CS610

Project Report

"You Sick Ah?" Audio-Based Respiratory Disease Detection Using Cough Sounds

Group 3

DAU VU DANG KHOI
LI XINJIE
SAMUEL LEE WEI SHENG
TAN ZONG YIN SHAWN
WU JIALU

Link to project (GitHub):
https://github.com/Jieoi/dual-channel-audio-fusion

05 April 2025

1. Introduction

Respiratory diseases are prevalent and require diagnostic methods that are scalable, accessible, and accurate. Traditional methods such as spirometry, chest X-rays, and oximetry have limitations including reliance on professionals, being dependent on specialised medical equipment, and being less effective at early stages of the disease. Current diagnostic tools are limited in their effectiveness due to scarcity of labelled data in healthcare-related use cases.

Based on these challenges, our project focuses on the following research questions:

1. Whether a machine learning model can accurately classify healthy vs. unhealthy individuals using respiratory audio; and
2. The effect of using synthetic audio data on model performance in the classification of respiratory disease vis-à-vis real audio data, to see if it can overcome the issue of limited labelled data.

A common marker of respiratory disease is the sound produced when a person is unwell—most notably, coughing. While untrained listeners may find it difficult to distinguish between healthy and unhealthy coughs, even medical professionals can struggle to make accurate assessments based solely on sound. This challenge presents a compelling opportunity to apply machine learning (ML), which can detect subtle and nuanced acoustic patterns that may elude human perception. Cough sound analysis is non-invasive, convenient for patients, and well-suited for large-scale screening as audio can be processed efficiently. Additionally, ML models can continually improve with access to more data and advancements in methodology over time.

As such, our project aims to deliver the following:

1. Develop a baseline ML model using expert-labelled cough data to classify healthy vs. unhealthy individuals; and
2. Compare model performance when the training set is expanded with real versus synthetic cough samples.

2. Literature Review

Cough sounds are valuable non-invasive biomarkers, reflecting changes in the respiratory tract. The COVID-19 pandemic spurred AI research using audio features as an additional support to diagnostic methods like PCR tests. In terms of analysing audio biomarkers, handcrafted features are commonly used, including MFCCs, which summarize spectral shape, and Log-Mel spectrograms, which visualize audio energy over time. These have been successfully applied to classifiers such as Support Vector Machines, Random Forests, and deep learning models (Erdogan & Narin, 2021; Sharma et al., 2020; 2023). These influenced our usage of MFCC and Log-Mel features as model inputs in our experiments.

The use of Convolutional Neural Networks (CNNs) is common in this domain, given its ability to effectively capture patterns across time and frequency dimensions. Due to data scarcity, the performance of standalone CNNs on relatively small datasets remain sub-optimal, with Sharma et al. (2020) reporting an accuracy of 66.7% on a dataset comprising 2800 samples. However, pretrained CNN models fine-tuned on COVID-19 cough data, have shown strong performance (Laguarta et al., 2020; Erdogan & Narin, 2021), while the usage of transfer learning further boosts this accuracy by leveraging unlabelled audio (e.g., cough, sneeze, speech) before fine-tuning on labelled data. Notably, applying transfer learning to ResNet50 improved AUC from 0.976 to 0.982, enhancing generalizability and reducing variance (Pahar et al., 2022). Given the effectiveness of CNN in this domain, we adopted a CNN-based architecture in our experiments.

With regards to the use of synthetic data, recent work has shown that it is able to provide promising improvements to model performance. For example, generative adversarial networks (GANs) were utilised to obtain synthetic samples for diseases such as asthma and COPD to help balance and increase the size of datasets, which led to improved classifier performance (Ramesh et al., 2020). A similar initiative was also undertaken improve speech-based diagnostics by enriching the dataset with varied and realistic waveforms (Reddy et al., 2025). These findings inspired our exploration into the use of GAN or other similar technologies to generate synthetic cough samples to augment our training set.

3. Dataset Exploration

We evaluated five publicly available datasets commonly used for cough audio classification and assessed them based on availability, label quality, audio quality, and sample size. Although the UK COVID-19 Vocal Audio dataset was large (72,999 participants) with verified labels, manual sampling revealed poor audio quality and models trained on it performed poorly despite preprocessing. Virufy and ComParE were well-cited but not readily accessible. In contrast, **COUGHVID** and **Coswara** met our criteria for availability and label reliability. Preliminary model training and testing on these datasets yielded performance close to reported benchmarks, leading to their selection. Full dataset comparisons are provided in Annex A.

4. Data Processing & Augmentation

4.1 Dataset: COUGHVID

We focused first on the COUGHVID dataset (Orlandic et al., 2021), which contained over 25,000 crowdsourced cough recordings. As majority of the data contained patients' self-reported conditions, which are not considered expert labelled and usable for ML, we focused on the 2,339 entries that had explicit expert labels by medical professionals, with 77.6% classified as unhealthy and 22.4% as healthy. This **imbalance** motivated the data expansion and augmentation strategies outlined later.

4.2 Data Processing Pipeline

Cough recordings are inherently short, non-verbal, and highly variable. To ensure reliable downstream analysis, a rigorous data preprocessing pipeline was applied:

1. **Filtering Short Samples**: Remove audio clips shorter than 0.1 seconds.
2. **Noise Filtering**: Calculate signal-to-noise ratio (SNR) and filter out samples with SNR below 10dB.
3. **Resampling**: Standardize audio sampling rate to 12kHz.
4. **Denoising**: Remove background noise using spectral noise estimation.
5. **Trimming**: Remove silent sections based on detection of low-energy regions.

This pipeline mitigates the effects of sub-optimal recording conditions, environmental noise, and inconsistencies in clip length.

4.3 Data Expansion Using Coswara

To address data scarcity and improve class balance, the training dataset was expanded using samples from the Coswara dataset (Bhattacharya et al., 2020).

- **Selection**: Healthy and unhealthy cough samples were selected at random.

- **Processing**: All selected samples were passed through the same preprocessing pipeline as the COUGHVID dataset to ensure consistency in format and quality.
- **Result**: This expansion introduced approximately 400 high-quality audio samples to the training set, enhancing both class balance and model generalizability.

4.4 Data Augmentation Using LDM

To further augment the training dataset, synthetic audio was generated using a Latent Diffusion Model (LDM), specifically AudioLDM2 (Liu, n.d). The workflow included the following steps (Fig. 4.1):

- **Prompt Generation**: A total of 80 text prompts for Audio LDM were crafted with the assistance of ChatGPT. After manual selection, 20 prompts for healthy coughs and 20 prompts for unhealthy coughs were selected.
- **Initial Audio Generation**: These prompts were fed into the AudioLDM2 model to generate 5-second audio samples at 16kHz.
- **Manual Review**: All generated audios were manually reviewed. Prompts that resulted in noisy, robotic, or unrealistic outputs were discarded.
- **Refined Generation**: The remaining prompts that generated high-quality audio were used to regenerate synthetic audio using the same model, improving consistency and quality.
- **Final Output**: A total of 300 healthy and 300 unhealthy samples were generated using the finalized prompts. After a second round of manual review, 185 healthy and 166 unhealthy samples were retained as usable for training.

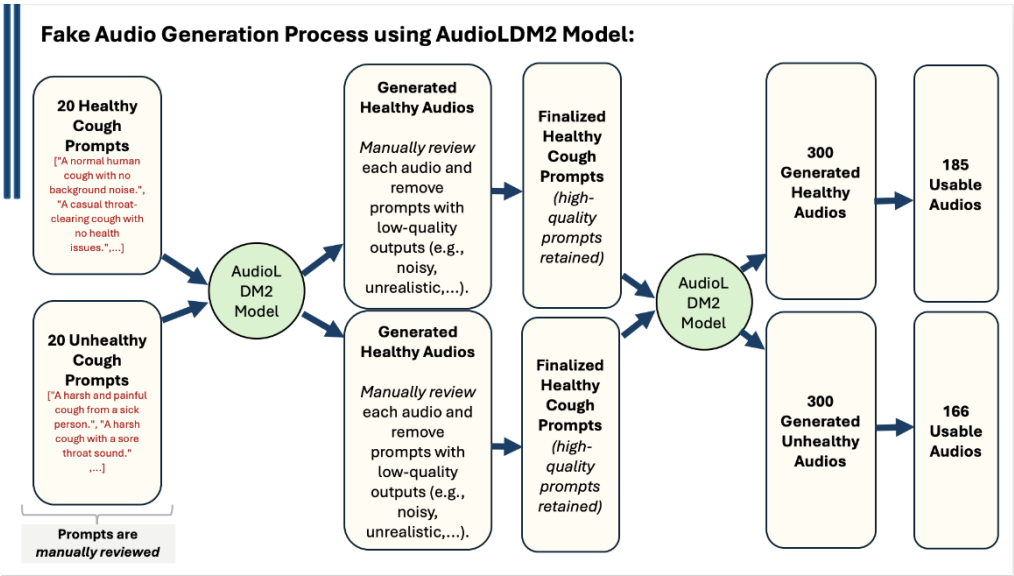

Fig. 4.1 Data generation steps

This expanded the dataset by approximately 15%, enhancing balance and model robustness.

4.5 Experiment Design

The experimental framework evolved through iterative refinements, transitioning from an initial conceptual design to a more pragmatic and empirically grounded final design. The initial experiments looked at creating a baseline classifier for the first phase by exploring CNN and CNN+Long Short-Term Memory (LSTM) architectures. The planned second phase was to explore data augmentation using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). However, during exploration of the said techniques in phase two,

which included generation of samples from VAE and an attempt to train our own GAN on samples from the COUGHVID dataset, our group found that our VAE-generated samples were distorted and of low-fidelity and hence not usable upon manual reviewing. We also underestimated the resources required to train a GAN based on samples from the COUGHVID dataset and thus failed the first time, later discovering that it would take over 100 hours to generate 100 audio samples using a T4 GPU or 14 hours with A100s.

Given the practical constraints and empirical observations, the eventual experiment design was streamlined to focus on two well-validated phases, with significant methodological refinements to improve effectiveness and efficiency.

- **Phase 1: Baseline Model** – The baseline model was developed using the pre-processed COUGHVID dataset. A CNN architecture was adopted. To enhance temporal feature capture, a LSTM layer was added. This hybrid CNN+LSTM model consistently outperformed CNN-only models in early testing and was therefore adopted for all subsequent evaluations.
- **Phase 2: Data Expansion and Augmentation** – To address data imbalance and improve model generalization, a targeted 15% dataset expansion was implemented using two sources:
    - **Coswara Integration**: Labelled cough samples were drawn from the Coswara dataset and processed using the same processing pipeline as the primary dataset, as described in section 4.3.
    - **Synthetic Generation via AudioLDM2**: In place of resource-intensive GANs, a LDM (AudioLDM2) was used to generate synthetic cough audio guided by carefully crafted textual prompts, as described in section 4.4. All generated audio was manually reviewed to filter out low-quality samples. High-quality prompts were retained for final data generation.

5. Project Design

The project was developed and evaluated using Python within a Jupyter Notebook / Google Colab environment. Due to local hardware constraints, cloud deployment and on-demand GPU resources were utilized. Key libraries included Librosa (for audio processing), NumPy, TensorFlow, and Keras.

Based on the datasets processed in Chapters 3 and 4, three training configurations were prepared:

1. A baseline model was explored using pre-processed data from COUGHVID. The primary goal of this step was to establish a local baseline for this project, which served as a point of comparison for subsequent evaluations.
2. COUGHVID dataset + Coswara samples (real data)
3. COUGHVID dataset expanded using **synthetic data,** generated by AudioLDM.

Due to resource limitations, the size of the dataset expansion was limited to 400 samples from the Coswara dataset and 351 sample generated by the LDM. Besides python, this project also utilised other related technology, such as:

- **Signal processing** was required for various audio processing tasks such as Short-time Fourier Transform, and discrete cosine transform to capture the features in audio. These technologies were found within the MFCC and log-Mel spectrogram, which were used for feature extraction.
- **Generative AI** was an important part of the project. Although still reliant of human verification, ChatGPT was used to generate prompts, which were then fed into the AudioLDM for cough audio generation.

- **Supervised Machine Learning**, however, remained the core of this project, where labelled datasets were used to train models for classification and prediction tasks. A comparison study of the models trained on data with expansion on real data and generated data was conducted.
- **Universal Workflow of Machine Learning** (Chollet, 2017) was adopted to push the models to its best performance based on the test dataset. This involved iteratively increasing and decreasing the complexity by adding more layers/neurons or regularise.

Model testing and evaluation were conducted using the same test dataset across all three training configurations. Given the class imbalance, the F1 score was used to provide a balanced measure of precision and recall, evaluated separately for each class. Overall accuracy was also reported to assess the model's general diagnostic reliability, given the prevalence of cough-related conditions. Fine tuning focused on improving the F1 score for the healthy class while minimizing any negative impact on the F1 score for unhealthy class and overall accuracy.

6. Implementation

Each dataset was organized into separate folders for training and testing, with the test set kept consistent across all models. Audio files were loaded by passing through a function, and the training set was further split into training and validation subsets. MFCC and Log-Mel spectrogram features were extracted, with feature frames padded or trimmed to the 95th percentile length to standardize input shapes.

6.1 Baseline Models Using COUGHVID

A simple CNN model was first developed using MFCC features from the COUGHVID dataset. Initial results showed 78% accuracy, with an F1 score of 10% for the healthy class and 87% for the unhealthy class—already outperforming the best performing model of 66.7% accuracy on 2,635 respiratory sounds (Sharma et al., 2020).

To address class imbalance, weight adjustments were applied. Further tuning included filter size optimization, learning rate scheduling, and early stopping to prevent overfitting. The best performing CNN only model achieved 71% accuracy, 41% F1 (healthy), and 81% F1 (unhealthy) which was a significant improvement for F1 on healthy class with minimal impact on the unhealthy class and overall accuracy.

LSTM layers were then added to extract time series features (Li & Wu, 2015). Both Bi-Directional LSTM and LSTM were explored, and regularisation was adjusted which resulted in improved performance for the best performing model: 74% accuracy, 48% F1 (healthy), and 83% F1 (unhealthy).

The same process was repeated for data with Log-Mel spectrogram features extracted. The best performing CNN only model had an accuracy of 78%, while the best performing CNN + LSTM model obtained 72% accuracy, 48% F1 (healthy) and 81% F1 (unhealthy).

Given the comparable performance of these two models, a Dual-Channel Audio Fusion Network that takes input from both MFCC and log-Mel spectrogram, fully connected dense layers was used to act as a feature fusion and transformation mechanism. An attention layer was also added at the start of the merged branch to selectively focus on the branch. This model achieved the best performance: 74.3% accuracy, 49.4% F1 (healthy) and 83.0% F1 (unhealthy).

6.2 Model Architecture

Eventually, we finalised our baseline model architecture as per the dual-channel CNN-LSTM network shown in Figure 6.1. MFCC and Log-Mel spectrogram each utilized one best-performed channel. Audio features were independently processed through a series of CNN+LSTM layers. Outputs from both channels were concatenated in a fusion layer, where an attention mechanism highlighted the most informative features. This design enabled

the model to effectively learn from both representations, enhancing cough classification and disease detection performance.

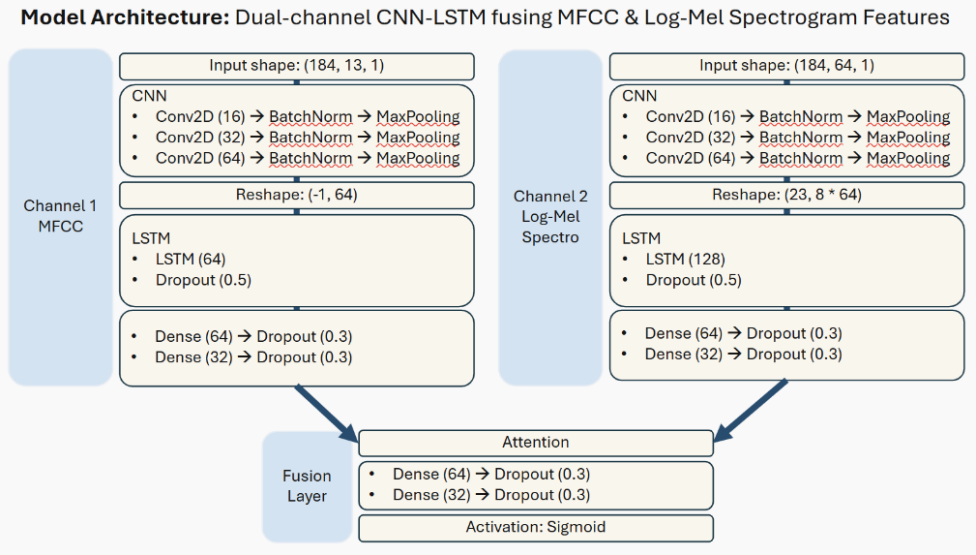

Figure. 6.1 Finalised baseline model architecture

6.3 COUGHVID + Coswara

To expand the original dataset and improve model generalization, COUGHVID dataset was combined with real cough recordings from the Coswara dataset. This expansion aimed to enhance diversity in cough sound samples and improve model performance. The following Iterative Development Process is as per our baseline model structure. For MFCC, after the model development process, we find a CNN + LSTM architecture achieving the best F1 score of 0.51 (healthy), while maintaining high F1 score on unhealthy cases. Removing regularization worsened results, affirming its importance.

Using Log-Mel spectrograms as features, CNN + LSTM model reached 0.48 F1 (healthy) and 0.72 accuracy. The final improvement comes from replacing LSTM with Bi-Directional LSTM, allowing the model to process cough sequences in both forward and backward directions. This achieved the best performance, with an increased in F1 score from 0.48 to 0.5 and test accuracy from 0.72 to 0.75. When Coswara data was used to expand the dataset, both MFCC and Log-Mel gave strong results. We combined the best MFCC and Log-Mel models into a dual-channel fusion network with an attention layer. This allowed the model to focus on the most informative features. The final model has a 72% accuracy and an 52.3% F1 (healthy), outperforming the baseline (F1 healthy: 49.4%).

Unlike the baseline, this model benefited from real-world data expansion using Coswara, extracting and learning more features during the iterative process.

6.4 COUGHVID + Synthetic Data

In the next experiment, we trained on an augmented dataset generated by LDM, instead of only using real-world data. The development process followed the same iterative approach as before, with regularization and weight adjustments applied to address class imbalance and reduce overfitting. When analysing MFCC features, the best model— CNN The same procedure was applied using Log-Mel features, with the best result again coming from a CNN+LSTM architecture.

The dual-channel fusion model achieved the best overall performance, with 76.3% accuracy, F1 healthy of 53.8%, and F1 unhealthy of 84.0%. These results highlighted that synthetic data could be highly effective, even more so than the real data from the Coswara samples, given that the real data also contained noise and displayed

inconsistencies. The cleaner and more consistent synthetic audio may have made it easier for the model to learn and generalize key patterns.

7. Results

Model evaluation followed the methodology outlined in the project design. This project had explored two basic model structures, namely LSTM and CNN. The models were trained using MFCC and Log-Mel spectrogram. A dual channel CNN+LSTM model with attention layer was also introduced in later stages. As described in section 6, the attention model performed the best, followed by the LSTM model and lastly the CNN model. This was largely expected given the attention model was able to learn more about the different features and thus generalise better in the test dataset. As cough audio data has a time domain, it was also expected that LSTM contributed to uplifting the performance of the overall model.

While some state-of-the-art models report over 95% accuracy, they relied on datasets 10–20 times larger than ours. In contrast, the best model trained on a similar sized dataset achieved only **66.7%** accuracy. All our models (Table 7.1) surpassed this baseline.

We compared the overall accuracy and F1 for each class and observed that the F1 for healthy class increased regardless of whether it was real or generated data that was added. F1 for unhealthy class and overall accuracy decreased when additional real data was added while it improved when synthetic cough sound were added. As Coswara performed worse than the models trained on COUGHVID and were biased towards F1, a finding based on our observations during our initial data exploration and model building phase, it was expected that the F1 for unhealthy class and overall accuracy decreased when it was added to COUGHVID data.

	F1 healthy	F1 unhealthy	Accuracy
Dataset 1 COUGHVID	49.4%	83.0%	74.3%
Dataset 2 COUGHVID + Coswara (real)	52.3%	80.1%	72.0%
Dataset 3 COUGHVID + LDM (generated)	53.8%	84.0%	76.3%

Table 7.1 Result comparison across different data source

8. Conclusion

8.1 Observation & Key Learnings

This project demonstrated that machine learning models can effectively classify respiratory health conditions using cough audio. Despite the challenges of limited and imbalanced data, meaningful classification performance was achieved through rigorous preprocessing, careful feature selection, and architectural refinements.

Our key learnings are as such:

- **Cough Audio is a Viable Diagnostic Feature**: With proper feature extraction (MFCC, Log-Mel), cough audio signals contained sufficient discriminative patterns for classification. The model achieved competitive performance even with a relatively small dataset.
- **Synthetic Data is a Valuable Resource**: The inclusion of synthetic cough samples generated via AudioLDM significantly improved model performance, particularly for the underrepresented healthy class. This suggests that well curated synthetic data could enhance model robustness and help address data scarcity especially relevant in healthcare contexts where labelled data is difficult to obtain.
- **Real vs. Synthetic Data**: Although real-world samples from Coswara added diversity, inconsistencies and noise slightly reduced model accuracy. In contrast, synthetic data from LDM, being cleaner and more uniformed, improved both class-level F1 scores and overall performance.

- **Model Architecture Matters**: CNN+LSTM models consistently outperformed CNN-only baselines. The introduction of a dual-channel attention fusion network leveraging both MFCC and Log-Mel features led to the highest performance, highlighting the value of combining temporal and spectral learning paths.

These insights reinforce the potential of using cough-based audio classification for scalable, non-invasive respiratory health screening, especially when complemented by synthetic data augmentation and thoughtfully designed neural architectures.

8.2 Gaps & Future Experiments

A key challenge encountered in this project was the generation of synthetic audio data. While Generative Adversarial Networks (GANs) were initially considered and even attempted, their computational demands made them impractical within the available resources. Latent Diffusion Models (LDMs) were adopted as a more efficient alternative. However, generative approaches remain resource-intensive, and future work could explore lighter-weight models to improve scalability.

Another limitation was the choice to focus exclusively on cough sounds. While the decision was deliberate to help us focus on our type of biomarker to streamline our experiments, we acknowledge the potential of other auditory biomarkers in enhance disease detection. Expanding the dataset to include other respiratory sounds such as breathing, speech, or wheezing may improve model performance and enable multi-condition classification.

Lastly, future research could potentially examine the impact of training with higher-quality cough data, as cleaner and more consistent samples may further improve model generalizability. A promising direction could involve training LDM or GAN models directly on the existing training dataset to generate domain-specific synthetic coughs, rather than relying on general audio-based generation techniques, though that would require greater access to computational resources.

References

Budd, J., Baker, K., Karoune, E., Coppock, H., Patel, S., Tendero Cañadas, A., Titcomb, A., Payne, R., Hurley, D., Egglestone, S., Butler, L., Mellor, J., Nicholson, G., Kiskin, I., Koutra, V., Jersakova, R., McKendry, R. A., Diggle, P., Richardson, S., Schuller, B. W., Gilmour, S., Pigoli, D., Roberts, S., Packham, J., Thornley, T., & Holmes, C. (2022). A large-scale and PCR-referenced vocal audio dataset for COVID-19. arXiv. https://doi.org/10.48550/arXiv.2212.07738

Chaudhari, G., Jiang, X., Fakhry, A., Han, A., Xiao, J., Shen, S., & Khanzada, A. (2020). Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. arXiv. https://doi.org/10.48550/arXiv.2011.13320

Chollet, F. (2017) Deep learning with python. New York , United States: Manning Publications.

Coppock, H., Jones, L., Karia, A., Schuller, B. W., & Gaskell, M. G. (2024). Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptom checkers. Nature Machine Intelligence, 6(2), 146–155. https://doi.org/10.1038/s42256-023-00773-8

Erdoğan, Y. E., & Narin, A. (2021).COVCoppock, H., Jones, L., Karia, A., Schuller, B. W., & Gaskell, M. G. (2024). Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptom checkers. Nature Machine Intelligence, 6(2), 146–155. https://doi.org/10.1038/s42256-023-00773-8ID-19 detection with traditional and deep features on cough acoustic signals.Computers in Biology and Medicine, 136, 104765. https://doi.org/10.1016/j.compbiomed.2021.104765

Husssain, S., Ayoub, M., Wahid, J. A., Khan, A., Alabrah, A., & Amran, G. A. (2024). Cough2COVID-19 detection using an enhanced multi-layer ensemble deep learning framework and CoughFeatureRanker. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-76639-9

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open Journal of Engineering in Medicine and Biology, 1, 275–281. https://doi.org/10.1109/OJEMB.2020.3026928

Li, X. and Wu, X. (2015) 'Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition', Interspeech 2015, pp. 3219–3223. doi:10.21437/interspeech.2015-648.

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, 156. https://doi.org/10.1038/s41597-021-00937-4

Pahar, M., Klopper, M., Warren, R., & Niesler, T. (2021). COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. arXiv preprint arXiv:2104.02477. https://doi.org/10.48550/arXiv.2104.02477

Ramesh, V., Vatanparvar, K., Nemati, E., Nathan, V., Rahman, M. M., & Kuang, J. (2020).CoughGAN: Generating synthetic coughs that improve respiratory disease classification.In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 5682–5688). IEEE. https://doi.org/10.1109/EMBC44109.2020.9175597

Reddy, R. P., Srikeerthi, S., Hema, C., & Latha, B. S. (2025)Enhancing Audio Synthesis with WAVEGAN: A Generative Adversarial Network (GAN) Approach.In A. Kumar et al. (Eds.), Proceedings of the Third International Conference on Cognitive and Intelligent Computing, Volume 2, pp. 107–113. Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9266-5_10

Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia,

T., Cicuta, P., & Leon, ... (2021). The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates. Proceedings of INTERSPEECH 2021.

Shati, A., Hassan, G. M., & Datta, A. (2023). COVID-19 detection system: A comparative analysis of system performance based on acoustic features of cough audio signals. arXiv preprint arXiv:2309.04505. https://arxiv.org/abs/2309.04505

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., & Ganapathy, S. (2020). Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv. https://doi.org/10.48550/arXiv.2005.10548

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., & Ganapathy, S. (2023). Coswara: A respiratory sounds and symptoms dataset for assessing COVID-19. Scientific Data, 10(1), Article 266. https://doi.org/10.1038/s41597-023-02266-0

Datasets

Coppock, H., Budd, J., Karoune, E., Holmes, C., Baker, K., Pigoli, D., Nicholson, G., Payne, R., Kiskin, I., Packham, J., Cañadas, A. T., Patel, S., Egglestone, S., Titcomb, A., Hurley, D., Butler, L., Thornley, T., Mellor, J., Roberts, S., Gilmour, S., Schuller, B., Koutra, V., Jersakova, R., Diggle, P., & Richardson, S. (2024). The UK COVID-19 Vocal Audio Dataset (Version openAccessv1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11167750

IISC LEAP. (n.d.). Coswara-Data [GitHub repository]. GitHub. Retrieved March 5, 2025, from https://github.com/iiscleap/Coswara-Data

Orlandic, L., Teijeiro, T., & Atienza, D. (2021). The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms (Version 3.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7024894

Virufy. (n.d.). Virufy-data [GitHub repository]. GitHub. Retrieved March 5, 2025, from https://github.com/virufy/virufy-data

Liu, H. (n.d.). *AudioLDM2: Text-to-Audio/Music Generation* [Computer software]. GitHub. https://github.com/haoheliu/AudioLDM2

Annex A – Comparison & Evaluation of Datasets Explored

Dataset	Source	Key Features	Y/N?	Reason	Existing Benchmarks / Remarks
COUGHVID (Orlandic et. al., 2021)	Embedded Systems Laboratory (ESL), Switzerland	• Crowdsourced 25,000 cough recordings • **Only 2,800 recordings labelled by experts**	**Y** **(Pri)**	• **Available** • **Expert / validated labels** • **Decent audio quality** • **Reasonable sized dataset (after preprocessing), ~2k**	• Accuracy = 96.8%, F1 = 91.0% using a InceptionFireNet (i.e., 3 CNNsblocks for feature extraction) and DeepConvNet (i.e., 4 CNN blocks for classifying) (Celik, 2023) • Note: dataset used was significantly large due to **inclusion of non-validated labels**
Coswara	Indian Institute of Science (IISc), Bangalore	• Crowdsourced **2,635** respiratory-related sounds • **Labels based on PCR tests**	**Y** **(Sec)**		• **Accuracy = 66.7% (2020)** using Random Forest based on MFCC, frequency-domain features (e.g., spectral bandwidth) & time-domain features (e.g., signal energy) (Bhattacharya et al., 2020)
UK COVID-19 Vocal Audio	UK Health Security Agency (UKHSA)	• Collected audio recordings from **72,999 participants** • Labels based on PCR tests	N	• Unable to process large dataset (130GB) • Poor performance of small sample (20%); likely due to quality	• **Unweighted Average Recall (UAR) = 68.1%, AUC = 75.0%** using self-supervised audio spectrogram transformer (Coppock et al., 2024)
Virufy	Virufy AI Research Group	• Combination of ~3.5k crowdsourced & clinical cough recordings	N	• Not readily available	• **AUC = 77.1%** using ensemble of three networks (MFCC + 2 layers, Mel-Spec + 3 CNN layers, other features + 2 layers) (Laguarta et al., 2020)
ComParE	Interspeech Computational Paralinguistics Challenge 2021	• Used in research competition • 929 cough recordings from 397 participants	N	• Not readily available	• **UAR = 73.9%** using a fusion model based on majority voting system (acoustic features SVM, BoAW SVM, MFCC + pre-trained CNN + SVM, S2SAE + SVM, CNN + LSTM RNN) (Schuller et al., 2021)