ± 0.07	0.71 ± 0.07	0.41 ± 0.06
	Vowel (Pat.)	0.72 ± 0.07	0.75 ± 0.05	0.41 ± 0.06
	Speech (Seg.)	**0.84 ± 0.07**	**0.84 ± 0.06**	**0.60 ± 0.06**
	Speech (Pat.)	**0.85 ± 0.07**	**0.85 ± 0.07**	**0.65 ± 0.07**

Table 5: Sample mean ± standard deviation across random seeds for ROC-AUC and PR-AUC in asthma vs. non-asthma classification.

Model	Input	AUC	PR-AUC
CNN10	Vowel (Seg.)	0.82 ± 0.02	0.92 ± 0.01
	Vowel (Pat.)	0.82 ± 0.01	0.93 ± 0.00
	Speech (Seg.)	**0.91 ± 0.01**	**0.97 ± 0.00**
	Speech (Pat.)	**0.91 ± 0.01**	**0.98 ± 0.00**
	Speech (scratch, Seg.)	0.84 ± 0.04	0.95 ± 0.02
	Speech (scratch, Pat.)	0.85 ± 0.05	0.94 ± 0.03
CNN14	Vowel (Seg.)	0.80 ± 0.02	0.91 ± 0.01
	Vowel (Pat.)	0.81 ± 0.02	0.93 ± 0.01
	Speech (Seg.)	**0.92 ± 0.01**	**0.98 ± 0.00**
	Speech (Pat.)	**0.93 ± 0.01**	**0.98 ± 0.00**

Table 6: McNemar's exact test on patient-level predictions (two-sided), aggregated across 10 seeds using Fisher's method.

Comparison (A vs. B)	N	Discordant	p
CNN10 Speech vs. CNN10 Vowel	107	26.3	<0.001
CNN14 Speech vs. CNN14 Vowel	107	34.2	<0.001
CNN10 Speech (pretrained) vs. scratch	111	26.4	<0.001
CNN10 vs. CNN14 (Speech)	111	14.3	0.454
CNN10 vs. CNN14 (Vowel)	107	12.2	0.040

Table 7: Patient-level performance with 95% bootstrap confidence intervals (1,000 patient-resamples). Values are mean [95% CI] across 10 seeds.

Model	Acc	Bal. Acc
CNN10 Vowel	0.76 [0.68, 0.82]	0.76 [0.68, 0.84]
CNN10 Speech	0.86 [0.80, 0.91]	0.81 [0.73, 0.89]
CNN10 Speech (scratch)	0.74 [0.68, 0.79]	0.75 [0.68, 0.81]
CNN14 Vowel	0.73 [0.65, 0.79]	0.73 [0.65, 0.82]
CNN14 Speech	0.86 [0.80, 0.91]	0.86 [0.79, 0.91]

CNN10 Speech, indicating comparable performance across the two pretrained models.

7.1 Fairness Analysis across Demographic Groups

We evaluate potential demographic biases through a subgroup analysis based on age and sex using patient-level predictions, considering all model and input configurations. Because basic voice characteristics such as pitch and formant frequencies differ between males and females, this may influence the acoustic patterns learned by the models, motivating explicit evaluation across sex groups (Pépiot, 2015).

Participants were grouped by age into four bins (≤30, 31–45, 46–60, >60 years) and by sex. For each subgroup, we report classification accuracy.

Table 8 summarizes the results. Across both models, performance varies with age, with lower accuracy in the youngest groups and higher accuracy for participants aged 46 years and above. This

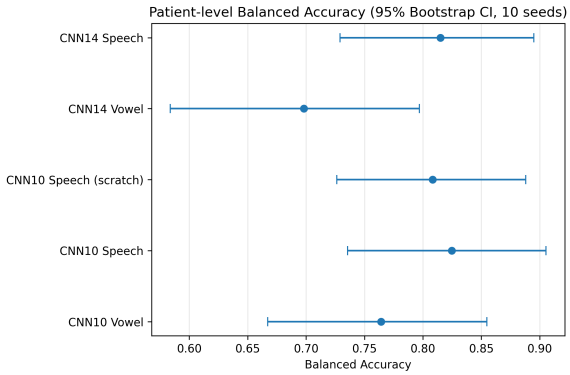

Figure 3: Patient-level balanced accuracy with 95% bootstrap confidence intervals for all evaluated models. CNN10 and CNN14 speech-based models achieve higher balanced accuracy than vowel-based models, while overlapping confidence intervals indicate comparable performance across several configurations.

effect is particularly pronounced for the Speech task, where both CNN10 and CNN14 achieve higher accuracy in older groups but perform poorly for younger participants. In contrast, the Vowel task exhibits more stable performance across age ranges, especially for CNN14, suggesting that sustained phonation may be less sensitive to age-related acoustic variability than read speech. However, estimates for the youngest age group should be interpreted with caution due to the small number of patients. Accuracy across sex is similar, although male participants are underrepresented, limiting statistical strength.

Overall, no strong bias is observed across sex, but the results indicate sensitivity to age differences and limited sample sizes.

8 Discussion

This work shows that pretrained convolutional audio models can effectively detect asthma from short voice and speech recordings. In all experiments, read speech outperformed sustained vowel phonation. This suggests that continuous speech contains richer information related to respiratory and phonatory behavior.

As mentioned in the Related Work section, previous studies have explored asthma detection from voice using handcrafted acoustic features and traditional machine learning models, mainly based on sustained vowels. Direct comparison with our study is not feasible due to differences in datasets, recording conditions, and languages. Nevertheless, our results are consistent with prior work in showing that asthma related information can be extracted from vocal signals. While sustained vowels were effective in earlier studies, our findings suggest that read speech provides even better discrimination. In contrast to these feature based approaches, we use pretrained convolutional audio models, highlighting the benefit of transfer learning.

Table 8: Sample mean ± standard deviation across random seeds for patient-level accuracy by age and sex groups.

Model	Input	Group	N	Acc
		Age groups		
CNN10	Vowel	≤30	4	0.38 ± 0.13
		31–45	9	0.84 ± 0.06
		46–60	15	0.71 ± 0.04
		>60	7	0.86 ± 0.12
	Speech	≤30	4	0.33 ± 0.12
		31–45	9	0.78 ± 0.13
		46–60	16	0.94 ± 0.05
		>60	8	0.99 ± 0.04
	Speech (scratch)	≤30	4	0.38 ± 0.24
		31–45	9	0.56 ± 0.23
		46–60	16	0.75 ± 0.17
		>60	8	0.80 ± 0.25
CNN14	Vowel	≤30	4	0.68 ± 0.24
		31–45	9	0.77 ± 0.10
		46–60	15	0.70 ± 0.13
		>60	7	0.76 ± 0.12
	Speech	≤30	4	0.38 ± 0.13
		31–45	9	0.67 ± 0.20
		46–60	16	0.89 ± 0.07
		>60	8	0.94 ± 0.11
		Sex		
CNN10	Vowel	Female	29	0.76 ± 0.03
		Male	6	0.75 ± 0.04
	Speech	Female	30	0.87 ± 0.01
		Male	7	0.81 ± 0.04
	Speech (scratch)	Female	91	0.74 ± 0.11
		Male	20	0.71 ± 0.11
CNN14	Vowel	Female	29	0.74 ± 0.05
		Male	6	0.69 ± 0.08
	Speech	Female	30	0.86 ± 0.05
		Male	7	0.83 ± 0.06

When comparing architectures, CNN10 and CNN14 exhibit comparable performance on read speech, with CNN14 achieving slightly higher discrimination metrics (e.g., ROC-AUC). This suggests that increasing model complexity does not necessarily yield consistent gains in accuracy in this setting. Given the current dataset size and overlapping confidence intervals, both architectures appear suitable, with CNN10 offering a simpler alternative and CNN14 benefiting from higher capacity in some metrics.

The comparison with a CNN10 trained from scratch highlights the importance of transfer learning. The pretrained CNN10 outperformed the ran-

domly initialized version, showing that large-scale audio pretraining is helpful when training data is limited. Even when the downstream task differs from the original objective, pretraining improves performance.

The fairness analysis showed differences across age groups, with lower accuracy for younger participants. This may be related to changes in speech with age and to data imbalance between groups. Performance across sex was more stable, although the small number of male participants limits stronger conclusions.

9 Conclusion

This work studied the use of pretrained neural audio models for asthma detection from voice and speech. By evaluating CNN10 and CNN14 models on vowel and read speech recordings, we showed that pretrained models perform better than models trained from scratch, highlighting the importance of transfer learning. The results also indicate that read speech provides stronger information for asthma detection than sustained vowels.

In addition, we analyzed performance at the patient level and across demographic groups. The results were stable across sex but varied across age groups, which emphasizes the importance of considering demographic factors in biomedical audio models.

Overall, the results confirm that pretrained audio models are a strong and effective approach for asthma detection from voice and speech. This work supports the potential of audio-based methods as a non-invasive tool for respiratory classification and motivates future studies with larger and more diverse datasets.

Limitations

This study has limitations. The dataset is relatively small and demographically imbalanced, which may limit generalization. We also consider only binary asthma classification, without severity or temporal modeling.

Ethical Considerations

This work uses voice and speech data from human participants collected with informed consent and handled in accordance with privacy and ethical guidelines. The work was approved by the Ethics Committee of Hospital (ommitted due to blind review). The data were anonymized, and the models are intended for research purposes only, not for clinical decision-making. We also examined demographic effects to help identify potential biases and support more responsible model development.

Acknowledgments

his study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001. The work was conducted at the Center for Artificial Intelligence (C4AI-USP), with support from the University of São Paulo and the São Paulo Research Foundation (FAPESP) under grant #2019/07665-4.

Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) through grants 2023/00488-5 (SPIRA-BM) and 2022/11254-2 (EMU), and by the National Council for Scientific and Technological Development (CNPq) under grant PQ1 302963/2022-7.

The authors used generative AI tools to assist with writing (paraphrasing and language refinement) and code development. All AI-generated suggestions were verified and approved by the authors.

References

Kawther S Alqudaihi, Nida Aslam, Irfan Ullah Khan, Abdullah M Almuhaideb, Shikah J Alsunaidi, Nehad M Abdel Rahman Ibrahim, Fahd A Alhaidari, Fatema S Shaikh, Yasmine M Alsenbel, Dima M Alalharith, and 1 others. 2021. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. *Ieee Access*, 9:102327–102344.

E. Casanova, L. Gris, A. Camargo, D. da Silva, M. Gazzola, E. Sabino, A. Levin, A. Candido Jr, S. Aluisio, and M. Finger. 2021. Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Online)*, pages 625–633. Association for Computational Linguistics.

Bradley E Chipps, Weily Soong, Reynold A Panettieri Jr, Warner Carr, Hitesh Gandhi, Wenjiong Zhou, Bill Cook, Jean-Pierre Llanos, and Christopher S Ambrose. 2023. Number of patient-reported asthma triggers predicts uncontrolled disease among specialist-treated patients with severe asthma. *Annals of Allergy, Asthma & Immunology*, 130(6):784–790.

Tusar Kanti Dash, Chinmay Chakraborty, Subhadip Mahapatra, and Ganapati Panda. 2022. Gradient boosting machine and efficient combination of features for speech-based detection of covid-19. *IEEE Journal*

of Biomedical and Health Informatics, 26(11):5364–5371.

Konstantinos P Exarchos, Maria Beltsiou, Chainti-Antonella Votti, and Konstantinos Kostikas. 2020. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. *European Respiratory Journal*, 56(3).

Anne Fuhlbrigge, David Peden, Andrea J Apter, Homer A Boushey, Carlos A Camargo Jr, James Gern, Peter W Heymann, Fernando D Martinez, David Mauger, William G Teague, and 1 others. 2012. Asthma outcomes: exacerbations. *Journal of Allergy and Clinical Immunology*, 129(3):S34–S48.

Marcelo Matheus Gauy, Larissa Cristina Berti, Arnaldo Cândido, Augusto Camargo Neto, Alfredo Goldman, Anna Sara Shafferman Levin, Marcus Martins, Beatriz Raposo-de Medeiros, Marcelo Queiroz, Ester Cerdeira Sabino, Flaviane Romani Fernandes Svartman, and Marcelo Finger. 2023. Discriminant audio properties in deep learning based respiratory insufficiency detection in brazilian portuguese. In *Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine, AIME 2023, Portorož, Slovenia, June 12–15, 2023, Proceedings*, page 271–275, Berlin, Heidelberg. Springer-Verlag.

Marcelo Matheus Gauy and Marcelo Finger. 2022. Pretrained audio neural networks for speech emotion recognition in portuguese. In *First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech Speech emotion recognition in Portuguese (SER 2022)*. CEUR-WS.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Murat Gezer, Mehmet Atilla Uysal, Neval Alagöz, Can Ortak, and Elif Yelda Niksarlıoğlu. 2025. Voices from the lungs: An innovative approach to asthma diagnosis using machine learning. *Acta Infologica*, 9(1):223–252.

Global Initiative for Asthma. 2024. Global strategy for asthma management and prevention. https://ginasthma.org.

M. A. Iqbal, Krishnamoorthy Devarajan, and Syed Musthak Ahmed. 2022. Real time detection and forecasting technique for asthma disease using speech signal and denn classifier. *Biomedical Signal Processing and Control*, 76:103637.

Md. Asim Iqbal, K. Devarajan, R. Lakshman Naik, and R. Sushmitha. 2024. Asthma detection from speech signals. In *Futuristic Trends in IoT*, volume 3 of *IIP Series*. IIP Series.

Panagiotis Kapetanidis, Fotios Kalioras, Constantinos Tsakonas, Pantelis Tzamalis, George Kontogiannis, Theodora Karamanidou, Thanos G Stavropoulos, and Sotiris Nikoletseas. 2024. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. *Sensors*, 24(4):1173.

Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. 2022. An overview of automatic speech recognition preprocessing techniques. In *2022 international conference on decision aid sciences and applications (DASA)*, pages 804–809. IEEE.

Sang Mee Lee, Hyein Ryu, Sunga Kong, Sun Hye Shin, Wooseong Huh, Myung Jin Chung, Juhee Cho, Taeyoung Kim, and Hye Yun Park. 2025. Voice as a digital biomarker: Foundation model-based copd assessment. *Research Square*. Preprint, under review at npj Digital Medicine.

Zhi Qing Looi, Zi Hao Ng, Rui Xiang Yak, Oren Rosen, and Anurag Kumar. 2024. Melodica ai-machine learning based detection of asthma via vocal audio analysis. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 905–910. IEEE.

Yi Lyu, Quan-Cheng Jiang, Shuai Yuan, Jing Hong, Chun-Feng Chen, Hai-Mei Wu, Yi-Qin Wang, Yu-Jing Shi, Hai-Xia Yan, and Jin Xu. 2025. Non-invasive acoustic classification of adult asthma using an xgboost model with vocal biomarkers. *Scientific Reports*, 15(1):28682.

Marcelo Matheus Gauy, Natália Hitomi Koza, Ricardo Mikio Morita, Gabriel Rocha Stanzione, Arnaldo Cândido Júnior, Larissa Cristina Berti, Anna Sara Shafferman Levin, Ester Cerdeira Sabino, Flaviane Romani Fernandes Svartman, and Marcelo Finger. 2026. Contrasting deep learning audio models for direct respiratory insufficiency detection versus blood oxygen saturation estimation. *Intelligence-Based Medicine*, 13:100331.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. 2025. Towards pre-training an effective respiratory audio foundation model. *arXiv preprint arXiv:2505.15307*.

Gaurav Pandey, Om P Pandey, Angela J Rogers, Mehmet E Ahsen, Gabriel E Hoffman, Benjamin A Raby, Scott T Weiss, Eric E Schadt, and Supinda Bunyavanich. 2018. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal rna sequence data. *Scientific reports*, 8(1):8826.

Erwan Pépiot. 2015. Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, HS-16.

Yasir Rahmatallah, Aaron S. Kemp, Anu Iyer, Lakshmi Pillai, Linda J. Larson-Prior, Tuhin Virmani, and Fred Prior. 2025. Pre-trained convolutional neural networks identify parkinson's disease from spectrogram images of voice samples. *Scientific Reports*, 15(1):7337.

Om Prakash Singh, Ramaswamy Palaniappan, and MB Malarvili. 2018. Automatic quantitative analysis of human respired carbon dioxide waveform for asthma and non-asthma classification using support vector machine. *IEEE Access*, 6:55245–55256.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Marko Topalovic, Stefan Laval, Jean-Marie Aerts, Thierry Troosters, Marc Decramer, Wim Janssens, Belgian Pulmonary Function Study investigators, and 1 others. 2017. Automated interpretation of pulmonary function tests in adults with respiratory complaints. *Respiration*, 93(3):170–178.

Laura Verde, Giuseppe De Pietro, and Giovanna Sannino. 2023. Artificial intelligence techniques for the non-invasive detection of covid-19 through the analysis of voice signals. *Arabian Journal for Science and Engineering*, 48(8):11143–11153.

Shengyun Wei, Shun Zou, Feifan Liao, and 1 others. 2020. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of physics: Conference series*, 1453(1):012085.

Liang Xu, Lizhong Wang, Sijun Bi, Hanyue Liu, and Jing Wang. 2023. Semi-supervised sound event detection with pre-trained model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yun Xu and Royston Goodacre. 2018. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3):249–262.

Peisi Yin, Xiaoyu You, Xinyue Cui, Zhipeng Tang, Shanshan Yu, Huaian Fu, Fei Song, Kai Zhang, Xin Zhao, Lipeng Wang, and 1 others. 2025. Clinically diagnose asthma and monitor its severity using an ultrasensitive chemiresistive nitric oxide (no) gas sensor via exhaled breath analysis assisted by pattern recognition. *ACS sensors*.

Jun Zhan, Wen Chen, Longsheng Cheng, Qiong Wang, Feifei Han, and Yubao Cui. 2020. Diagnosis of asthma based on routine blood biomarkers using machine learning. *Computational intelligence and neuroscience*, 2020(1):8841002.