FINE-TUNING PRE-TRAINED AUDIO MODELS FOR COVID-19 DETECTION: A TECHNICAL REPORT

Daniel Oliveira de Brito
Institute of Biosciences, Languages, and Exact Sciences
São Paulo State University
São José do Rio Preto, Brazil
daniel.o.brito@unesp.br

Letícia Gabriella de Souza
Institute of Biosciences, Languages, and Exact Sciences
São Paulo State University
São José do Rio Preto, Brazil
leticia.gabriella@unesp.br

Marcelo Matheus Gauy
Institute of Natural Sciences and Engineering
São Paulo State University
Itapeva, Brazil
marcelo.gauy@unesp.br

Marcelo Finger
Institute of Mathematics and Statistics
University of São Paulo
São Paulo, Brazil
mfinger@ime.usp.br

Arnaldo Candido Junior
Institute of Biosciences, Languages, and Exact Sciences
São Paulo State University
São José do Rio Preto, Brazil
arnaldo.candido@unesp.br

November 20, 2025

ABSTRACT

This technical report investigates the performance of pre-trained audio models on COVID-19 detection tasks using established benchmark datasets. We fine-tuned Audio-MAE and three PANN architectures (CNN6, CNN10, CNN14) on the Coswara and COUGHVID datasets, evaluating both intra-dataset and cross-dataset generalization. We implemented a strict demographic stratification by age and gender to prevent models from exploiting spurious correlations between demographic characteristics and COVID-19 status. Intra-dataset results showed moderate performance, with Audio-MAE achieving the strongest result on Coswara (0.82 AUC, 0.76 F1-score), while all models demonstrated limited performance on Coughvid (AUC 0.58-0.63). Cross-dataset evaluation revealed severe generalization failure across all models (AUC 0.43-0.68), with Audio-MAE showing strong performance degradation (F1-score 0.00-0.08). Our experiments demonstrate that demographic balancing, while reducing apparent model performance, provides more realistic assessment of COVID-19 detection capabilities by eliminating demographic leakage - a confounding factor that inflate performance metrics. Additionally, the limited dataset sizes after balancing (1,219-2,160 samples) proved insufficient for deep learning models that typically require substantially larger training sets. These findings highlight fundamental challenges in developing generalizable audio-based COVID-19 detection systems and underscore the importance of rigorous demographic controls for clinically robust model evaluation.

1 Introduction

Audio-based detection of respiratory conditions using deep learning has emerged as a promising diagnostic approach, with numerous studies reporting high accuracy for COVID-19 detection from cough, voice, and speech recordings [1]. While these advances have demonstrated the potential of acoustic biomarkers for respiratory assessment, important questions remain about the relationship between different respiratory diagnostic tasks—particularly whether models

trained for COVID-19 detection can inform broader respiratory conditions such as respiratory insufficiency (RI) or blood oxygen saturation (SpO2) estimation.

Our recent work [1] explored this question by applying pre-trained audio models to both respiratory insufficiency detection and SpO2 estimation tasks. We found that while models achieved near-perfect accuracy (99.9%) for RI detection, the same architectures struggled significantly with SpO2 estimation, achieving less than 70% accuracy. This performance disparity suggests that the acoustic features indicative of respiratory insufficiency may be fundamentally different from those needed to estimate blood oxygenation levels, raising important questions about the specificity and generalizability of audio-based respiratory diagnostics.

This technical report investigates how the same pre-trained models—Audio-MAE [2] and PANNs (CNN6, CNN10, CNN14) [3]—perform on established COVID-19 detection benchmarks. We evaluate these models on two widely-used datasets: Coswara [4] and Coughvid [5], conducting both intra-dataset and cross-dataset experiments.

Our experiments examine whether models can generalize across different COVID-19 datasets and how demographic factors affect reported performance. These findings contribute to understanding the capabilities and limitations of audio-based respiratory assessment, which is essential for developing reliable clinical applications.

2 Related Work

Islam et al. [6] conducted a comprehensive study using five datasets, including Coswara and Coughvid. Their approach employed an extensive range of techniques: threshold moving for optimal threshold selection based on ROC-AUC, feature dimension reduction using RFECV with Extra-Trees classifier, Bayesian Optimization for hyperparameter tuning, and SMOTE for handling class imbalance. For the Coswara dataset, samples were preprocessed and labeled based on healthy (COVID-19 negative) and heavy cough variations (COVID-19 positive). For the Coughvid dataset, the authors preprocessed and labeled samples into two groups: those from healthy individuals (COVID-19 negative) and those from individuals with notable cough variations (COVID-19 positive). Notably, no mention was found regarding demographic balancing (age, gender) in their train/test splits.

Table 1: Intra-dataset performance on Coswara dataset from literature

Study	AUC	Precision	Recall
Zhang et al. [7]	0.86	–	0.60
Sobahi et al. [8]	–	0.90	0.88
Chowdhury et al. [9]	0.66	0.76	0.47
Islam et al. (DNDT) [6]	0.84	0.52	0.80
Islam et al. (DNDF) [6]	**0.92**	0.72	**0.93**

Table 2: Intra-dataset performance on Coughvid dataset from literature

Study	AUC	Precision	Recall
Orlandic et al. [10]	0.88	–	–
Sunitha et al. [11]	–	0.76	0.77
Hamdi et al. [12]	0.91	0.91	0.90
Hamdi et al. [13]	0.91	0.95	0.86
Aytekin et al. [14]	0.83	0.77	1.00
Pavel and Ciocoiu [15]	0.76	0.69	0.68
Islam et al. (DNDT) [6]	0.81	0.83	0.79
Islam et al. (DNDF) [6]	**0.93**	**0.93**	**0.94**

Table 3: Cross-dataset performance from Islam et al. [6]

Training → Test	AUC	Precision	Recall	F1-Score
Coswara → Coughvid	0.53	0.57	0.24	0.33
Coughvid → Coswara	0.57	0.17	0.85	0.28

Islam et al. achieved 0.92 of ROC-AUC on Coswara (Table 1) and 0.93 on Coughvid (Table 2). The models trained on Coswara and tested on Coughvid had a AUC of to 0.53 with an F1-Score of 0.33 (Table 3). Similarly, models trained on

Coughvid and tested on Coswara achieved an AUC of 0.57 with an F1-Score of 0.28, despite maintaining high recall (0.85) at the cost of precision (0.17). This substantial performance drop in cross-dataset scenarios highlights the limited generalization capability of current COVID-19 detection models across different audio datasets.

3 Methodology

3.1 Datasets

3.1.1 Coswara Dataset

The Coswara dataset [4] is a publicly available COVID-19 audio dataset developed by the Indian Institute of Science. We utilized version 1.0.0 of this dataset, focusing specifically on cough-heavy audio recordings for COVID-19 detection.

Preprocessing The original dataset contained 2,746 unique participant IDs with cough-heavy audio recordings. Our preprocessing pipeline involved the following steps:

1. **Audio Quality Filtering**: We removed 43 recordings with zero duration, ensuring only valid audio samples were retained.
2. **Label Mapping**: To create a binary classification task, we consolidated the COVID status labels as follows:
   - COVID-positive class: combined *positive moderate*, *positive mild*, and *positive asymptomatic* categories
   - COVID-negative class: *healthy* participants only
   - Excluded categories: *no respiratory illness exposed*, *recovered full*, *respiratory illness not identified*, and *under validation*
3. **Metadata Completeness**: We excluded participants with missing demographic information (age or gender), ensuring robust stratification capabilities.

After preprocessing, the dataset comprised 2,083 participants with confirmed COVID-positive or COVID-negative status.

Table 4 presents the demographic distribution of the filtered dataset before balancing. The dataset exhibited significant class imbalance, with COVID-negative samples substantially outnumbering COVID-positive samples across most demographic groups.

Table 4: Dataset Balancing: Original vs. Stratified Undersampled Counts for Coswara

Age Group	Gender	COVID-19 (Template)	Healthy (Unbalanced)	Healthy (Balanced)
0–17	Female	7	7	7
	Male	10	20	10
18–29	Female	90	182	90
	Male	135	470	135
30–39	Female	54	78	54
	Male	97	297	97
40–49	Female	30	47	30
	Male	61	131	61
50–59	Female	43	30	30
	Male	56	102	56
60+	Female	43	12	12
	Male	51	28	28
	Other	1	1	1
Total Samples		**648**	**1,435**	**571**

To address the class imbalance and ensure fair model training, we implemented a stratified undersampling approach. The balancing strategy preserved the demographic distribution of the minority class (COVID-positive) while undersampling the majority class (COVID-negative) to achieve equal representation across age and gender strata.

Balancing Methodology For each demographic stratum (age group $\times$ gender combination), we:

1. Retained all COVID-positive samples
2. Randomly undersampled COVID-negative samples to match the COVID-positive count within that stratum
3. For strata where COVID-negative samples were fewer than COVID-positive samples (e.g., 50–59 females, 60+ age groups), all available COVID-negative samples were retained

Table 4 presents the final balanced dataset distribution, which maintains demographic representativeness while addressing class imbalance.

Final Dataset Characteristics The balanced dataset consists of 1,219 samples (571 COVID-negative and 648 COVID-positive), with slight overall class imbalance retained in older age groups where COVID-positive cases were more prevalent. This approach ensures that the model training reflects realistic demographic patterns while mitigating severe class imbalance issues.

3.1.2 Coughvid

Coughvid[5] is a public dataset with crowdsourced cough recordings of more than 30.000 recordings collected by the École Polytechnique Fédérale de Lausanne (EPFL) to help develop research on audio-based respiratory disease diagnosis such as COVID-19 and the dataset contains various features such as ages, genders, geographic locations, and COVID-19 status.

The dataset is unbalanced with a considerable number of audio samples for `healthy` and `symptomatic` compared to `covid-19 positive`. Furthermore, a considerable amount of the recordings are missing demographic metadata. In order to create a robust and unbiased dataset for the models trainings a preprocessing and balancing methodology was executed. Coughvid dataset contains a total of 34.434 audio samples. However, it presents two major challenges for these type of complex deep learning models such as PANNs:

1. **Missing Metadata:** A significant amount of audio samples metadata are missing essential labels. As shown in Table 5, 13.770 audio samples (39.9%) have no healthy, COVID-19 or symptomatic label information.
2. **Class Unbalance:** For the labeled data, it also shows unbalance data. The `healthy` class with 15.476 audio samples and the `COVID-19` class with 1.315 audio samples.

Table 5: Class Distribution in the Coughvid Dataset.

Status	**Audio Sample Count**	**Percentage**
`healthy`	15.476	44.9%
`symptomatic`	3.873	11.2%
`COVID-19`	1.315	3.8%
`NaN (Missing)`	13.770	39.9%
Total	**34.434**	**100.0%**

Data Preprocessing and Balancing Our preprocessing pipeline consisted of the following steps:

1. **Data Filtering:** The dataset was first filtered to select only audio samples that contained valid self-reported `age` and `gender` metadata. This step is crucial for demographic analysis and stratified sampling.
2. **Class Selection:** We isolated the two primary target classes for this report: `COVID-19` as (COVID positive), `healthy` as (COVID negative) and `symptomatic` was completely ignored and excluded.
3. **Undersampling:** To address the class unbalance, we implemented a stratified undersampling strategy. The `covid-19` class was identified as the minority class. The `healthy` classe were then undersampled to match the exact demographic distribution by age and gender strata of the `covid-19` class.
4. **Random Window Sampling**: A fixed windowing of 4 seconds was applied in the audio samples.

After preprocessing the dataset for available age and gender the `covid-19` minority class consisted of 1.080 audio samples, so for `covid-19` and `healthy` classes with a total of 2.160 audio samples. The specific demographic distribution of `covid-19` class was used as a template for undersampling the `healthy` class which is detailed in Table 6.

Table 6: Dataset Balancing: Original vs. Stratified Undersampled Counts.

Age	Gender	COVID-19 (Template)	Healthy (Unbalanced)	Healthy (Balanced)
0–17	female	65	314	65
	male	67	563	67
18–29	female	126	1.264	126
	male	216	2.811	216
30–44	female	120	1.490	120
	male	203	3.613	203
45–59	female	78	1.024	78
	male	130	2.005	130
60+	female	30	533	30
	male	45	825	45
Total Samples		**1.080**	**14.442**	**1.080**

3.2 Models

3.2.1 Audio-MAE

Audio-MAE (Audio Masked Autoencoders) [2] extends the Masked Autoencoder framework to audio spectrograms using a Vision Transformer encoder-decoder architecture. The model employs asymmetric processing: during pre-training, the encoder operates on only 20% of visible spectrogram patches (with 80% masked), while a decoder with local attention mechanisms reconstructs the complete input by minimizing mean squared error on masked regions.

Unlike image-based MAE, Audio-MAE incorporates local attention in the decoder to better capture the localized time-frequency correlations characteristic of audio spectrograms. Pre-trained on AudioSet-2M (near 2 million samples), the model achieved state-of-the-art results on speech classification tasks with 98.3% accuracy on Speech Commands V2 (84k samples) and 94.8% on VoxCeleb (138k samples).

3.2.2 PANN

based on the use of Pre-trained Audio Neural Networks (PANNs) [3], deep Convolutional Neural Network (CNN) architectures available in the `audioset_tagging_cnn` repository [16]. The fine-tuning strategy and training pipeline were adapted from reference implementations, such as those found in the `audio-pann-train` repository [17], and customized for this specific task as described below.

3.2.3 Model Architecture: PANNs (Pre-trained Audio Neural Networks)

The backbone of our model is an architecture from the PANNs family [3], originating from Qiuqiang Kong's repository [16]. PANNs are a set of deep CNNs pre-trained on the task of audio event tagging using the massive **AudioSet** dataset [18], which contains over 5.000 hours of audio and 527 sound classes.

This pre-training endows the model with a robust ability to extract hierarchical and generalizable acoustic features, ranging from simple textures (early layers) to complex sound patterns (deep layers). The [16] repository offers several architectures; the most notable are CNN6, CNN10, and CNN14.

CNN6, CNN10, and CNN14: The number in each architecture's name refers to the number of convolutional layers in the network.

- **CNN6:** A lighter architecture with 6 convolutional layers. It is faster for inference but has a smaller representational capacity (fewer parameters).
- **CNN10:** An intermediate model, composed of 4 convolutional blocks and 2 linear layers in the backbone, totaling 10 convolutional layers.
- **CNN14:** The deepest architecture. It is composed of 6 convolutional blocks and 2 linear layers in the backbone, totaling 14 convolutional layers. Its depth allows for a larger receptive field over the spectrogram and the ability to learn more complex and abstract features, which led to its selection for this task.

Table 7: Intra-Dataset Performance

Dataset/Model	Acc	AUC	Prec	Rec	F1
Coughvid					
Audio-MAE	0.57	0.60	0.58	0.55	0.56
CNN6	0.63	0.63	0.62	0.62	0.67
CNN10	0.59	0.62	0.59	0.56	0.63
CNN14	0.59	0.58	0.58	0.58	0.65
Coswara					
Audio-MAE	**0.75**	**0.82**	**0.77**	**0.75**	**0.76**
CNN6	0.66	0.61	0.57	0.56	0.73
CNN10	0.65	0.33	0.50	0.39	0.66
CNN14	**0.75**	**0.75**	0.67	0.68	**0.80**

The CNN6, CNN10, and CNN14 models was specifically chosen, loaded with the pre-trained weights (`Cnn6_mAP=0.343.pth`, `Cnn10_mAP=0.380.pth`, and `Cnn14_16k_mAP=0.438.pth`), which were optimized for 16 kHz audio.

3.2.4 Fine-Tuning Strategy

The fine-tuning process involved adapting the pre-trained CNN6,CNN10 and CNN14 backbone for the specific COVID-19 classification task. This adaptation was twofold: architectural and parametric. Architecturally, the original 527-class classifier was removed and replaced with a new classification head. This head consists of a two-layer MLP which projects the 2048-dimensional backbone embedding to a 256-dimensional latent space (using `ReLU` activation and `Dropout` with $p = 0.5$ for regularization), followed by a final linear layer mapping to the $COVID - negative$ and $COVID - positive$ target classes. A selective fine-tuning strategy was implemented by freezing the entire backbone (`freeze_panns=True`) except for the last two convolutional blocks (`conv_block6` and `conv_block5`). This approach allows the model to adapt its high-level feature extractors to the new task while preserving the robust, low-level representations learned from AudioSet, thus mitigating catastrophic forgetting.

The optimization was conducted using the `Adam` optimizer [19] with differential learning rates to account for the different initialization states of the parameters. The randomly initialized classification head (`fc1`, `fc2`) was trained with a relatively high learning rate of 5×10^{-4} for rapid convergence. In contrast, the pre-trained, unfrozen backbone layers were fine-tuned with a very conservative learning rate of 5×10^{-6} for gentle adjustment. The model was trained using `nn.CrossEntropyLoss` for a maximum of 100 epochs. To prevent overfitting and select the best performing model, an early stopping mechanism was employed, monitoring the validation accuracy (`eval_acc`) with a patience of 50 epochs.

3.3 Experimental Setup

All audio samples were processed using a fixed 4-second windowing strategy to standardize input lengths across both datasets. During training, random 4-second windows were extracted from each audio sample to provide data augmentation and reduce overfitting. During evaluation, systematic non-overlapping 4-second windows were extracted from each sample, with predictions aggregated to obtain a final per-sample classification decision.

For both Coswara and Coughvid datasets, we employed an 80/20 train-test split, stratified by COVID-19 status to maintain class balance across partitions. The Audio-MAE model was fine-tuned for 60 epochs with a learning rate of 2e-4. Similarly, the PANN models (CNN6, CNN10, and CNN14) were fine-tuned for [100] epochs with a learning rate of [5e-4].

Model performance was evaluated using multiple complementary metrics: accuracy, AUC-ROC, precision, recall and F1-score. Confusion matrices were computed for each model-dataset combination to enable detailed analysis of classification errors and facilitate comparison with related work in the literature.

4 Experiments and Results

4.1 Intra-dataset

Table 7 presents the intra-dataset performance of all models trained and evaluated on the same dataset. The results demonstrate considerable variability across models and datasets, with Audio-MAE achieving the strongest performance on Coswara (AUC 0.82, F1-score 0.76), while PANNs showed more balanced but moderate performance across both datasets.

4.1.1 Coswara Dataset Results

On the Coswara dataset, Audio-MAE achieved the highest AUC (0.82) and precision (0.77), indicating strong discriminative capability and reliable positive predictions. CNN14 matched Audio-MAE's accuracy (0.75) and achieved a competitive F1-score (0.80), though with lower AUC (0.75). CNN10 showed the poorest performance with notably low AUC (0.33) and recall (0.39), suggesting difficulty in identifying COVID-19 positive cases. CNN6 demonstrated intermediate performance across all metrics.

The confusion matrices for Coswara (Table 8 and Table 9) reveal varying error patterns across models. Audio-MAE achieved the most balanced confusion matrix with 102 true positives and 92 true negatives, though it still misclassified 34 COVID-positive cases as negative. CNN14 showed high specificity (92 true negatives) but poor sensitivity (only 22 true positives), indicating a strong bias toward predicting the negative class. CNN10 failed completely on Coswara, predicting all samples as negative (0 true positives, 53 false negatives). CNN6 demonstrated intermediate performance with better balance than CNN14 but worse than Audio-MAE.

4.1.2 Coughvid Dataset Results

On the Coughvid dataset, all models demonstrated more modest performance compared to Coswara, with AUC scores ranging from 0.58 to 0.63. CNN6 achieved the highest accuracy (0.63) and AUC (0.63), followed closely by CNN10 (AUC 0.62). Audio-MAE and CNN14 showed similar performance (AUC 0.58-0.60), suggesting that the Coughvid dataset may not provide sufficient discriminative features or that demographic diversity poses additional challenges.

The confusion matrices (Table 8 and Table 9) show more balanced but still suboptimal classification patterns on Coughvid. Audio-MAE achieved 118 true positives and 129 true negatives, representing the most balanced performance despite modest AUC. CNN14 showed 60 true positives versus 57 true negatives with relatively balanced error distribution. CNN6 achieved 69 true positives and 56 true negatives, while CNN10 demonstrated severe imbalance with only 33 true positives against 84 true negatives.

Table 8: Consolidated Confusion Matrix Results for CNN 14, CNN10 and CNN6

	CNN14				**CNN10**				**CNN6**			
Datasets(Train x Test)	**TP**	**FN**	**FP**	**TN**	**TP**	**FN**	**FP**	**TN**	**TP**	**FN**	**FP**	**TN**
CoughVid x CoughVid	60	41	42	57	33	68	15	84	69	32	43	56
CoughVid x Coswara	27	26	50	49	16	37	44	55	23	30	46	53
Coswara x Coswara	22	31	7	92	0	53	0	99	13	40	11	88
Coswara x CoughVid	24	77	12	87	19	82	4	95	25	76	14	85

Table 9: Consolidated Confusion Matrix Results for Audio-MAE

Datasets (Train × Test)	**TP**	**FN**	**FP**	**TN**
CoughVid × CoughVid	118	98	87	129
CoughVid × Coswara	0	136	3	119
Coswara × Coswara	102	34	30	92
Coswara × CoughVid	10	206	12	204

4.2 Cross-dataset evaluation

Table 10 presents the cross-dataset performance, revealing severe degradation when models are evaluated on datasets different from their training data. This finding aligns with previous literature [6] and highlights fundamental challenges in developing generalizable audio-based COVID-19 detection systems.

4.2.1 Models Trained on Coughvid, Tested on Coswara

When trained on Coughvid and tested on Coswara, Audio-MAE completely failed to generalize, with undefined accuracy and all metrics at or near zero. The confusion matrix in Table 8

The PANN models showed marginally better but still poor generalization. CNN14 achieved the highest cross-dataset performance in this direction (AUC 0.50, F1-score 0.49), essentially performing at chance level. CNN6 and CNN10 performed similarly (AUC 0.43-0.49), with confusion matrix in Table 8.

Table 10: Cross-Dataset Generalization Performance

Train/Model	Acc	AUC	Prec	Rec	F1
Tested on Coswara					
Trained on Coughvid:					
Audio-MAE	–	0.47	0.00	0.00	0.00
CNN6	0.50	0.49	0.48	0.48	0.44
CNN10	0.47	0.43	0.43	0.43	0.43
CNN14	0.50	0.50	0.50	0.49	0.49
Tested on Coughvid					
Trained on Coswara:					
Audio-MAE	0.50	0.51	0.45	0.05	0.08
CNN6	0.55	0.58	0.55	0.51	0.56
CNN10	0.57	**0.68**	0.57	0.50	**0.59**
CNN14	0.56	0.60	0.56	0.51	**0.59**

When trained on Coughvid and tested on Coswara, Audio-MAE failed to generalize. The confusion matrix in Table 9 shows this failure: 0 true positives, 136 false negatives, with only 3 false positives and 119 true negatives, indicating the model predicted nearly all Coswara samples as COVID-negative.

4.2.2 Models Trained on Coswara, Tested on Coughvid

The reverse direction (training on Coswara, testing on Coughvid) yielded slightly better but still inadequate results. CNN10 achieved the best cross-dataset performance with AUC 0.68 and F1-score 0.59, followed by CNN14 (AUC 0.60, F1-score 0.59) and CNN6 (AUC 0.58, F1-score 0.56). Audio-MAE again showed severe generalization failure with AUC 0.51 and F1-score 0.08, demonstrating extreme bias toward the negative class as evidenced by recall of only 0.05.

Training on Coswara and testing on Coughvid yielded slightly better but still inadequate results. CNN10 achieved the best cross-dataset performance with AUC 0.68 and F1-score 0.59, followed by CNN14 (AUC 0.60, F1-score 0.59) and CNN6 (AUC 0.58, F1-score 0.56). Audio-MAE again showed severe generalization failure (Table 9) with only 10 true positives against 206 false negatives (recall 0.05), demonstrating extreme bias toward the negative class.

5 Discussion

5.1 Impact of Demographic Stratification on Model Performance

A critical methodological distinction between our work and previous studies is our approach to dataset balancing. Most prior works, including Islam et al. [6], achieved high intra-dataset performance (AUC 0.92 on Coswara, 0.93 on Coughvid) without explicit demographic stratification in their train/test splits. While these results appear impressive, they may be confounded by demographic biases present in the unbalanced datasets.

Our analysis of the raw Coswara dataset (Table 4) and Coughvid (Table 5.1) reveals substantial demographic imbalances. Without stratified sampling, models may inadvertently learn to exploit acoustic properties associated with age and gender distributions rather than COVID-19-specific respiratory features.

To investigate this hypothesis, we conducted two additional experiments: fine-tuning Audio-MAE on the unbalanced Coswara and Coughvid datasets, without demographic stratification. This approach resulted in improved AUC compared to our demographically balanced split supporting our concern about demographic leakage: On Coswara, we went from 0.82 to 0.85 of AUC (3.7% increase change); on Coughvid, from 0.60 of AUC to 0.63 (4.5% change). The model likely leveraged spurious correlations between age/gender acoustic characteristics and COVID-19 labels present in the unbalanced datasets, achieving higher performance metrics that do not reflect true COVID-19 detection capability.

This finding has important implications for interpreting results from the literature. The high intra-dataset performance reported in previous studies may partially reflect the models ability to discriminate demographic groups rather than COVID-19 status. This demographic leakage would also explain the severe performance degradation observed in cross-dataset scenarios (our results: AUC 0.46-0.51; Islam et al.: AUC 0.53-0.57), where the demographic distributions differ between training and testing datasets.

Our decision to employ stratified undersampling, while resulting in lower intra-dataset performance, ensures that models cannot exploit demographic shortcuts. This approach provides a more realistic assessment of COVID-19 detection capabilities and better reflects the challenge of developing clinically deployable systems that must generalize across diverse patient populations.

5.2 Dataset Size as a Limiting Factor

The modest performance of Audio-MAE on COVID-19 detection tasks can be primarily attributed to the limited size of available training data. After demographic balancing, our datasets contain only 1,219 samples (Coswara) and 2,160 samples (Coughvid)—approximately 40-70 times smaller than the datasets where Audio-MAE demonstrated strong performance (84k-138k samples for speech tasks).

Comparing with Table 1, we see that Islam et al. [20] achieved superior results (AUC 0.92 on Coswara) using traditional machine learning with extensive feature engineering, threshold optimization, bayesian hyperparameter tuning, and SMOTE for class balancing.

These techniques may be more suitable for small-scale datasets, whereas deep learning models like Audio-MAE typically require larger datasets to reach their full potential.

5.3 Generalization Challenges

The poor cross-dataset performance (AUC 0.46-0.51) indicates that models trained on such limited data fail to learn generalizable features. This aligns with findings from Islam et al. [20], who reported similar degradation in cross-dataset scenarios despite achieving high intra-dataset performance with traditional ML approaches and extensive optimization.

6 Conclusion

This technical report evaluated pre-trained audio models Audio-MAE [2] and PANNs [3] on COVID-19 detection using the Coswara [4] and Coughvid [5] datasets. Intra-dataset performance varied, with Audio-MAE achieving the strongest result on Coswara (0.82 AUC), while all models showed modest and limited performance on Coughvid (AUC 0.58–0.63). The most critical finding was a severe performance degradation in cross-dataset evaluation. This finding aligns with previous literature, such as Islam et al. [6], and highlights a fundamental failure to generalize. Audio-MAE, in particular, failed generalization tests, yielding an F1-score of just 0.08 when trained on Coswara and tested on Coughvid.

A primary conclusion is that demographic stratification significantly impacts reported performance. This work utilized strict stratified undersampling by age and gender to prevent models from exploiting demographic shortcuts. The high performance (e.g., 0.92–0.93 AUC) reported in such studies [6] may be inflated by this *demographic leakage*. Furthermore, the modest performance of these deep learning models is attributed to the limited size of the balanced datasets (1.219 - 2.160 samples), which is insufficient compared to the large-scale data (84k - 138k samples) on which models like Audio-MAE typically excel.

References

[1] Marcelo Matheus Gauy, Natalia Hitomi Koza, Ricardo Mikio Morita, Gabriel Rocha Stanzione, Arnaldo Candido Junior, Larissa Cristina Berti, Anna Sara Shafferman Levin, Ester Cerdeira Sabino, Flaviane Romani Fernandes Svartman, and Marcelo Finger. Contrasting deep learning models for direct respiratory insufficiency detection versus blood oxygen saturation estimation, 2024.

[2] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

[3] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[4] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R, Prasanta Kumar Ghosh, and Sriram Ganapathy. Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Interspeech 2020*, pages 4811–4815, October 2020. arXiv:2005.10548 [eess].

[5] Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, 2021.

[6] Rofiqul Islam, Nihad Karim Chowdhury, and Muhammad Ashad Kabir. Robust covid-19 detection from cough sounds using deep neural decision tree and forest: A comprehensive cross-datasets evaluation, 2025.

[7] Xueshuai Zhang, Jiakun Shen, Jun Zhou, Pengyuan Zhang, Yonghong Yan, Zhihua Huang, Yanfen Tang, Yu Wang, Fujie Zhang, Shaoxing Zhang, and Aijun Sun. Robust Cough Feature Extraction and Classification Method for

COVID-19 Cough Detection Based on Vocalization Characteristics. In *Interspeech 2022*, pages 2168–2172. ISCA, September 2022.

[8] Nebras Sobahi, Orhan Atila, Erkan Deniz, Abdulkadir Sengur, and U. Rajendra Acharya. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. *Biocybernetics and Biomedical Engineering*, 42(3):1066–1080, July 2022.

[9] Nihad Karim Chowdhury, Muhammad Ashad Kabir, Md. Muhtadir Rahman, and Sheikh Mohammed Shariful Islam. Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Computers in Biology and Medicine*, 145:105405, June 2022.

[10] Lara Orlandic, Tomas Teijeiro, and David Atienza. A semi-supervised algorithm for improving the consistency of crowdsourced datasets: The COVID-19 case study on respiratory disorder classification. *Computer Methods and Programs in Biomedicine*, 241:107743, November 2023.

[11] Gurram Sunitha, Rajesh Arunachalam, Mohammed Abd-Elnaby, Mahmoud M. A. Eid, and Ahmed Nabih Zaki Rashed. A comparative analysis of deep neural network architectures for the dynamic diagnosis of COVID-19 based on acoustic cough features. *International Journal of Imaging Systems and Technology*, 32(5):1433–1446, September 2022.

[12] Skander Hamdi, Mourad Oussalah, Abdelouahab Moussaoui, and Mohamed Saidi. Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound. *Journal of Intelligent Information Systems*, 59(2):367–389, October 2022.

[13] Skander Hamdi, Abdelouahab Moussaoui, Mourad Oussalah, and Mohamed Saidi. Autoencoders and Ensemble-Based Solution for COVID-19 Diagnosis from Cough Sound, October 2022. Accepted: 2025-01-03T06:56:00Z Publisher: Springer.

[14] Idil Aytekin, Onat Dalmaz, Kaan Gonc, Haydar Ankishan, Emine Ulku Saritas, Ulas Bagci, Haydar Celik, and Tolga Cukur. COVID-19 Detection From Respiratory Sounds With Hierarchical Spectrogram Transformers. *IEEE journal of biomedical and health informatics*, 28(3):1273–1284, March 2024.

[15] Irina Pavel and Iulian B. Ciocoiu. Covid-19 detection from cough recordings using bag-of-words classifiers. *Sensors*, 23(11), 2023.

[16] Qiuqiang Kong et al. audioset_tagging_cnn. `https://github.com/qiuqiangkong/audioset_tagging_cnn`, 2019.

[17] Martin Hodges. audio-pann-train: Training and fine-tuning panns. `https://github.com/MartinHodges/audio-pann-train`, 2020.

[18] Jort F Gemmeke, Daniel PW Ellis, David Freedman, Aren Jansen, Wade Lawrence, R Chris Moore, Manoj Plakal, and Marvin Ritter. 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, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Rofiqul Islam, Nihad Karim Chowdhury, and Muhammad Ashad Kabir. Robust COVID-19 Detection from Cough Sounds using Deep Neural Decision Tree and Forest: A Comprehensive Cross-Datasets Evaluation, January 2025. arXiv:2501.01117 [cs].

[21] George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014.

[22] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014.

[23] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

[24] Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, June 2021. arXiv:2009.11644 [cs].

[25] Marcelo Matheus Gauy, Marcelo Finger, Flaviane Romani Fernandes Svartman, Ester Sabino, Anna Levin, Arnaldo Cândido Júnior, Larissa Cristina Berti, Gabriel Rocha Stanzione, Ricardo Mikio Morita, and Natália Hitomi Koza. Contrasting Deep Learning Audio Models for Direct Respiratory Insufficiency Detection Versus Blood Oxygen Saturation Estimation, June 2025.

[26] Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, Sadhana Gonuguntla, and Murali Alagesan. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection | Scientific Data, June 2023.

[27] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition, August 2020. arXiv:1912.10211 [cs].

[28] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked Autoencoders that Listen, January 2023. arXiv:2207.06405 [cs].

[29] Jobie Budd, Kieran Baker, Emma Karoune, Harry Coppock, Selina Patel, Richard Payne, Ana Tendero Cañadas, Alexander Titcomb, David Hurley, Sabrina Egglestone, Lorraine Butler, Jonathon Mellor, George Nicholson, Ivan Kiskin, Vasiliki Koutra, Radka Jersakova, Rachel A. McKendry, Peter Diggle, Sylvia Richardson, Björn W. Schuller, Steven Gilmour, Davide Pigoli, Stephen Roberts, Josef Packham, Tracey Thornley, and Chris Holmes. A large-scale and PCR-referenced vocal audio dataset for COVID-19. *Scientific Data*, 11(1):700, June 2024. Publisher: Nature Publishing Group.

[30] Harry Coppock, George Nicholson, Ivan Kiskin, Vasiliki Koutra, Kieran Baker, Jobie Budd, Richard Payne, Emma Karoune, David Hurley, Alexander Titcomb, Sabrina Egglestone, Ana Tendero Cañadas, Lorraine Butler, Radka Jersakova, Jonathon Mellor, Selina Patel, Tracey Thornley, Peter Diggle, Sylvia Richardson, Josef Packham, Björn W. Schuller, Davide Pigoli, Steven Gilmour, Stephen Roberts, and Chris Holmes. Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. *Nature Machine Intelligence*, 6(2):229–242, February 2024. Publisher: Nature Publishing Group.

[31] Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method - ScienceDirect.

[32] The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms | Scientific Data.

[33] Bagus Tris Atmaja, Zanjabila, Suyanto, and Akira Sasou. Cross-dataset COVID-19 Transfer Learning with Cough Detection, Cough Segmentation, and Data Augmentation, October 2022. arXiv:2210.05843 [eess].

[34] Rofiqul Islam, Nihad Karim Chowdhury, and Muhammad Ashad Kabir. Robust COVID-19 Detection from Cough Sounds using Deep Neural Decision Tree and Forest: A Comprehensive Cross-Datasets Evaluation, January 2025. arXiv:2501.01117 [cs].

[35] Nebras Sobahi, Orhan Atila, Erkan Deniz, Abdulkadir Sengur, and U. Rajendra Acharya. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. *Biocybernetics and Biomedical Engineering*, 42(3):1066–1080, 2022.

[36] Nihad Karim Chowdhury, Muhammad Ashad Kabir, Md. Muhtadir Rahman, and Sheikh Mohammed Shariful Islam. Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Computers in Biology and Medicine*, 145:105405, June 2022.

[37] Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms, September 2020.

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037, 2019.

[40] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, 2015.

[41] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[42] Neeraj Sharma, P Krishnan, R Kumar, S Ramoji, S R Chetupalli, R Nirmala, P K Ghosh, and Sriram Seshadri. Coswara—a database of breathing, cough, and voice sounds for covid-19 diagnosis. *arXiv preprint arXiv:2005.10548*, 2020.

[43] Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid dataset: A large-scale pre-processed and augmented dataset of coughs. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3361–3365, 2021.