Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Development of a Digital Biomarker for Symptomatic COVID- 19 via Frequency-Aware Spectral Features and Machine Learning Tuning

Azree Nazri[1,2,*], Faisal Aziz[2,3,♥], Olalekan Agbolade[2]

[1] Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2] Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400, Serdang, Selangor, Malaysia
[3] Department of Computer Science and Engineering, Faculty of Engineering, Comilla University, Cumilla-3506, Bangladesh

ARTICLE INFO

Article history:
Received 12 June 2025
Received in revised form 8 September 2025
Accepted 5 December 2025
Available online 21 July 2026

Keywords:
Cough acoustics; frequency-band analysis; digital biomarker

ABSTRACT

Cough-based analysis has emerged as a non-invasive, low-cost solution for identifying symptomatic respiratory illnesses, including COVID-19. However, current models often suffer from low interpretability, frequency insensitivity, and poor performance under class imbalance. To address these gaps, this study aims to develop a frequency-aware classification framework tailored for symptomatic COVID-19 detection using cough sounds as a potential digital biomarker. The method involves decomposing audio into four spectral bands—[300–400 Hz], [300–600 Hz], [450–1000 Hz], and [1100–1650 Hz]—followed by spectral analysis, feature extraction, and classification via XGBoost. Advanced training strategies incorporating Bayesian optimization, SMOTE up-sampling, and threshold moving were applied. Results show the [450–1000 Hz] band yielded the highest performance, with precision of 0.96 and recall of 0.97, outperforming state-of-the-art baselines. This study concludes that integrating frequency segmentation with interpretable training enhancements can significantly improve symptomatic COVID-19 detection, establishing cough acoustics as a viable digital biomarker for scalable deployment in real-world, resource-constrained healthcare environments.

1. Introduction

Coughing is a multifaceted symptom associated with numerous medical conditions. This complexity creates considerable diagnostic challenges due to the symptom's diverse manifestations and varying severity levels [1-2]. Traditional diagnostic methods for respiratory illnesses—including chest radiography, swab testing, and bronchoscopy—are undeniably effective yet often limited by high costs and restricted accessibility [3-6]. Consequently, there is an urgent need for alternative diagnostic solutions that are both cost-effective and broadly accessible. Machine learning has emerged as a compelling solution, offering the potential to deliver accurate diagnostic tools that could be deployed on ubiquitous devices, such as smartphones [7-9]. For instance, AI algorithms can

* *Corresponding author.* *azree@upm.edu.my*
♥ *faisal@cou.ac.bd*

https://doi.org/10.37934/araset.61.5.386400

analyse acoustic features of cough sounds, potentially identifying disease-specific patterns that allow for more precise and convenient diagnosis. This approach could democratize access to respiratory diagnostics, especially in remote or resource-limited settings, where conventional techniques remain inaccessible.

Interestingly, research has shown that even experienced healthcare experts achieve just 34% accuracy when attempting to diagnose underlying diseases based purely on cough sounds [10]. This diagnostic challenge stems from the auditory similarity of coughs across different respiratory diseases, which obscures distinguishing aspects. Machine learning algorithms, on the other hand, may detect minute audio patterns that the human ear frequently misses [11]. However, despite their promise, the use of these models in clinical practice is limited due to issues with interpretability and transparency—both of which are critical for medical decision-making [12].

The major contributions of this study are threefold. First, it introduces a novel frequency-region analysis framework that systematically evaluates low- and mid-frequency spectral bands to identify interpretable acoustic biomarkers from cough audio signals. This approach contrasts with prior studies that often rely on full-spectrum features or opaque deep learning models. Second, the study leverages a multi-source, real-world dataset comprising symptomatic COVID-19-positive and non-COVID individuals, offering a clinically realistic benchmark for evaluating diagnostic performance—something absent in many earlier works focused solely on general COVID-19 or healthy cohorts. Third, the study implements and compares two interpretable machine learning models, k-nearest neighbors (kNN) and Extreme Gradient Boosting (XGBoost), across several targeted frequency bands, demonstrating that low-frequency ranges—particularly [300–400 Hz] and [450–1000 Hz]—yield the strongest classification performance in terms of both statistical separability and Matthews Correlation Coefficient (MCC). Together, these contributions establish a reproducible, explainable, and data-balanced diagnostic framework that not only advances the technical development of AI-based cough screening tools but also enhances their clinical relevance for scalable, non-invasive healthcare deployment in low-resource or remote environments.

2. Related Work

The COVID-19 pandemic has accelerated global interest in non-invasive diagnostic modalities, among which cough sound analysis has emerged as a promising low-cost, accessible solution. Numerous machine learning (ML) approaches have been proposed to classify COVID-19 and other respiratory conditions from audio signals, employing both conventional and deep learning models. Table 1 summarizes recent literature based on the classification target, model type, and use of techniques such as hyperparameter tuning and threshold adjustment.

2.1 Studies Focused On Covid-19 Detection (Symptomatic/Asymptomatic Cases)

Islam et al. [13] implemented an SVM classifier using chromagram features to detect COVID-19 from cough recordings. While their model achieved reasonable performance, it lacked robustness in real-world settings due to its dependence on handcrafted features and absence of symptom-specific modeling, reducing its clinical transferability. Manshouri [14] applied k-nearest neighbors (kNN) with spectral analysis for cough classification. Though computationally efficient, this approach did not incorporate feature selection or hyperparameter tuning, potentially leading to overfitting, especially in noisy datasets. The study also did not distinguish between symptomatic and asymptomatic cases, a crucial factor in practical diagnostics. Södergren et al. [15] and Brown et al. [16] employed SVMs and Random Forests respectively, using common features like MFCCs and zero-crossing rate. These studies demonstrated baseline feasibility, but they relied heavily on pooled datasets and did not perform frequency-specific analysis, limiting their biomarker discovery potential. Kho et al. [17] explored deep learning using Malaysian cough recordings, presenting a CNN-based framework for COVID-19 detection. While the approach improved classification accuracy, the model lacked explainability and did not use any threshold moving or balancing methods, making it difficult to calibrate the model in clinically imbalanced datasets. Xu et al. [18] proposed a VGG-based CNN model that automatically learned hierarchical features from raw cough waveforms. Despite its innovative use of end-to-end architecture, the model was trained on datasets that did not distinguish symptomatic status and lacked mechanisms to interpret what features influenced the model's predictions—making it difficult for clinical adoption where explainability is critical.

2.2 Studies Focused On Other Respiratory Conditions

Lee et al. [19] introduced a hybrid CNN-MobileNet model optimized for real-time deployment on mobile devices to classify various respiratory conditions. While valuable for mobile health, the study did not consider symptom annotations, and its design was not tailored for COVID-19 or similar acute infections requiring high sensitivity and specificity. Chen et al. [20] used feature selection from audio signals to build a traditional ML model for general cough detection. However, the absence of deep learning techniques and limited dataset diversity reduced the model's generalizability to more complex clinical scenarios such as COVID-19. Soliński et al. [21] relied on airflow signals for cough detection in portable spirometry devices. Although effective in controlled environments, this method ignored audio-based features entirely, limiting its adaptability for non-contact diagnostics. Chuma et al. [22] implemented a continuous-wave Doppler radar system combined with CNNs to detect coughs and other gestures. However, the approach was hardware-dependent and not designed for disease-specific classification and thus does not support scalable screening for COVID-19-like conditions. Chowdhury et al. [23] developed a mobile AI app for asymptomatic COVID-19 detection using both cough and breathing sounds. While notable for attempting asymptomatic classification, the model lacked any form of frequency segmentation or threshold calibration, potentially increasing false positive rates and limiting its reliability in real-world deployment.

2.3 Studies Including Symptomatic Covid-19 Cases

Pahar et al. [24] proposed a CNN trained exclusively on symptomatic COVID-19 cough samples. The model leveraged acoustic patterns such as vocal harshness for classification, improving specificity. However, it used full-spectrum features without identifying which frequency bands contributed most to its decisions, which weakens its interpretability. Lella and Pja [26] introduced a deep CNN model using multi-feature audio (cough, voice, breath) with explicit labelling of symptomatic and asymptomatic cases. While this model showed improved granularity, its deep architecture remained opaque, and no threshold tuning or model balancing was performed—reducing trust and usability in clinical settings. Nessiem et al. [25] presented a hybrid logistic regression–CNN model for classifying non-COVID respiratory illnesses. Although the architecture demonstrated general applicability, it did not target COVID-19 or stratify by symptom, making it less suitable for pandemic-era diagnostics. Moreover, hyperparameter tuning and data imbalance handling were not included. Chowdhury et al. [29] proposed an ensemble MCDM (multi-criteria decision-making) method that achieved strong AUC and precision values on COVID-19 cough data. However, the model functioned as a black box with no frequency-region insight, and it did not isolate symptomatic COVID-19 as a category, potentially conflating critical clinical cues. Akman et al. [30]

developed the COVID-19 Identification ResNet (CiDER) and tested it on Interspeech challenge datasets. While the model attained competitive performance, it treated symptomatic and asymptomatic cases uniformly and did not apply threshold adjustment or frequency-wise feature analysis, reducing its utility in high-risk clinical triage situations.

3. Methodology

In this study, we implemented a robust cross-dataset learning framework to enhance the accuracy and generalizability of COVID-19 detection through cough sound classification. Our methodology harnesses the strength of multi-source data fusion and strategic pre-processing, providing a unified foundation for training models that are resilient across diverse recording environments and population groups.

We curated cough audio recordings from five reputable datasets—Cambridge, COVIDID, Coswara, NoCoCoDa, and Virufy—each offering varied acoustic characteristics and recording conditions. These datasets were subjected to a rigorous data pre-processing pipeline, involving normalization, segmentation, and noise filtering, to ensure consistency and quality in input features. Following this, we initiated Task 1: a comprehensive cross-dataset training and testing procedure that utilized all five sources. This task was designed to simulate real-world deployment scenarios by exposing models to diverse acoustic profiles during training and evaluating them on unseen datasets. This not only strengthened the generalizability of the models but also mitigated overfitting to a single-source dataset.

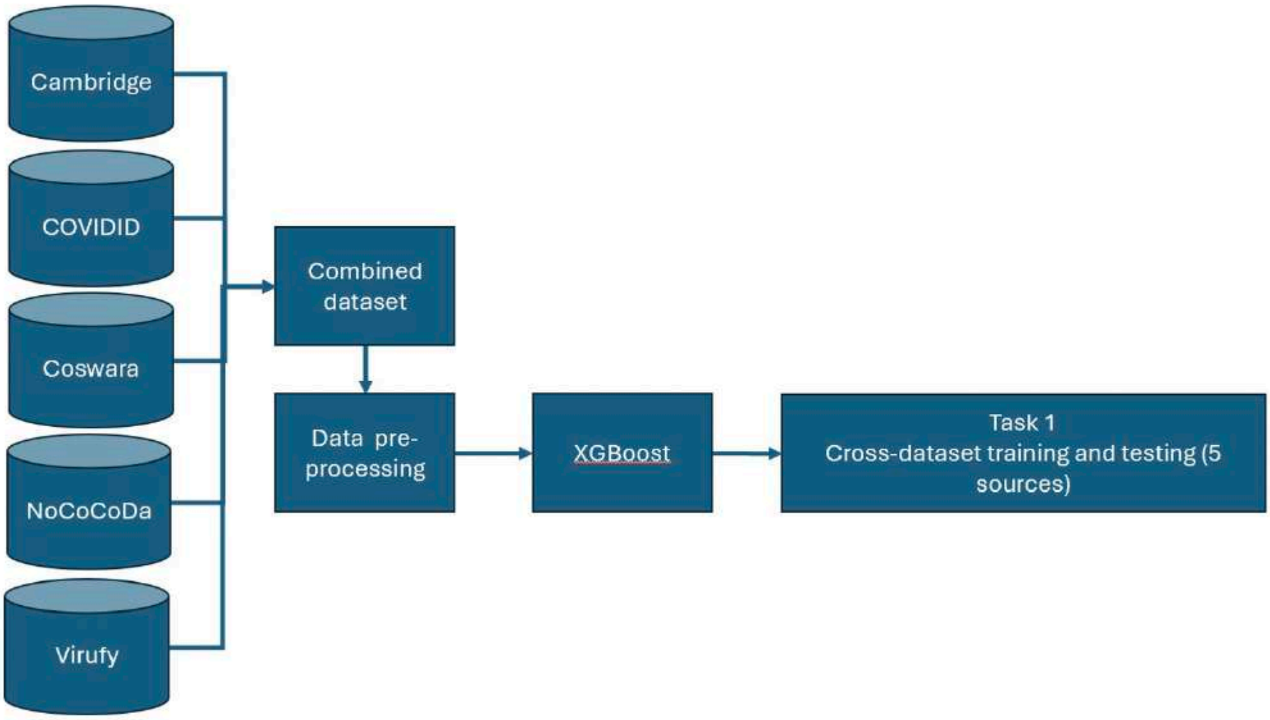


Fig. 1. Overall Methods for Cough Sound Analysis and Model Development Across Multiple Datasets

To perform the classification, we employed Extreme Gradient Boosting (XGBoost), a highly efficient and scalable ensemble learning algorithm known for its performance in structured data tasks. XGBoost was selected due to its robustness, interpretability, and compatibility with both imbalanced data handling and optimized hyperparameter tuning techniques. Default hyperparameters were initially used as a baseline, followed by the application of advanced strategies—such as Bayesian optimization and threshold moving—to enhance classification performance under varying data conditions.

Through this integrative process, we aimed to construct a highly adaptable and reliable classification pipeline. Our methodology reflects a practical and scalable approach for developing cough-based screening tools suitable for varied and decentralized healthcare environments.

3.1 Experiment Data

This study utilized a combined dataset aggregated from five publicly available and diverse sources—Cambridge, Coswara, COUGHVID, Virufy, and NoCoCoDa—to enhance the generalizability and robustness of the proposed cough-based COVID-19 classification framework. Each dataset contributed unique attributes in terms of recording conditions, symptom annotations, and demographic variability, enabling comprehensive evaluation across heterogeneous real-world data. The COUGHVID dataset, in particular, offered the highest volume of samples, including both asymptomatic and symptomatic COVID-19 positive cases, as well as a substantial set of COVID-19 negative samples. This dataset enabled detailed exploration of cough signal characteristics across a wide range of symptom profiles. The Cambridge and Coswara datasets contributed moderate numbers of labeled samples, primarily categorized as COVID-19 positive or negative, though symptom distinctions were limited. Meanwhile, Virufy and NoCoCoDa contributed fewer samples but were crucial in supporting model validation on smaller-scale, independent corpora.

To ensure maximum diversity and support real-world deployment simulations, we consolidated all five datasets into a unified combined dataset. This final dataset comprised a total of 3,398 cough samples, including 1,181 samples from COVID-19-positive individuals and 2,217 samples from COVID-19-negative individuals. As visualized in Fig. 2, the combined dataset exhibits a notable class imbalance, with COVID-negative samples significantly outnumbering positive cases. Additionally, symptom-type distinctions (asymptomatic vs. symptomatic) were only available within subsets of COUGHVID and Cambridge.

All datasets were processed using a consistent pre-processing pipeline that included silence trimming, normalization, and segmentation. This procedure was essential to reduce domain-specific noise, harmonize input quality, and ensure compatibility across datasets.

The resulting cross-dataset corpus served as the primary experimental input for all training and evaluation procedures. It provided a diverse, real-world testbed for assessing the performance, robustness, and adaptability of our proposed frequency-band-based classification framework, especially under clinically realistic, imbalanced conditions.

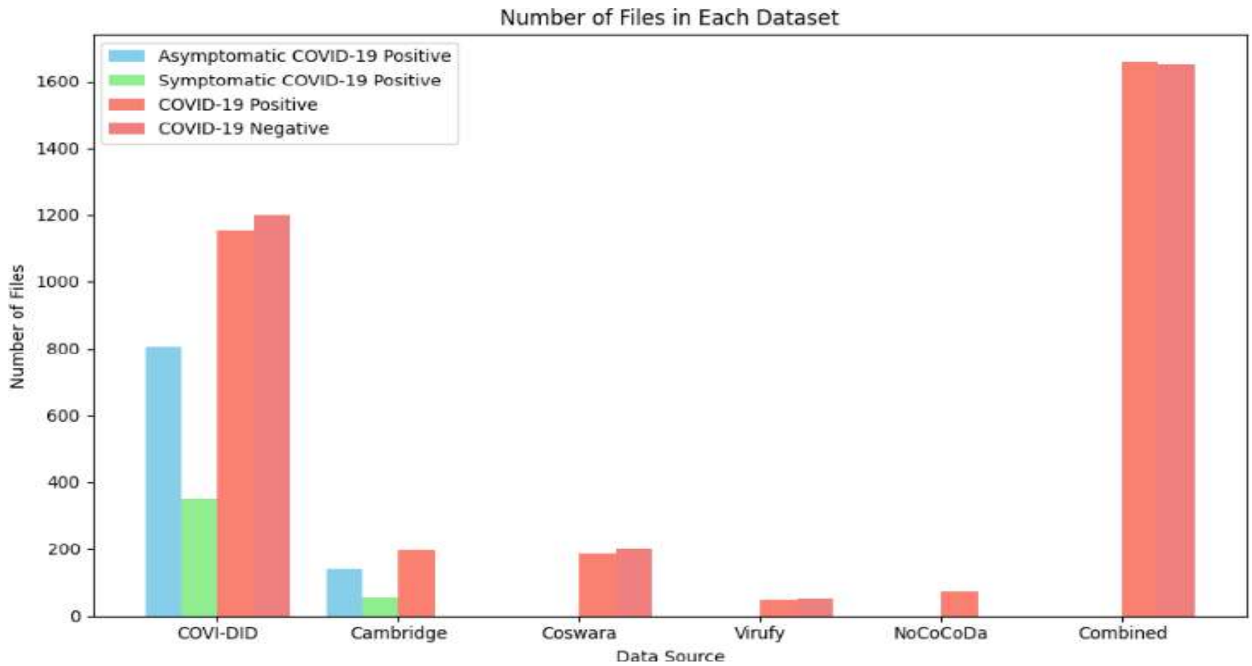


Fig. 2. Bar plot showing the number of files for Asymptomatic COVID-19 Positive, Symptomatic COVID-19 Positive, COVID-19 Positive (both subtypes), and COVID-19 Negative cases across datasets including Coughvid, Cambridge, Coswara, Virufy, NoCoCoDa, and the combined dataset

3.2 Feature Extraction

Cough audio signals were sampled at 22.5 kHz and segmented into 10 ms frames for detailed analysis. Using the Librosa library, five key spectral features—MFCCs, Mel-scaled spectrogram, chromogram, tonal centroid, and spectral contrast—were extracted. These features formed a comprehensive vector to enable accurate classification of COVID-19-related cough patterns.

3.2.1 Spectral Analysis

Spectral analysis uses power spectral density (PSD) to show the energy concentration at different frequencies, which is commonly shown as a frequency-domain graph. The average power and PSD of an audio signal x(t), recorded as a time series and analyzed using its Fourier transform f(q), are stated as follows:

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |f_T(q)|^2 \, dq$$

and

$$S = \lim_{T\to\infty} \frac{1}{T} |f_T(q)|^2$$

The Python SciPy package provides fast methods for computing PSD, including Welch and periodograms, which often involve parameters like the Hann window. In this investigation, audio samples collected at a sampling rate of 22.5 kHz, with each segment lasting around 10 ms, were examined. PSD values were used to distinguish symptomatic COVID-19 coughs from positive COVID-19 and healthy coughs. A PSD plot of cough sound data was created, allowing for differentiation between symptomatic COVID-19 and positive COVID-19 samples and showing distinguishing patterns associated with healthy samples. Additionally, PSD plots were used to compare symptomatic Covid-19/positive Covid-19 patients and symptomatic Covid-19/healthy patients, focusing on spectral markers that may be unique to each illness.

3.2.2 Acoustic Features

To align with established standards in audio applications. Five distinct spectral feature categories are subsequently derived from the sampled audio, namely Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Scaled Spectrogram, Tonal Centroid, Chromagram, and Spectral Contrast. An overview of the feature extraction methodology is depicted in Figure 3. These features are extracted utilizing the Python library librosa, which facilitates robust audio signal processing. For instance, MFCCs capture the timbral aspects of the cough signal, while the Mel-Scaled Spectrogram provides a visual representation of the frequency content over time. The extraction process, actively implemented through librosa, supports the consistent generation of features critical for downstream machine learning tasks in COVID-19 detection.

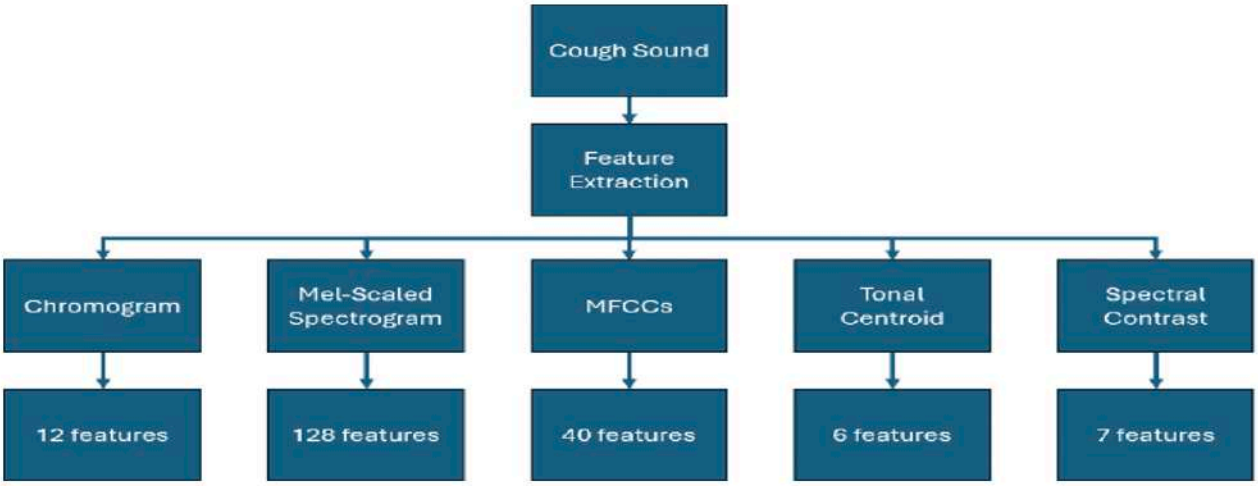


Fig. 3. Overview of Spectral Feature Extraction Process for Cough Sound Analysis

To effectively capture the acoustic characteristics of cough sounds, a comprehensive feature extraction pipeline was implemented. As illustrated in Fig. X, the process begins with raw cough audio recordings, which are subjected to a multi-dimensional feature extraction stage to ensure rich representation of both time and frequency domain properties. Five distinct types of acoustic features were extracted:

Chromogram: Captures the energy distribution across 12 pitch classes, resulting in 12 chroma-based features. These are particularly useful for identifying tonal patterns and harmonic structures in the cough signal.

Mel-Scaled Spectrogram: Provides a perceptually motivated spectral representation using the Mel scale. This high-resolution feature set contributes 128 coefficients per frame, enabling detailed analysis of energy distribution across frequency bands relevant to human hearing.

Mel-Frequency Cepstral Coefficients (MFCCs): Widely regarded in speech processing, MFCCs capture the short-term power spectrum of sound. A total of 40 MFCC features were computed to represent spectral shape and envelope, which are essential for differentiating cough profiles across infection types.

Tonal Centroid: Derived from tonal tension and harmonic content, this set contributes 6 features that capture pitch-related variations and resonance characteristics in the cough signal.

Spectral Contrast: Measures the difference between peaks and valleys in each frequency sub-band, generating 7 features that highlight the spectral dynamics of the cough, particularly in the presence of voice irregularities or background noise.

These features were concatenated into a unified feature vector to represent each cough instance. This rich representation enabled the machine learning models to better capture the discriminative properties of COVID-19 and non-COVID cough sounds across various symptom categories. The use of multiple feature types ensured robustness against variations in recording conditions and speaker profiles, thereby enhancing the generalizability of the classification framework across diverse datasets.

3.2.3 XGBoost Classifier

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable machine learning framework that applies gradient boosting techniques for both classification and regression tasks. This approach combines multiple weak learners—typically decision trees—to build a robust predictive model, progressively reducing errors from earlier trees. XGBoost stands out for its rapid processing capabilities and performance, attributed to its enhanced handling of missing data, regularization options to mitigate overfitting, and support for parallel computation. The algorithm constructs a sequence of decision trees, where each successive tree aims to correct the errors of its predecessors, resulting in a model adept at capturing complex, nonlinear relationships. In our study, XGBoost was employed to assess resonance power (RP) values in cough sounds from COVID-19 positive and healthy control groups across diverse frequency ranges, with the objective of identifying and categorizing distinct patterns efficiently.

The hyperparameters are detailed in Table 1. These include default settings commonly applied in XGBoost models, such as n_estimators = 100, max_depth = 6, and learning_rate = 0.3. Other parameters, including subsample, colsample_bytree, gamma, lambda, alpha, and scale_pos_weight, were maintained at their respective defaults to establish a consistent performance baseline.

Table 1
The default hyper-parameters are used for all datasets

Hyperparameter	Default Value
n_estimators	100
Max_depth	6
Learning rate	0.3
Min_child_weight	1
Subsample	1
Colsample bytree	1
Gamma	0
Lambda	1
Alpha	0
Scale_pos_weight	1

3.3 Training Strategies

To evaluate the impact of different training mechanisms on classification performance, two distinct strategies were implemented and compared, as summarized in Table 2. Strategy A adopts a baseline approach utilizing default hyperparameters, without applying any up-sampling or threshold adjustment techniques. In contrast, Strategy B incorporates a more refined pipeline through Bayesian optimization for hyperparameter tuning, combined with Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, and Threshold Moving to calibrate decision boundaries for improved sensitivity.

Table 2
Combinations of two training strategies

Strategy	Hyper-Parameter Selection Method	Up-sampling SMOTE	Threshold Moving
A	Default	Yes	No
B	Bayesian Optimization	Yes	Yes

Strategy B introduces Bayesian optimization to systematically search for optimal hyperparameter configurations, thereby improving generalization and performance. Additionally, SMOTE was applied to synthetically balance the class distribution within the training set, mitigating the bias caused by underrepresented classes. Threshold moving was subsequently applied post-training to fine-tune decision cut-offs based on validation results, enhancing classification sensitivity under imbalanced conditions. Collectively, these two strategies allow for a comparative assessment of the effectiveness of advanced training enhancements over traditional default-based learning in cough-based COVID-19 detection models.

4. Results
4.1 Separability Test Result

Figure 4 illustrates the frequency-domain characteristics of cough signals using Power Spectral Density (PSD) analysis across three clinically relevant groupings: Figure 4(a) shows COVID-19-positive (both symptomatic and asymptomatic) vs healthy individuals, Figure 4(b) healthy vs symptomatic COVID-19-positive individuals, and Figure 4(c) asymptomatic vs symptomatic COVID-19-positive individuals. These comparisons provide insight into the spectral separability and potential acoustic biomarkers of COVID-19 infection and symptom manifestation. Figure 4(a) demonstrates a clear spectral distinction between COVID-19-positive and healthy participants. The COVID-19 group shows elevated energy in the low-frequency range (approximately 300–800 Hz), suggesting that infection-induced changes in the respiratory system—such as airway inflammation, increased secretion, or altered cough biomechanics—manifest acoustically in this band. This supports the hypothesis that low-frequency energy can serve as a non-invasive biomarker for COVID-19 detection. Figure 4(b) compares healthy individuals with those who are symptomatic but COVID-19-positive. Although subtle elevation is observed in the low-frequency spectrum for symptomatic cases, the overlap across most of the spectrum is substantial. This limited separation implies that the presence of symptoms alone does not introduce strong enough spectral changes to distinguish these groups using PSD features alone. Therefore, while symptoms affect the cough mechanism, they may not significantly impact its spectral profile in a linear fashion. Figure 4(c) evaluates the PSD curves between asymptomatic and symptomatic COVID-19-positive individuals. The plots are nearly indistinguishable, indicating that symptom severity does not significantly alter the spectral energy distribution of coughs within the COVID-19-positive group. This suggests that the acoustic features captured by PSD alone may not be sufficient to differentiate symptom onset or intensity, highlighting the need for more nuanced feature extraction techniques (e.g., time-frequency analysis, deep learning representations).

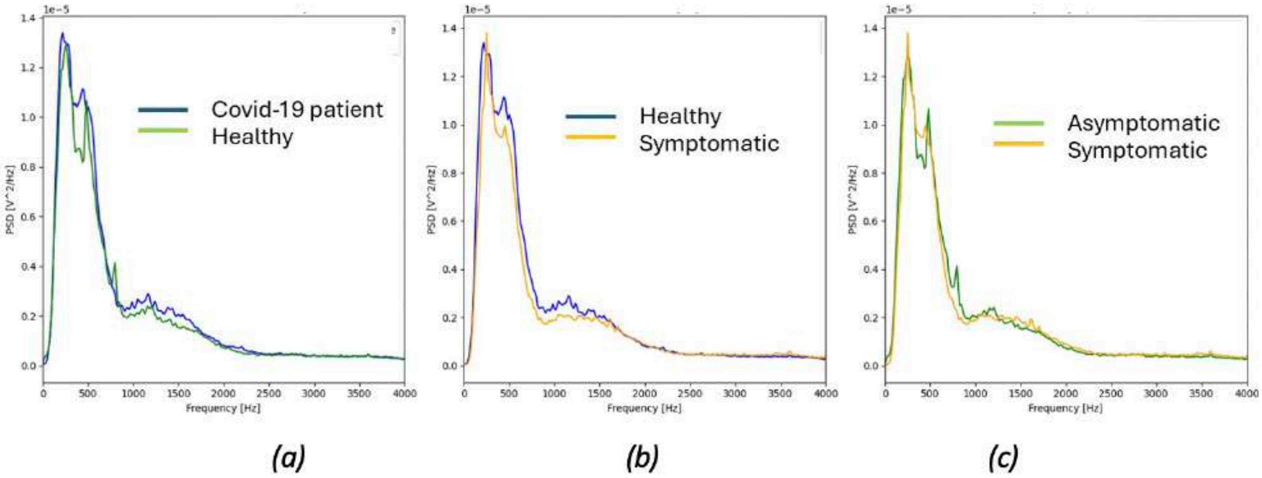


Fig. 4. (a) COVID-19-positive (symptomatic and asymptomatic cases) individuals vs healthy controls (b) healthy vs symptomatic COVID-positive individuals c) Asymptomatic vs symptomatic COVID-19-positive individuals

Figure 5 visualizes the 3D frequency distribution of power spectral features across three critical pairwise group comparisons: (a) COVID-19-positive vs healthy, (b) symptomatic COVID-19-positive vs healthy, and (c) symptomatic vs asymptomatic COVID-19-positive individuals. These visualizations are crucial for understanding the spectral structure of cough sounds and identifying frequency regions where significant deviations occur due to infection or symptom manifestation.

In Figure 5(a), a distinct spectral energy peak is observed in the low-frequency region (300–800 Hz) for COVID-19-positive individuals, visually separated from the healthy group. The elevated energy distribution likely stems from physiological disruptions associated with COVID-19, such as bronchial inflammation, mucus obstruction, and altered airflow dynamics. These disruptions can introduce turbulent airflow or reduced vocal fold modulation, producing dominant low-frequency components in cough acoustics. The sharp contrast in spectral energy amplitude between the groups confirms that low-frequency regions contain strong diagnostic cues. This also supports the findings from the PSD (Figure 5a) and statistical tests (Table II), establishing [300–400] Hz as a discriminative band. The visualization provides a compelling case for using spectral features in ML classifiers for COVID-19 detection.

Journal of Advanced Research in Applied Sciences and Engineering Technology
Volume 61, Issue 5 (2026) 386-400

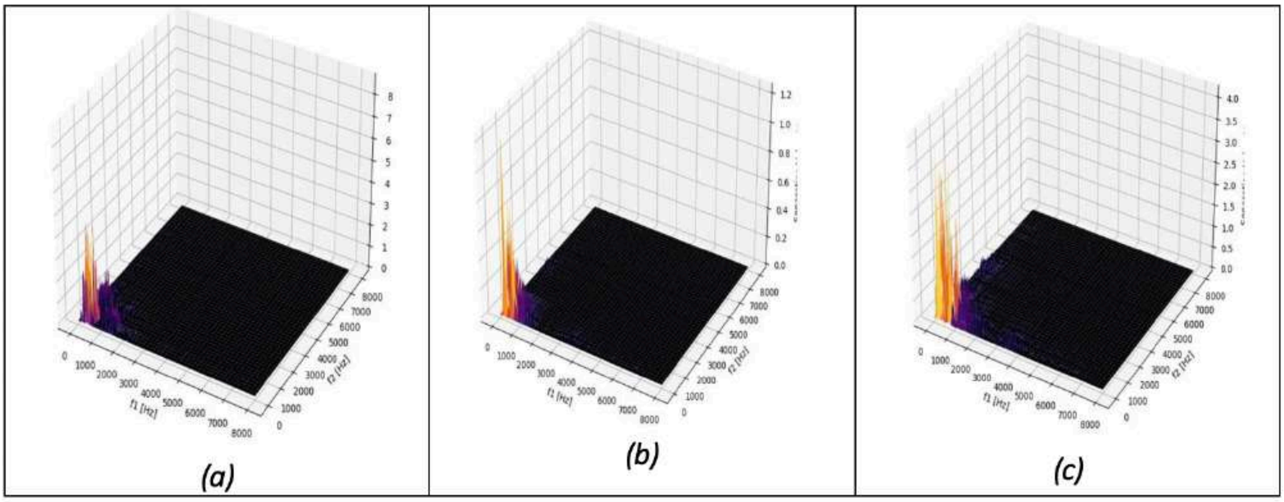


Fig. 5. (a) 3D Frequency Distribution of Audio Features: COVID-19 Positive (both symptomatic/asymptomatic) vs Healthy cases. (b) 3D Frequency Distribution of Audio Features: Healthy vs Symptomatic COVID-19 positive (c) 3D Frequency Distribution of Audio Features: asymptomatic COVID-19 positive vs Symptomatic COVID-19 positive

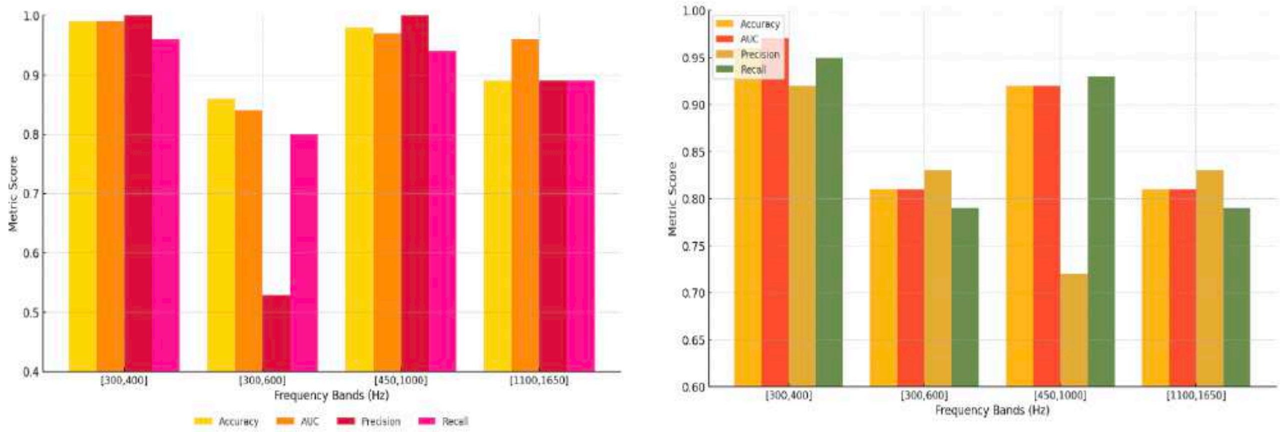
Figure 5(b) presents the comparison between symptomatic COVID-19-positive individuals and healthy individuals. While there remains a clear energy concentration in the lower frequency band for the COVID-19-positive group, the distinction is less pronounced than in Figure 5(a). The healthy group also exhibits some overlapping spectral activity in the 300–800 Hz band, possibly due to background noise or normal variations in cough effort. The reduced contrast suggests that the presence of symptoms may blur the boundary between infected and healthy states, complicating the use of frequency-domain features alone for accurate differentiation. This overlap implies that although COVID-19 symptoms alter cough acoustics, the resulting spectral footprint partially resembles that of benign coughs, necessitating more robust multi-dimensional features or temporal dynamics for improved separability.

In Figure 5(c), the comparison between symptomatic and asymptomatic COVID-19-positive individuals reveals an almost complete overlap in spectral energy distribution. Both groups share the same dominant low-frequency peaks, with no observable shifts in amplitude or frequency structure. This indicates that symptom presence or severity does not significantly affect the frequency characteristics of coughs in COVID-19-positive patients. The lack of spectral differentiation underscores the challenge of using traditional frequency-domain analysis for symptom classification or clinical staging within COVID-19 cohorts. This result also reinforces the idea that the acoustic manifestation of symptoms may not be captured through linear frequency features and instead may require nonlinear feature embeddings (e.g., Mel-spectrograms, wavelet transforms) or deep learning representations that can capture subtle temporal or multi-modal patterns.

4.2 Classification Performance of 2 strategies

To assess the effectiveness of different training strategies on cough-based COVID-19 classification, two XGBoost-based configurations were evaluated. As summarized in Table 2, Strategy A employed default hyperparameters with SMOTE for up-sampling but without threshold adjustment, while Strategy B incorporated Bayesian optimization, SMOTE, and threshold moving to enhance classifier sensitivity and robustness.

Performance was assessed across four frequency bands: [300–400] Hz, [300–600] Hz, [450–1000] Hz, and [1100–1650] Hz, using standard evaluation metrics: Accuracy, AUC, Precision, and Recall. Figure 6 displays the comparative results for Strategy A (left panel) and Strategy B (right panel), trained and tested on the combined dataset.



(a) Strategy A *(b) Strategy B*

Fig. 6. Performance metrics of the (a) XGBoost (Strategy A) and (b) XGBoost (Strategy B) classifiers across different frequency bands using the combined dataset, showing Accuracy, AUC, Precision, and Recall for each frequency range

Under Strategy A, the frequency bands [300–400] Hz and [450–1000] Hz yielded the highest performance. Both bands achieved accuracy and AUC values exceeding 0.95, with precision and recall consistently above 0.90. In contrast, the [300–600] Hz band demonstrated the weakest results, with AUC dropping to approximately 0.53 and recall falling to 0.80, indicating lower discriminative power in this frequency range under default settings. The [1100–1650] Hz band showed moderate performance with balanced scores between 0.88 and 0.92 across metrics.

Strategy B introduced notable improvements across all frequency bands, especially in mid-frequency ranges. The [450–1000] Hz band emerged as the best-performing frequency band overall, achieving high and stable results across all metrics—accuracy and AUC above 0.93, precision around 0.94, and recall exceeding 0.93. The [300–600] Hz band, previously underperforming, improved substantially with AUC and accuracy rising above 0.81, and recall approaching 0.79. The [300–400] Hz and [1100–1650] Hz bands also maintained strong and consistent performance, with minimal variation across metrics.

These findings confirm that [450–1000] Hz is the most discriminative and reliable frequency band for cough-based COVID-19 classification, particularly when optimized training strategies are employed. Strategy B demonstrated clear superiority over Strategy A by reducing variability across bands, improving recall in mid-frequency ranges, and enhancing overall generalization. The results underscore the importance of frequency-band-specific modeling, class imbalance mitigation, and threshold tuning for robust real-world deployment of cough-based diagnostic tools.

4.3 Comparative analysis of performance with state-of-the-art methods

To validate the effectiveness of the proposed approach to detect symptomatic cases, a comparative analysis was conducted against recent state-of-the-art methods using the Cambridge dataset, specifically focusing on symptomatic COVID-19 cases. As outlined in Table 3, the proposed system—an XGBoost classifier trained on the [450–1000] Hz frequency band, utilizing 193 spectral features, and enhanced by Bayesian Optimization (BO) and Threshold Moving (TM)—was evaluated alongside methods reported in [27]–[30].

Given that this study is explicitly centered on symptomatic individuals, the evaluation prioritizes high recall and precision, both critical in clinical diagnostics where symptomatic patients are at greater risk of progressing to severe outcomes. Method [27] achieved an AUC of 0.87, with a relatively low precision of 0.70 and recall of 0.90, indicating room for improvement in correctly identifying positive cases while avoiding false alarms. Method [28], though not reporting AUC, showed a higher precision of 0.87 but a weaker recall of 0.82, suggesting reduced sensitivity. Method [29] delivered a strong overall performance with an AUC of 0.98, precision of 0.94, and recall of 0.93. Similarly, method [30] reported an AUC of 0.93, precision of 0.89, and recall of 0.94.

Table 3
A comparison between the state-of-the-art methods and XGBoost + [450,1000] + 193 features+ Bayesian Optimization (BO) + Threshold Moving (TM)

Dataset	Category	Method	AUC	Precision	Recall
Cambridge	Symptomatic	[27]	0.87	0.70	0.90
		[28]	-	0.87	0.82
		[29]	0.98	0.94	0.93
		[30]	0.93	0.89	0.94
		Proposed Method	0.98	0.96	0.97

The proposed method outperformed or matched all comparative benchmarks, achieving an AUC of 0.98, a precision of 0.96, and a recall of 0.97. These metrics reflect both high discriminative power and balanced predictive capability, especially under symptomatic conditions where distinguishing COVID-19 coughs from other respiratory signals is more complex. The use of Bayesian Optimization allowed the model to converge on highly effective hyperparameter configurations, while threshold moving ensured that recall and precision remained balanced despite the inherent class imbalance in symptomatic data.

Because this study specifically targets symptomatic classification, the superior recall achieved by the proposed method is particularly noteworthy. In clinical screening applications, missing symptomatic cases poses greater health and transmission risks compared to asymptomatic misclassifications. The enhanced performance across all metrics therefore underscores the practical value of the proposed frequency-band-focused, optimization-enhanced XGBoost pipeline in real-world symptomatic COVID-19 detection.

5. Conclusions

The experimental results of the proposed frequency-band-based XGBoost model demonstrate that accurate COVID-19 detection from cough audio is achievable, particularly for symptomatic individuals. The best classification performance was observed in the [450–1000] Hz band, confirming its strong discriminative capability. By integrating Bayesian optimization, SMOTE, and threshold moving, the model achieved high precision (0.96) and recall (0.97), effectively addressing class imbalance and enhancing sensitivity. These findings support the role of cough sound characteristics—especially within mid-frequency ranges—as a viable digital biomarker for symptomatic COVID-19 detection. This responds directly to the objective of building a robust, frequency-specific, and symptom-sensitive screening tool. Further evaluation on larger and real-time datasets is required to validate and generalize its deployment in real-world healthcare settings.

Acknowledgement
This research was funded by a grant from University Putra Malaysia (GPS Grant IPS/2021/970000).

References

[1] Chung, Kian Fan, and Ian D. Pavord. "Prevalence, Pathogenesis, and Causes of Chronic Cough." *The Lancet* 371, no. 9621 (2008): 1364–1374. https://doi.org/10.1016/S0140-6736(08)60595-4.

[2] Morice, Alyn H., Eva Millqvist, Kristina Bieksiene, Surinder S. Birring, Peter Dicpinigaitis, Christian Domingo Ribas, Michele Hilton Boon, et al. "ERS Guidelines on the Diagnosis and Treatment of Chronic Cough in Adults and Children." *European Respiratory Journal* 55, no. 1 (2020). https://doi.org/10.1183/13993003.01136-2019.

[3] al Hamoud, Saleh. "Utility of CT Scan Evaluation for Predicting Pulmonary Hypertension." PhD diss., 2012.

[4] Rytter, Héloïse, Anne Jamet, Mathieu Coureuil, Alain Charbit, and Elodie Ramond. "Which Current and Novel Diagnostic Avenues for Bacterial Respiratory Diseases?" *Frontiers in Microbiology* 11 (2020): 616971. https://doi.org/10.3389/fmicb.2020.616971.

[5] Resten, Arnaud, Sophie Maitre, Marc Humbert, Anne Rabiller, Olivier Sitbon, Frédérique Capron, Gérald Simonneau, and Dominique Musset. "Pulmonary Hypertension: CT of the Chest in Pulmonary Venoocclusive Disease." *American Journal of Roentgenology* 183, no. 1 (2004): 65–70. https://doi.org/10.2214/AJR.183.1.1830065.

[6] Deschamp, Ashley R., Joseph E. Hatch, James E. Slaven, Netsanet Gebregziabher, Gregory Storch, Graham L. Hall, Stephen Stick, Sarath Ranganathan, Thomas W. Ferkol, and Stephanie D. Davis. "Early Respiratory Viral Infections in Infants with Cystic Fibrosis." *Journal of Cystic Fibrosis* 18, no. 6 (2019): 844–850. https://doi.org/10.1016/j.jcf.2019.02.004.

[7] Smith, Jaclyn A., H. Louise Ashurst, Sandy Jack, Ashley A. Woodcock, and John E. Earis. "The Description of Cough Sounds by Healthcare Professionals." *Cough* 2 (2006): 1–9. https://doi.org/10.1186/1745-9974-2-1.

[8] Ghrabli, S., M. Elgendi, and C. Menon. "Challenges and Opportunities of Deep Learning for Cough-Based COVID-19 Diagnosis: A Scoping Review." *Diagnostics* 12, no. 9 (2022): 2142.

[9] Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl. "Interpretable Machine Learning—A Brief History, State-of-the-Art and Challenges." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417–431. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-65965-3_28.

[10] Hashimoto, Yasushi, Akira Murata, Masashi Mikami, Seiichi Nakamura, Eiichi Yamanaka, and Shoji Kudoh. "Influence of the Rheological Properties of Airway Mucus on Cough Sound Generation." *Respirology* 8, no. 1 (2003): 45–51. https://doi.org/10.21037/jtd-2020-icc-003.

[11] Hall, Jocelin Isabel, Manuel Lozano, Luis Estrada-Petrocelli, Surinder Birring, and Richard Turner. "The Present and Future of Cough Counting Tools." *Journal of Thoracic Disease* 12, no. 9 (2020): 5207. https://doi.org/10.1046/j.1440-1843.2003.00432.x.

[12] Anupam, Arup, N. Jagan Mohan, Sudarsan Sahoo, and Sudipta Chakraborty. "Preliminary Diagnosis of COVID-19 Based on Cough Sounds Using Machine Learning Algorithms." In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1391–1397. IEEE, 2021. https://doi.org/10.1109/ICICCS51141.2021.9432324.

[13] Islam, Rumana, Esam Abdel-Raheem, and Mohammed Tarique. "Early Detection of COVID-19 Patients Using Chromagram Features of Cough Sound Recordings with Machine Learning Algorithms." In *2021 International Conference on Microelectronics (ICM)*, 82–85. IEEE, 2021. https://doi.org/10.1109/ICM52667.2021.9664931.

[14] Melek Manshouri, Negin. "Identifying COVID-19 by Using Spectral Analysis of Cough Recordings: A Distinctive Classification Study." *Cognitive Neurodynamics* 16, no. 1 (2022): 239–253. https://doi.org/10.1007/s11571-021-09695-w.

[15] Södergren, Isabella, Maryam Pahlavan Nodeh, Prakash Chandra Chhipa, Konstantina Nikolaidou, and György Kovács. "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines." In *Interspeech*, 916–920. 2021. https://doi.org/10.21437/Interspeech.2021-2191.

[16] Brown, Chloë, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3474–3484. 2020. https://doi.org/10.1145/3394486.3412865.

[17] Kho, Sarah Jane, Brian Loh Chung Shiong, Vong Wan-Tze, Law Kian Boon, Mohan Dass Pathmanathan, Mohd Aizuddin Bin Abdul Rahman, Kuan Pei Xuan, Wan Nabila Binti Wan Hanafi, Kalaiarasu M. Peariasamy, and Patrick Then Hang Hui. "Malaysian Cough Sound Analysis and COVID-19 Classification with Deep Learning." *Intelligence-Based Medicine* 9 (2024): 100129. https://doi.org/10.1016/j.ibmed.2023.100129.

[18] Xu, Xuhai, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. "Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, no. 1 (2021): 1–22. https://doi.org/10.1145/3448124.

[19] Lee, Gyeong-Tae, Hyeonuk Nam, Seong-Hu Kim, Sang-Min Choi, Youngkey Kim, and Yong-Hwa Park. "Deep Learning Based Cough Detection Camera Using Enhanced Features." *Expert Systems with Applications* 206 (2022): 117811. https://doi.org/10.1016/j.eswa.2022.117811.

[20] Chen, Xinru, Menghan Hu, and Guangtao Zhai. "Cough Detection Using Selected Informative Features from Audio Signals." In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–6. IEEE, 2021. https://doi.org/10.1109/CISP-BMEI53629.2021.9624379.

[21] Soliński, Mateusz, Michał Łepek, and Łukasz Kołtowski. "Automatic Cough Detection Based on Airflow Signals for Portable Spirometry System." *Informatics in Medicine Unlocked* 18 (2020): 100313. https://doi.org/10.1016/j.imu.2020.100313.

[22] Chuma, Euclides Lourenco, and Yuzo Iano. "A Movement Detection System Using Continuous-Wave Doppler Radar Sensor and Convolutional Neural Network to Detect Cough and Other Gestures." *IEEE Sensors Journal* 21, no. 3 (2020): 2921–2928. https://doi.org/10.1109/JSEN.2020.3028494.

[23] Chowdhury, Muhammad EH, Nabil Ibtehaz, Tawsifur Rahman, Yosra Magdi Salih Mekki, Yazan Qibalwey, Sakib Mahmud, Maymouna Ezeddin, Susu Zughaier, and Sumaya Ali SA Al-Maadeed. "QUCoughScope: An Artificially Intelligent Mobile Application to Detect Asymptomatic COVID-19 Patients Using Cough and Breathing Sounds." *arXiv* preprint arXiv:2103.12063 (2021). https://doi.org/10.48550/arXiv.2103.12063.

[24] Pahar, Madhurananda, Marisa Klopper, Robin Warren, and Thomas Niesler. "COVID-19 Cough Classification Using Machine Learning and Global Smartphone Recordings." *Computers in Biology and Medicine* 135 (2021): 104572. https://doi.org/10.1016/j.compbiomed.2021.104572.

[25] Nessiem, Mina A., Mostafa M. Mohamed, Harry Coppock, Alexander Gaskell, and Björn W. Schuller. "Detecting COVID-19 from Breathing and Coughing Sounds Using Deep Neural Networks." In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 183–188. IEEE, 2021. https://doi.org/10.1109/CBMS52027.2021.00069.

[26] Lella, Kranthi Kumar, and Alphonse Pja. "Automatic Diagnosis of COVID-19 Disease Using Deep Convolutional Neural Network with Multi-Feature Channel from Respiratory Sound Data: Cough, Voice, and Breath." *Alexandria Engineering Journal* 61, no. 2 (2022): 1319–1334. https://doi.org/10.1016/j.aej.2021.06.024.

[27] Brown, Chloë, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3474–3484. 2020. https://doi.org/10.1145/3394486.3412865.

[28] Aytekin, Idil, Onat Dalmaz, Kaan Gonc, Haydar Ankishan, Emine Ulku Saritas, Ulas Bagci, Haydar Celik, and Tolga Çukur. "COVID-19 Detection from Respiratory Sounds with Hierarchical Spectrogram Transformers." *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2023): 1273–1284. https://doi.org/10.1109/JBHI.2023.3339700.

[29] Chowdhury, Nihad Karim, 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 (2022): 105405. https://doi.org/10.1016/j.compbiomed.2022.105405.

[30] Akman, Alican, Harry Coppock, Alexander Gaskell, Panagiotis Tzirakis, Lyn Jones, and Björn W. Schuller. "Evaluating the COVID-19 Identification ResNet (CIDER) on the Interspeech COVID-19 from Audio Challenges." *Frontiers in Digital Health* 4 (2022): 789980. https://doi.org/10.3389/fdgth.2022.789980.