Robust COVID-19 Detection from Cough Sounds using Deep Neural Decision Trees and Forests: A Comprehensive Cross-Dataset Evaluation

ROFIQUL ISLAM, Department of Computer Science and Engineering, University of Chittagong, Bangladesh
NIHAD KARIM CHOWDHURY*, Department of Computer Science and Engineering, University of Chittagong, Bangladesh
ASHAD KABIR, School of Computing, Mathematics, and Engineering, Charles Sturt University, Australia

This research introduces robust COVID-19 cough classification using advanced ML techniques, showcasing consistent performance across diverse datasets with deep neural decision trees and forests. We comprehensively extract features to capture diverse audio traits from both COVID-19 positive and negative individuals. Next, we use recursive feature elimination with cross-validation to find critical features, followed by Bayesian optimization to fine-tune hyper-parameters of deep neural decision trees and forests. Moreover, we incorporate synthetic minority oversampling during training to achieve balanced representation of positive and negative data. Model performance is refined via threshold optimization to maximize the ROC-AUC score. Our approach is comprehensively evaluated across five datasets: Cambridge, Coswara, COUGHVID, Virufy, and the combined Virufy with NoCoCoDa. Consistently outperforming state-of-the-art methods, our proposed approach yields AUC scores of 0.89, 0.95, 0.73, 0.69, 0.97, and 0.99, alongside recall scores of 0.90, 0.93, 0.78, 0.74, 1, and 0.98 across the respective datasets. Combining all datasets into one, our method, using a deep neural decision tree classifier, achieves an accuracy of 0.76, AUC of 0.75, precision of 0.65, recall of 0.69, F1-score of 0.66, and specificity score of 0.80. Also, our study conducts an extensive cross-dataset analysis, unveiling demographic and geographic variations in COVID-19-associated cough sounds.

CCS Concepts: • **Computing methodologies → Supervised learning by classification**; **Machine learning approaches**; • **Applied computing → Computational biology**; **Consumer health**; **Health care information systems**; **Health informatics**.

Additional Key Words and Phrases: Bayesian Optimization, Cough, COVID-19, Cross Dataset Study, Feature Extraction, Feature Selection, RFECV, SMOTE, Threshold Moving

ACM Reference Format:
Rofiqul Islam, Nihad Karim Chowdhury, and Ashad Kabir. 2024. Robust COVID-19 Detection from Cough Sounds using Deep Neural Decision Trees and Forests: A Comprehensive Cross-Dataset Evaluation. *ACM Trans. Comput. Healthcare* XX, X, Article XXX (May 2024), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding Author

Authors' Contact Information: Rofiqul Islam, Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh, rofiqcsecu101@gmail.com; Nihad Karim Chowdhury, Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh, nihad@cu.ac.bd; Ashad Kabir, School of Computing, Mathematics, and Engineering, Charles Sturt University, Bathurst, NSW, Australia, akabir@csu.edu.au.

ACM 2637-8051/2024/5-ARTXXX
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

COVID-19 is a highly contagious disease caused by the SARS-CoV-2 virus. It can cause serious illness, death, and economic disruption. The surge in COVID-19 cases placed an overwhelming burden on medical diagnostic laboratories and healthcare facilities, underscoring the limitations of conventional diagnostic methods such as clinical examinations, CT scans, and PCR tests [2]. Although PCR is a highly accurate method for COVID-19 detection, its cost and time requirements render it inaccessible to a significant portion of the population [113]. Furthermore, the utility of CT scan imaging for COVID-19 diagnosis is constrained by the potential overlap of symptoms with influenza, delayed detection of lung manifestations in early stages of infection, and restrictions on its use, notably for infants and pregnant women [56, 64].

Studies have indicated the potential of utilizing the human voice as a primary diagnostic tool for conditions associated with voice production, including the detection of pathological voice [40], pertussis [87], asthma [4], and respiratory diseases [104]. These instances underscore the promising capabilities of the human voice in diagnosing diseases, particularly those related to voice production. Researchers have explored acoustic analysis as a means to extend COVID-19 detection beyond voice-related conditions. They have investigated various modalities, including cough sounds, breathing patterns, and speech or voice, encompassing vowels. The initial two modalities present indicators for the symptoms of COVID-19, namely persistent coughing and breathlessness. We opted for the first modality due to the substantial volume of available cough data. Numerous universities around the world, such as the University of Cambridge in the UK [18], the Massachusetts Institute of Technology (MIT) in the United States [23], and the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland [77], are actively involved in researching the application of machine learning (ML) techniques for the diagnosis of COVID-19 using cough sound data.

To expedite research focused on the detection of COVID-19 through cough sound analysis, we introduce an approach that leverages deep neural decision trees (DNDT) and deep neural decision forests (DNDF). Initially, audio samples are pre-processed to extract various audio characteristics from cough sounds of individuals with confirmed positive and negative results of the COVID-19 test. Subsequently, we employ Recursive Feature Elimination with Cross-Validation (RFECV) in combination with the Extra Trees classifier to identify the most critical features for our classifiers. Following this, Bayesian optimization (BO) is utilized to fine-tune the hyperparameters of our proposed method. To enhance the classification performance of our models, we incorporate threshold moving (TM) techniques to ascertain the optimal threshold value. The dataset prominently displays an imbalance, notably with a limited representation of positive instances for COVID-19, potentially posing a detrimental effect on the performance of the ML classifier. In response to this imbalance, we have implemented the synthetic minority oversampling technique (SMOTE) [24] during the training process. This strategic inclusion aims to rectify the dataset imbalance, thereby enhancing the performance of the ML classifier.

The evaluation of our classification models extends across multiple datasets, including Cambridge [18], Coswara [95], COUGHVID [77], Virufy [23], and Virufy merged with NoCoCoDa [27]. Additionally, we conduct a comprehensive cross-dataset study, in which our proposed approach is trained on one dataset and its performance is evaluated across several other datasets. Furthermore, all the five datasets are consolidated into a combined dataset, which serves as the basis for the COVID-19 classification. We conducted extensive experiments on this unified dataset to evaluate the effectiveness of our proposed method, thoroughly evaluating its performance and robustness across diverse data sources to ensure comprehensive validation of our approach. The key contributions of this paper are summarized as follows:

- We propose a pioneering method utilizing deep neural decision trees and decision forests tailored specifically for detecting COVID-19 from cough sounds. This approach offers a unique perspective on leveraging machine learning techniques for early diagnosis.

- To enhance prediction performance, we employ advanced feature dimension reduction techniques, including recursive feature elimination with cross-validation and the extra trees classifier. By identifying and prioritizing the most informative features, we significantly boost the accuracy and reliability of our detection model.
- We explore five distinct training strategies encompassing various frameworks to optimize the efficacy of our detection model. Leveraging Bayesian Optimization, we systematically evaluate each strategy to identify the most effective approach, ensuring robust performance across different scenarios.
- To validate the effectiveness and generalizability of our proposed approach, we conduct extensive evaluations on five diverse cough datasets. This rigorous evaluation framework ensures the reliability of our findings and underscores the versatility of our approach across different data sources.
- We empirically assess the performance of our proposed approach by benchmarking it against state-of-the-art models for distinguishing COVID-19 from non-COVID-19 cases. Through comprehensive comparative analyses, we demonstrate the superiority of our method in terms of accuracy and efficiency.
- To assess the generalizability of our approach, we perform cross-dataset evaluations wherein the model is trained on one dataset and evaluated on four additional datasets. This rigorous evaluation paradigm ensures the robustness and versatility of our detection method across diverse data sources and scenarios.

The remainder of the paper is organized as follows. Section 2 presents the related work, while Section 3 outlines the research questions addressed in this study. Section 4 details the methodology and provides an in-depth explanation of our proposed approach. Section 5 presents the experimental results. Finally, Section 6 concludes the article summarizing the key findings and directions for future research.

2 RELATED WORK

Numerous researchers have concentrated on detecting COVID-19 using sounds such as coughs, breath, voice, and speech [1, 5, 31, 39, 46, 49, 55, 62, 93]. In this study, we focus specifically on research related to cough sounds. Many researchers have invested significant efforts in developing datasets, including Cambridge [18], Coswara [95], COUGHVID [77], Virufy [23], Novel Coronavirus Cough Database (NoCoCoDa) [27], Cough against COVID [14], AI4COVID-19 [50], MIT-Covid-19 [103], IATos [86], Sarcos [82], ComParE [94], COVID-19 Cough [60], COVID-19 Sounds [112], and DICOVA Challenge [72]. Next, we investigate studies exclusively focusing on the utilization of cough sounds for the classification of COVID-19. We explore the methodologies, findings, and implications of these studies to gain a comprehensive understanding of the role and effectiveness of cough-based approaches in COVID-19 detection.

There are two classifications of COVID-19 datasets: publicly available and those not accessible to the public. Publicly available datasets include Coswara [95], COUGHVID [77], Virufy [23], IATos [86], and others. Conversely, datasets that are not publicly available include Cambridge [18], COVID-19 Sounds [112], DICOVA Challenge [72], Novel Coronavirus Cough Database (NoCoCoDa) [27], Cough against COVID [14], AI4COVID-19 [50], MIT-Covid-19 [103], Sarcos [82], ComParE [94], COVID-19 Cough [60], and others.

A comparison of our research and previous studies on COVID-19 identification using cough sound analysis is shown in Table 1. It is worth noting that some researchers limited their analyses to specific datasets, such as Cambridge [9, 18, 25, 63], Coswara [6, 8, 15, 16, 48, 100], COUGHVID [10, 11, 19, 29, 36, 37, 42, 68, 78, 92, 98, 109], and Virufy [35, 52, 53, 101]. A few others use only their proprietary datasets [7, 47, 50, 74, 83, 116]. Furthermore, in addition to using well-known datasets like Cambridge, Coswara, COUGHVID, and Virufy, some research included the use of other publicly and non-publicly accessible datasets [13, 20, 32, 33, 59, 61, 65, 71, 76, 91, 107, 111, 115, 118] to validate the robustness of their classification models.

Numerous research works have opted to evaluate the effectiveness of their suggested techniques by utilizing several datasets. Of these, a few research worked with two datasets [12, 14, 38, 45, 51, 66, 67, 70, 80, 84, 85, 88–90, 96, 105, 110, 114, 117, 119], while others evaluated from three [57, 75, 79, 81, 97, 99, 102] or four datasets [21, 26, 43, 106, 108]. Nevertheless, what makes our work noteworthy is that it is the only one that fully incorporates each of the five well-known COVID-19 cough datasets: Cambridge, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa dataset. In addition, some researchers choose to take a combined strategy, combining two or more datasets in order to strengthen the validity of their research. Among these research, several combined two datasets [67, 70, 88, 89, 108], while several investigated combining three datasets [75, 79] or four datasets [26, 43, 66, 106] for a more thorough analysis. It is imperative to emphasize that, except from our research, no one else has attempted the thorough integration of the five well-known COVID-19 cough datasets: Cambridge, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa dataset.

The methodology of the cross-dataset study involves sequentially training the proposed method on specific datasets, one dataset used for training at a time, and subsequently evaluating its effectiveness across diverse datasets, excluding those employed during the training phase. Several studies have undertaken partial cross-dataset investigations, where the proposed method is trained on a specified dataset and subsequent assessment across either a singular dataset or a variety of datasets different from the original training dataset. The researchers in Ulukaya et al. [108] trained their suggested technique using a combined dataset that included COUGHVID and Coswara. Following the training phase, they conducted separate evaluations of their method using the Virufy and NoCoCoDa datasets. In the study conducted by Pahar et al. [80], the researchers developed their suggested approach using the Coswara dataset and then tested it on the Sarcos dataset. Accordingly, the Coswara dataset was used by Zhang et al. [119] and Feng et al. [38] to design and train their respective methods, with the Virufy dataset being used for performance assessments thereafter. Similar to this, Nguyen et al. [76] developed their suggested model with a dataset that was made available to the public and evaluated its performance with the AICovidVN dataset. The studies mentioned earlier do not utilize all of their datasets for cross-dataset investigations. Instead, they choose either a single dataset or a combined dataset for training, and one or two distinct datasets for testing. In contrast, our research stands out as it performs a distinctive cross-dataset analysis using all five datasets (Cambridge, Coswara, COUGHVID, Virufy, and Virufy with NoCoCoDa). Notably, we employ one of the five datasets for training our proposed methods, while the remaining four datasets are independently used to validate each method.

Audio features are commonly classified into different types, encompassing time domain features like Zero Crossing Rate (ZCR), Energy, Amplitude based features, Root Mean Square (RMS) Energy, etc. Additionally, frequency domain features include Power Spectral Density (PSD), Spectral Bandwidth, Spectral Contrast, Spectral Centroid, Spectral Roll-Off, Spectral Kurtosis, Spectral Flux, Spectral Spread, Chromagram, Tonal Centroid, and others. Time-frequency representations involve Spectrogram, Mel-Spectrogram, Constant-Q Transform, Scattering Embeddings, etc. Cepstral domain features comprise MFCC, Δ-MFCC, Δ^2-MFCC, Inverted Mel-Frequency Cepstral Coefficients (IMFCCs), Gammatone Frequency Cepstral Coefficients (GFCC), Gammatone Cepstral Coefficients (GTCCs), and so forth. Deep features include characteristics obtained through deep learning models such as Convolutional Neural Networks (CNNs), YAMNet, VGGish, and similar architectures.

In the field of COVID-19 classification, studies use a variety of audio features for COVID-19 classification. Certain studies, such as [19, 52], focus solely on frequency domain features, providing useful insights into the different patterns found in this domain. Studies such as [6, 12, 20, 21, 25, 35, 36, 42, 68, 75, 88, 89, 91, 97, 106, 111, 115], on the other hand, focus primarily on time-frequency representations, uncovering the unique properties hidden in these representations. Another set of investigations, exemplified by [9, 15, 16, 45, 59, 61, 63, 65–67, 71, 96, 101, 118], focus their attention on cepstral domain features, shedding light on the pertinent information concealed within cepstral analyses. Furthermore, a subset of researchers opts for deep features, harnessing the power of deep learning techniques to uncover intricate patterns within the audio data [18, 110].

A significant number of studies are adopting a comprehensive approach by integrating multiple types of audio features into their analyses. For instance, studies such as [7, 13, 23, 32, 41, 50, 57, 76, 84, 107, 114, 119] incorporate two different types of audio features. In contrast, [8, 10, 11, 14, 18, 26, 38, 43, 47, 51, 53, 78, 80, 81, 83, 102, 105, 108, 110] employ three different types of features in their studies. Notably, studies [70, 74, 79, 90, 109] utilizing a combination of four distinct types of audio features in their research. In our study, we utilized three distinct feature categories. These include features derived from the frequency domain, such as tonal centroid, chromagram, and spectral contrast; features from the cepstral domain, such as Mel-Frequency Cepstral Coefficients (MFCC); and features from the time-frequency representation domain, such as the Mel-Scaled Spectrogram. This approach diversity exemplifies the detailed examination of numerous feature domains in the pursuit of robust COVID-19 classification models.

The technique of Optimal Feature Selection has been utilized in numerous research investigations. To choose the best features for training classification models, some studies [26, 78, 79, 106] used the Recursive Feature Elimination with Cross-Validation (RFECV) technique. To determine the optimal feature set, the Sequential Forward Selection (SFS) technique has been used in a number of research [67, 80, 81]. A small number of studies [15, 35] selected the best features using the Relief feature selection approach. A variety of feature selection techniques, including the Modified Cat and Mouse Based Optimizer (MCMBO), Stacked autoencoder architecture, Kruskal-Wallis test, Mutual information criterion, principal component analysis (PCA), and GridSearch, were used by a number of other studies [10, 13, 16, 23, 33, 70, 98] to optimize their feature sets. In this work, we use an Extra Trees classifier in conjunction with Recursive Feature Elimination with Cross-Validation (RFECV) to perform optimal feature selection. The choice to utilize RFECV is substantiated by the results presented in a study by [69], where they illustrated improved classification accuracy through the use of Recursive Feature Elimination with Cross-Validation (RFECV).

To determine which parameters are optimal for their classification models, some researchers have performed hyperparameter tuning. Numerous studies [9, 23, 26, 29, 33, 36, 42, 84, 90, 98, 101, 108] employed Grid Search for hyperparameter tuning. Optimal hyper-parameters for their suggested models are selected by a cross-validation procedure in a few studies [12, 14, 50, 114]. The optimal hyper-parameters were chosen using 5-fold cross-validation in the research by [92] and [100]. Nested cross-validation has been used in certain studies [18, 81] as a technique for hyperparameter tuning in their analyses. Hyperparameter optimization was done using a variety of methods in a number of other studies [10, 11, 67, 78, 80]. These methods included MCMBO (Modified Cat and Mouse Based Optimizer), LOOCV (Leave-One-Out Cross-Validation), The Leave-p-out cross-validation, Leave-p-out nested cross-validation, and Mayfly optimization (MFO). Bayesian Optimization (BO) is more effective than more brute-force methods like Grid Search (GS) and Random Search (RS) in determining the ideal hyper-parameter combination [34].

Techniques for shifting thresholds have been used in various studies. Threshold moving based on accuracy scores was used in the studies by Hamdi et al. [42]. ROC-AUC values were used to determine threshold movement in the [26] study. The study [14] chose a threshold based on which the maximum Sensitivity score was obtained. In their individual investigations, Zhang et al. [118] and Mouawad et al. [71] both used threshold moving based on F1 score. In our study, we employ the threshold moving technique based on ROC-AUC scores as it proficiently strikes a balance between precision and recall.

Unlike prior investigations, our study takes a method by thoroughly assessing model performance using cross-dataset study, a facet overlooked by previous research efforts. Additionally, we extend our analysis beyond the traditional scope of testing model performance across diverse datasets. Delving into the intricacies of training strategies, we scrutinize how different training methodologies significantly influence the overall performance of our models. This comprehensive examination ensures a holistic understanding of our models' capabilities, emphasizing the importance of both testing across varied datasets and adopting nuanced training strategies for robust and informed outcomes.

Table 1. A comparison to previous studies related to COVID-19 detection from cough sounds.

References	Dataset			Feature Extraction	Approach		
	Name	M	CDS		FS	HT	TM
Orlandic et al. [78]	COUGHVID	-	-	Power Spectral Density (PSD), Zero Crossing Rate (ZCR), MFCC, etc	✓	✓	✗
Sobahi et al. [99]	COUGHVID,Virufy, Coswara	✗	✗	Fractal Dimensions (FD)	✗	✗	✗
Rayan et al. [91]	MIT-Covid-19	-	-	Mel-Spectrogram	✗	✗	✗
Awais et al. [10]	COUGHVID	-	-	MFCC, Spectral and Statistical Features	✓	✓	✗
Aytekin et al. [12]	Cambridge, COUGHVID	✗	✗	Mel-Spectrogram	✗	✓	✗
Padmalatha et al. [79]	Own dataset, COUGHVID, Cambridge	✓	✗	MFCC, Δ-MFCC, Δ^2-MFCC, ZCR, etc	✓	✗	✗
Pavel and Ciocoiu [85]	COUGHVID, COVID-19 Sounds	✗	✗	YAMNet, MFCC, VGGish, x-Vecs	✗	✗	✗
Kim et al. [57]	Cambridge, Coswara, COUGHVID	✗	✗	MFCC, Δ-MFCC, Δ^2-MFCC, Spectral Contrast	✗	✗	✗
Son and Lee [102]	COUGHVID, Cambridge, Coswara	✗	✗	MFCC, Spectrogram, Spectral Centroid, etc	✗	✗	✗
Soltanian and Borna [101]	Virufy	-	-	MFCCs	✗	✓	✗
Aly and Alotaibi [6]	Coswara	-	-	Mel-Scale Spectrogram	✗	✗	✗
Rahman et al. [88]	Cambridge, Qatari	✓	✗	Spectrogram	✗	✗	✗
Ulukaya et al. [108]	COUGHVID, Coswara, Virufy, NoCoCoDa	✓	✗	MFCC, Spectrogram, Chromagram	✗	✓	✗
Wang et al. [110]	AICovidVN, Cambridge	✗	✗	Mel Spectrum, MFCC, VGG Embeddings	✗	✗	✗
Nguyen et al. [75]	Own dataset, Coswara, COUGHVID	✓	✗	Mel Spectrogram	✗	✗	✗
Islam et al. [53]	Virufy	-	-	Time and Frequency Domain Features	✗	✗	✗
Hamdi et al. [42]	COUGHVID	-	-	Mel-Spectrogram	✗	✓	✓
Ren et al. [92]	COUGHVID	-	-	OpenSMILE	✗	✓	✗
Andreu-Perez et al. [7]	Own dataset	-	-	Mel-Scaled Spectrogram, Linear Predictive Coding Spectrum (LPCS), MFCC	✗	✗	✗
Zealouk et al. [116]	Own dataset	-	-	Hidden Markov Model (HMM), Gaussian Mixture Distributions (GMMs), MFCC	✗	✗	✗
Xue and Salim [114]	Coswara, COVID-19 Sounds	✗	✗	MFCC, Log compressed mel-filterbank	✗	✓	✗
Zewail et al. [117]	Virufy, Cambridge	✗	✗	Deep Wavelet Scattering Network (DWSN)	✗	✗	✗
Skander et al. [98]	COUGHVID	-	-	OpenSMILE	✓	✓	✗
Hemdan et al. [48]	Coswara	-	-	Genetic Algorithm (GA)	✗	✗	✗
Södergren et al. [100]	Coswara	-	-	OpenSMILE	✗	✓	✗
Melek [67]	Virufy, NoCoCoDa	✓	✗	MFCC	✓	✓	✗
Pahar et al. [80]	Coswara, Sarcos	✗	✗	MFCC, Δ-MFCC, Δ^2-MFCC, ZCR, etc	✓	✓	✗
Pahar et al. [81]	Coswara, ComParE, Sarcos	✗	✗	MFCC, Δ-MFCC, Δ^2-MFCC, ZCR, etc	✓	✓	✗

Table 1 – *(Continued .....)*

References	Dataset			Feature Extraction	Approach		
	Name	M	CDS		FS	HT	TM
Chowdhury et al. [26]	Cambridge, Coswara Virufy, NoCoCoDa	✓	✗	MFCC, Chromagram, Tonal Centroid, Spectral Contrast, Mel-Scaled Spectrogram	✓	✓	✓
Vinod et al. [109]	COUGHVID	-	-	ZCR, MFCC, Chroma STFT, Roll-Off, Spectral Centroid, Spectral Bandwidth	✗	✗	✗
Brown et al. [18]	Cambridge	-	-	MFCC, Δ-MFCC, Δ^2-MFCC, RMS Energy, Spectral Centroid, Roll-Off Frequency, etc	✗	✓	✗
Hassan et al. [47]	Own dataset	-	-	MFCC, Δ-MFCC, Δ^2-MFCC, ZCR, Spectral Centroid, Spectral Roll-Off	✗	✗	✗
Despotović et al. [33]	CDCVA	-	-	GeMaps, eGeMaps, ComParE	✓	✓	✗
Mohammed et al. [70]	Coswara, Virufy	✓	✗	MFCC, Spectrogram, Power Spectrum, Chroma, etc	✓	✗	✗
Chaudhari et al. [23]	Coswara, COUGHVID, Virufy	✗	✗	MFCCs, Mel-Spectrogram	✓	✓	✗
Fakhry et al. [37]	COUGHVID	-	-	MFCC, Mel-Spectrogram, Clinical Features	✗	✗	✗
Laguarta et al. [61]	COVID-19 Cough	-	-	MFCC	✗	✗	✗
Bagad et al. [14]	Cough against COVID, Coswara	✗	✗	MFCC, Mel-Spectrogram, RMS Energy, Tempo	✗	✓	✓
Imran et al. [50]	Own dataset	-	-	MFCC, Mel-Spectrogram	✗	✓	✗
Tena et al. [106]	Lleida, Cambridge, Coswara,Virufy	✓	✗	Time–frequency Features	✓	✗	✗
Nasab et al. [74]	Own dataset	-	-	MFCC, Δ-MFCC, Δ^2-MFCC, Chromagram, ZCR, etc	✗	✗	✗
Chowdhury et al. [25]	Cambridge	-	-	Spectrogram	✗	✗	✗
Lella and Pja [63]	Cambridge	-	-	De-noising Auto Encoder (DAE), Gamma-tone Frequency Cepstral Coefficients (GFCC), Improved Multi-frequency Cepstral Coefficients (IMFCC)	✗	✗	✗
Pal and Sankarasubbu [83]	Own dataset	-	-	MFCC, ZCR, Kurtosis, etc	✗	✗	✗
Rao et al. [90]	DICOVA, COUGHVID	✗	✗	MFCC Δ-MFCC, Δ^2-MFCC, RMS energy, ZCR, etc	✗	✓	✗
Mouawad et al. [71]	Voca.ai and Carnegie Mellon University	-	-	MFCC	✗	✗	✓
Gupta et al. [41]	COUGHVID	-	-	MFCC, Mel Frequency Spectrogram	✗	✗	✗
Zhang et al. [119]	Coswara, Virufy	✗	✗	Log Mel Spectrograms, Time-frequency Differential Feature, Energy Ratio Feature	✗	✗	✗
Shen et al. [97]	ComParE, Coswara, Cambridge	✗	✗	Log Mel Spectrograms, Time-frequency Differential Feature	✗	✗	✗
Pavel and Ciocoiu [84]	COUGHVID, IATos	✗	✗	MFCC, Mel Spectrogram	✗	✓	✗
Ashby et al. [9]	Cambridge	-	-	MFCC	✗	✓	✗

Table 1 – *(Continued .....)*

References	Dataset			Feature Extraction	Approach		
	Name	M	CDS		FS	HT	TM
Ayappan and Anila [11]	COUGHVID	-	-	MFCC, Log Frame Energies, ZCR, Kurtosis	✗	✓	✗
Trang et al. [107]	AICV115M	-	-	MFCC, Δ-MFCC, Δ^2-MFCC, Log Frame Energies	✗	✗	✗
Meng et al. [68]	COUGHVID	-	-	Mel Spectrogram	✗	✗	✗
Malviya et al. [65]	Pfizer digital medicine challenge	-	-	MFCC	✗	✓	✗
Harvill et al. [45]	COUGHVID, DICOVA	✗	✗	MFCC, Δ-MFCC, Δ^2-MFCC	✗	✗	✗
Chang et al. [21]	Flusense, COUGHVID, ComParE, DICOVA	✗	✗	Log Mel Spectrogram	✗	✗	✗
Zhang et al. [118]	Biovitals	-	-	MFCCs	✗	✗	✓
Cesarelli et al. [19]	COUGHVID	-	-	Spectral Roll-Off	✗	✗	✗
Yan et al. [115]	ComParE	-	-	3-channel Log-mel Spectrograms	✗	✓	✗
Haritaoglu et al. [43]	COUGHVID, Coswara, Virufy, IATos	✓	✗	MFCC, Δ-MFCC, Δ^2-MFCC, Spectral Centroid, etc	✗	✓	✓
Tawfik et al. [105]	Coswara, Virufy	✗	✗	MFCC, Chroma, ZCR, Spectral Centroid, Spectral Roll-Off, Spectral Bandwidth, Constant Q Transform	✗	✗	✗
Erdoğan and Narin [35]	Virufy	-	-	Features on Intrinsic Mode Functions (IMFs)	✓	✗	✗
Anupam et al. [8]	Coswara	-	-	MFCCs, Spectral Centroid, Spectral Roll-Off Point, Spectral Kurtosis, ZCR	✗	✓	✗
Rao et al. [89]	DICOVA, COUGHVID	✓	✗	Spectrograms	✗	✓	✗
Islam et al. [52]	Virufy	-	-	Chromagram	✗	✗	✗
Nguyen et al. [76]	AICovidVN	-	-	Log Mel Spectrograms, Wavegram-Log-Mel-CNN	✗	✗	✗
Esposito et al. [36]	COUGHVID	-	-	Log Mel Spectrograms	✗	✓	✗
Irawati and Zakaria [51]	Virufy, Coswara	✗	✗	MFCC, Chroma, ZCR, Spectral Centroid, Spectral Bandwidth, Spectral Roll-Off, Root Mean Square	✓	✓	✗
Shen et al. [96]	DICOVA, Own dataset	✗	✗	MFCC, Δ-MFCC, Δ^2-MFCC	✗	✗	✗
Benmalek et al. [15]	Coswara	-	-	MFCC, Gamma-tone Cepstral Coefficients (GTCC)	✓	✗	✗
Kumawat et al. [59]	DICOVA	-	-	MFCC, Δ-MFCC, Δ^2-MFCC	✗	✗	✗
Benmalek et al. [16]	Coswara	-	-	MFCC	✓	✗	✗
Mehta et al. [66]	Coswara, COUGHVID, Virufy, Owndataset	✓	✗	MFCC	✗	✗	✗
Chang et al. [20]	DICOVA	-	-	Mel Spectogram	✗	✗	✗
Wullenweber et al. [111]	DICOVA	-	-	Spectrogram	✗	✗	✗
Feng et al. [38]	Coswara, Virufy	✗	✗	MFCC, Energy, Entropy of Energy, ZCR, Spectral Centriod, Spectral Spread, Spectral Entropy, Spectral Flux	✗	✓	✗

Table 1 – *(Continued .....)*

References	Dataset			Feature Extraction	Approach		
	Name	M	CDS		FS	HT	TM
Our study	Cambridge, Coswara, COUGHVID, Virufy, NoCoCoDa	✓	✓	MFCC, Chromagram, Tonal Centroid, Spectral Contrast, Mel-Scaled Spectrogram	✓	✓	✓

M- Merged, **CDS**- Cross Dataset Study, **FS**- Feature Selection, **HT**- Hyper-parameters Tuning and **TM**- Threshold Moving.

3 RESEARCH QUESTIONS

In our pursuit of advancing the cutting-edge in cough-based COVID-19 detection, we present a ML-based architecture tailored for the analysis of cough sounds. Furthermore, our research focuses on pinpointing the most successful techniques for precise COVID-19 detection using cough data. As a result, we have devised a series of research questions (RQs) dedicated to the field of cough-based COVID-19 detection:

- **RQ1:** How do different training strategies impact the classification performance of detecting COVID-19 from cough sounds?
  — We provide various training strategies to enhance the effectiveness of the proposed method. The significance of these training techniques is elucidated in Section 4.7, with a detailed analysis presented in Section 5.2.
- **RQ2:** Should the construction of detection models take into account any demographic or geographic variations in cough sounds associated with COVID-19?
  — Yes, demographic and geographic variations in cough sounds related to COVID-19 exist. These variations should be considered when developing detection models to ensure that they are effective and generalizable across different populations. We explore the specifics of this cross-dataset investigation and its resulting findings in Section 5.4.

4 METHODOLOGY

Inspired by the advancements in ML-based audio applications, we have created a comprehensive ML framework capable of taking cough samples and making direct predictions of binary classification labels, hinting at the potential presence of COVID-19. Figure 1 displays the overall system methodology for the COVID-19 detection from cough sound system developed in this study. This method comprises several stages, including data collection, preprocessing, feature extraction, and model evaluation. During the data collection phase, we utilized cough data samples sourced from various crowdsource datasets. The core of our proposed method relies on audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Scaled Spectrogram, Tonal Centroid, Chromagram, and Spectral Contrast, followed by a feature fusion process. We use feature dimension reduction techniques, such as RFECV and Extra-Trees classifier, to identify the most important features. This feature representation and the hyper-parameter space are fed into a Bayesian Optimization function, which optimizes the hyper-parameters of our proposed classifiers. Bayesian Optimization function then calculates the best hyper-parameters before training our proposed classifiers. The dataset clearly exhibits an under-representation of the positive category for COVID-19, which could potentially impact the ML classifier's performance negatively. To address this imbalance, we have incorporated SMOTE during training, aiming to balance the dataset and improve the ML classifier's performance. In classification tasks, using the default threshold (i.e., 0.50) often results in subpar performance, especially when dealing with class imbalance. As a remedy, we employ the threshold-moving technique to adapt the probability threshold that determines the assignment of class labels. The optimal threshold, hyper-parameters, and selected features are subsequently input into classifiers (namely, DNDF and DNDT) for COVID-19 detection

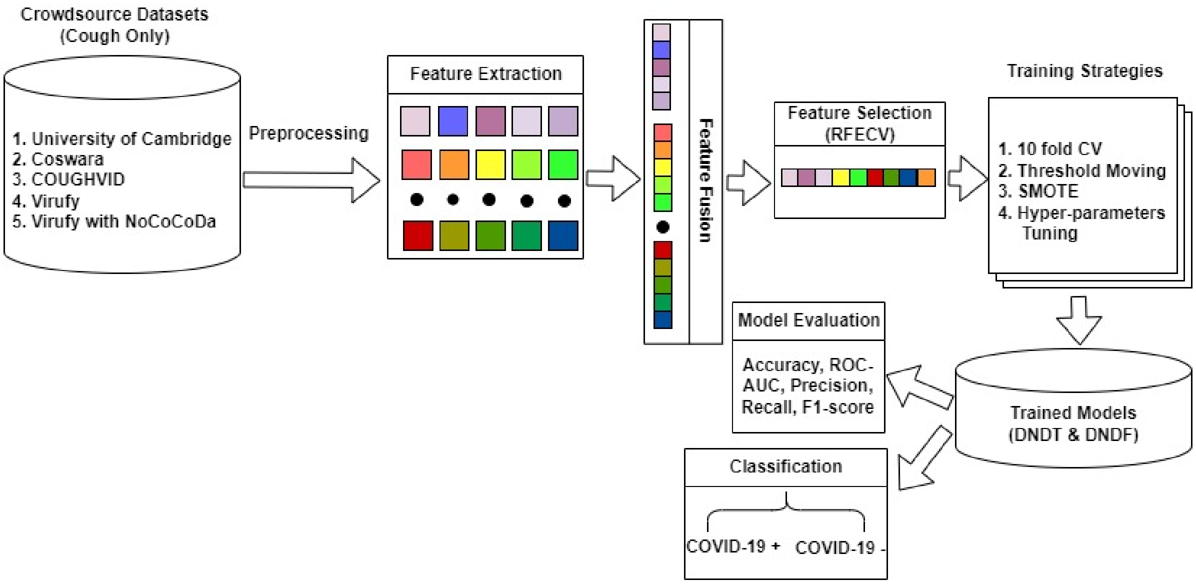


Fig. 1. A synopsis of the proposed method to classify COVID-19 based on the analysis of cough sound audio recordings.

and to assess the proposed method. A comprehensive explanation of each phase in the proposed method is provided in the subsequent sections.

4.1 Dataset description

In this section, we describe the datasets we use to validate and evaluate the effectiveness of our COVID-19 classification models based on cough sound analysis. We utilize five different datasets in our experiment: Cambridge [18], Coswara [95], COUGHVID [77], Virufy [23], and Virufy merged with NoCoCoDa [27]. Cambridge has a dataset of two different categories (asymptomatic and symptomatic). Every cough sample undergoes resampling with a sampling rate of 22.5 kHz, using a Hann window type. We also combine all five datasets to create a combined dataset with a variety of cough samples from both COVID-19 positive and negative individuals, which is shown in Table 2.

Table 2. Cough samples from both COVID-19 positive and negative cases are utilized in separate datasets.

Dataset	Category	COVID-19 Positive	Non COVID-19	Total
Cambridge [18]	Asymptomatic	141	298	439
	Symptomatic	54	32	86
Coswara [95]	-	185	1,134	1,319
COUGHVID [77]	-	680	680	1,360
NoCoCoDa [27]	-	73	-	73
Virufy [23]	-	48	73	121
Virufy + NoCoCoDa [27]	-	121	73	194
Combined	-	1,181	2,217	3,398

4.1.1 Cambridge dataset. The University of Cambridge has developed an online platform and mobile app that allows people to submit recordings of their coughs, inhalations, and voices while reciting a specific phrase. The

Cambridge dataset [18] is divided into two groups: asymptomatic and symptomatic, to distinguish between people who have tested positive for COVID-19 and those who have not. We acknowledge the limitations imposed by the authors of the Cambridge dataset, which is only available under a bilateral legal agreement for research purposes and not for commercial use.

- **Asymptomatic:** To distinguish between people who have tested positive for COVID-19 and those who have tested negative, the Cambridge Asymptomatic dataset contains 141 cough samples from COVID-19 positive people and 298 cough samples from people who have tested negative for COVID-19. The people in the dataset have no notable medical conditions, do not smoke, and are asymptomatic (show no symptoms).
- **Symptomatic:** To distinguish between people who have tested positive for COVID-19 and people who have tested negative, both with a cough, but no other medical conditions or smoking history, the Cambridge Symptomatic dataset contains 54 cough samples from COVID-19 positive people and 32 cough samples from people who have tested negative for COVID-19.

4.1.2 Coswara dataset. The Coswara dataset [95] is a publicly available dataset of cough samples developed by the Coswara project, a collaboration between the Indian Institute of Science and the Indian Institute of Technology Palakkad. It was collected between April 2020 and May 2021, and contains 1,319 cough samples, with 185 samples from COVID-19 positive individuals and 1,134 samples from COVID-19 negative individuals. The samples were preprocessed and labeled based on the healthy (COVID-19 negative) and heavy cough variations (COVID-19 positive).

4.1.3 COUGHVID dataset. The COUGHVID dataset [77] was collected by researchers at the Embedded System Laboratory (ESL) in Switzerland in 2022. We preprocessed and labeled the samples into two groups: those from healthy individuals (COVID-19 negative) and those from individuals with notable cough variations (COVID-19 positive). The COUGHVID dataset contains 1,360 cough samples, with 680 samples from COVID-19 positive individuals and 680 samples from COVID-19 negative individuals.

4.1.4 Virufy dataset. The Virufy dataset [27] is the first publicly available dataset of cough sounds from COVID-19 patients. The sounds were recorded in a hospital with the patient's consent, under the supervision of a physician, and in accordance with standard operating procedures. The Virufy dataset contains 121 cough samples from 16 patients, with 48 samples from COVID-19 positive patients and 73 samples from COVID-19 negative patients.

4.1.5 NoCoCoDa dataset. The NoCoCoDa dataset [27] is a collection of cough sounds from COVID-19 patients recorded during interviews and news programs. It contains 73 cough sounds from 10 participants who attended 13 interviews. To provide a more comprehensive dataset for experiments, we combine the NoCoCoDa dataset with the Virufy dataset, which contains cough samples from both COVID-19 positive and negative people. The combined dataset contains 194 samples, with 121 from COVID-19 positive people and 73 from COVID-19 negative people.

4.1.6 Combined dataset. We consolidate the Cambridge (both asymptomatic and symptomatic), Coswara, COUGHVID, and Virufy merged with NoCoCoDa datasets to form a combined dataset. This unified dataset comprises 3,398 samples, including 1,181 cough samples from COVID-19-positive individuals and 2,217 cough samples from COVID-19-negative individuals.

4.2 Feature extraction

To maintain consistency with a common standard in audio applications, we capture the acoustic signal used for feature extraction at a frequency of 22 kHz. We then compute five spectral features from the sampled audio: MFCCs, Tonal Centroid, Chromagram, Mel-Scaled Spectrogram, and Spectral Contrast. We use the Python library librosa [17] to extract these features.

4.2.1 Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs have demonstrated their effectiveness in distinguishing between dry and wet coughs [22], and are well-regarded as valuable spectral features for audio analysis. We derive 40 MFCCs features from an audio signal. The process of extracting MFCCs encompasses several key steps. Initially, audio signals are divided into frames, with each frame being subject to a windowing function to mitigate noise stemming from sudden changes at its start and end. Subsequently, the Fast Fourier Transform (FFT) is applied to compute the power spectrum of each frame post-windowing. This power spectrum is further manipulated using a filter bank designed based on the Mel scale, as depicted in Equation 1, to obtain Mel-scaled filters from the original frequency (f). Ultimately, the Discrete Cosine Transform (DCT) is employed to derive a set of MFCCs (MFCC coefficients) for each frame from the audio input, following the transformation of the power spectrum into a logarithmic scale.

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{1}$$

4.2.2 Mel-Scaled Spectrogram. The Mel-Scaled Spectrogram is a widely adopted technique in ML for audio analysis, serving as a prevalent method for feature extraction from audio data. This process involves converting the power spectrogram into the Mel scale domain through the utilization of a set of Mel filters. To generate a Mel-scaled Spectrogram, the initial step is to divide the signal into small frames using windowing. For audio processing, a window size of 2048 samples and a hop length of 512 samples are typically set. Subsequently, the power spectrum is computed for each frame using the Fourier Transform. The number of Mel filters, usually set to 128, is evenly spaced in terms of frequencies on the Mel scale. Finally, the power spectrum is passed through these 128 Mel filters, followed by the application of a logarithmic transformation to the resultant values. We obtain 128 Mel-scaled Spectrogram features from an audio signal.

4.2.3 Tonal Centroid. The tonal centroid, a feature employed in audio analysis, is created by projecting a 12-bin chroma vector onto a six-dimensional vector using a transformation matrix, as denoted by Equation 2 [44].

$$\zeta_n(d) = \frac{1}{||c_n||} \sum_{l=0}^{11} \phi(d,l) c_n(l), \quad 0 \le d < 5, \quad 0 \le l \le 11 \tag{2}$$

where, ζ_n is the 6-dimensional tonal centroid vector, computed by multiplying the transformation matrix Φ by the chroma vector c during the specified time frame n and dividing the resulting vector by the L1-norm of the chroma vector to ensure proper scaling of the values.

4.2.4 Chromagram. In the realm of ML applied to audio analysis, chromagrams serve as fundamental input features. We extract 12 chromagram features from an audio signal. The process of generating a chromagram from an acoustic signal involves the utilization of the frequency power spectrum derived from the Short-Time Fourier Transform (STFT). The STFT is computed by employing a sliding window over the audio signal and performing the Fourier transform for each window, effectively representing the audio stream as a time-frequency wave. Subsequently, the power spectrum is derived by squaring the magnitudes of the STFT coefficients.

The chromagram itself is derived from the power spectrum of an acoustic signal through a mapping of the frequency bins. In this context, a specific hop length of 512 and a window size of 2048 are chosen, resulting in the creation of 12 chroma bins. Finally, the feature vector is compiled by obtaining the normalized energy of each chroma bin for every frame in the audio signal.

4.2.5 Spectral Contrast. Spectral contrast features find application in ML for audio analysis. We obtain 7 spectral contrast features from an audio signal. The procedure for deriving spectral contrast features from an audio signal encompasses several sequential stages. Firstly, a Fast Fourier Transform (FFT) is applied to the digital audio clips, capturing the spectral distribution of the audio signal. Subsequently, the frequency spectrum is partitioned into a collection of sub-bands using octave band filters. The number of these sub-frequency bands is standardized at 6. The evaluation of spectral valleys, peaks, and their disparities is performed within each sub-band, as described in Equations 3, 4 and 5 [54]. The initial spectral contrast values are then transformed into a logarithmic representation. Lastly, through the utilization of a Karhunen-Loeve transform, the Log-frequency contrast values are projected into an orthogonal subspace.

$$Peak_k = log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i}\right\} \tag{3}$$

$$Valley_k = log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,N-i+1}\right\} \tag{4}$$

$$SC_k = Peak_k - Valley_k \tag{5}$$

Here, N represents the overall count within the k-th sub-frequency band, k ranging from 1 to 6, and α is invariant with a range of 0.02 to 0.2.

4.3 Feature selection

We extract 193 features from each audio signal, including 40 MFCCs, 12 chromagram, 128 mel-scaled spectrogram, 7 spectral contrast, and 6 tonal centroid features. However, it's worth noting that not all of these features are optimal. One method that's often used in ML is feature selection. The combination of RFECV technique and Extra Trees Classifier is one efficient method for feature dimension reduction (optimal feature selection). The most appropriate features to be chosen automatically are found using RFECV, which uses cross-validation and feature significance weights. Less significant characteristics are removed iteratively, and cross-validation is used to assess the model's performance. Finds the set of features that are most important for classification by using this method. Following the utilization of the RFECV technique and Extra Trees Classifier, we obtain optimal features of 71, 182, 33, 172, 46, and 188 for Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa dataset, respectively. To obtain information about the significance of each feature, we employ the Extra-Trees estimator. The method's main objective is to use the RFECV method and the Extra-Trees classifier to examine feature importance and minimize feature dimension.

4.4 Hyper-parameters tuning

Hyper-parameters are parameters that control the learning process of a ML model, such as the number of trees, depth, used features rate, and epochs. Hyperparameter tuning is the process of finding the best values for these hyper-parameters to optimize the model's performance on a given dataset. Bayesian Optimization (BO) is a popular and effective technique for hyperparameter tuning. BO works by building a probabilistic model of the relationship between the hyper-parameters and the model's performance. It then uses this model to select the next set of hyper-parameters to try, with the goal of finding the values that lead to the best performance. BO uses a model to make informed decisions about which hyper-parameter values to test next, leveraging past results to make more efficient choices. This approach tends to require fewer iterations to find the optimal hyper-parameter combination compared to the more brute-force methods of Grid Search (GS) and Random Search (RS) [34]. This efficiency is a key advantage of BO in hyperparameter tuning. The performance of DNDT and DNDF for classification is significantly impacted by hyper-parameters. In BO, we define hyper-parameter space for hyper-parameters of our classifiers. Extracted input features and their labels give to Bayesian Optimization function for getting the most effective hyperparameter values. The default hyper-parameters employed for all

Table 3. The default hyper-parameters are used for all datasets.

Num_trees	Depth	Features rate	Learning rate	Batch size	Num_epochs
10	10	1	0.01	256	10

Table 4. The optimized hyper-parameters are used for different datasets.

Dataset	Category	Hyper-parameters					
		Num_trees	Depth	Features rate	Learning rate	Batch size	Num_epochs
Cambridge	Asymptomatic	18	8	0.8	0.01	32	17
	Symptomatic	25	9	0.8	0.01	8	13
Coswara	-	25	11	0.6	0.01	16	14
COUGHVID	-	17	6	1.0	0.01	32	19
Virufy	-	16	16	0.8	0.01	8	22
Virufy + NoCoCoDa	-	29	5	0.6	0.01	32	34
Combined	-	42	13	0.7	0.01	256	40

datasets in both DNDT and DNDF classifiers are presented in Table 3. Additionally, Table 4 displays the optimized hyper-parameters for various datasets in both DNDT and DNDF classifiers.

4.5 Trained classifiers

4.5.1 Deep Neural Decision Tree (DNDT). The Deep Neural Decision Tree is a classifier structured as a tree, encompassing both decision and prediction nodes. Decision nodes, positioned within the tree but not at its leaves, serve as points where the tree assesses data features or conditions to make determinations. The decisions at each node guide the path a sample takes through the tree. On the other hand, prediction nodes are the leaf nodes of the tree, where the final prediction is generated. These nodes serve as the terminal points for predictions. Each prediction node corresponds to a specific class or outcome the classifier aims to predict. To classify a sample, the Deep Neural Decision Tree guides it to a leaf node, employing a probability distribution to make the final prediction. The final prediction for a sample is determined by Equation 6 [58],

$$\mathbb{P}_T[y|x,\theta,\pi] = \sum_{l\in L} \pi_{l_y}\mu_l(x|\theta) \tag{6}$$

where, π_{l_y} is the likelihood of a sample arriving leaf node l to get placed in class y and routing function $\mu_l(x|\theta)$, is the likelihood of a sample x will arrive leaf node l.

4.5.2 Deep Neural Decision Forest (DNDF). The Deep Neural Decision Forest is a classifier consisting of multiple Deep Neural Decision Trees trained simultaneously. The Neural Decision Forest produces its final output by averaging the individual outputs from each of the trees within the forest. The output of the Neural Decision Forest represents by Equation 7 [58].

$$\mathbb{P}_F[y|x] = \frac{1}{k}\sum_{h=1}^{k} \mathbb{P}_{Th}[y|x] \tag{7}$$

Here, k represents how many trees are in the forest, $\mathbb{P}_F[y|x]$ output for sample x produced by Deep Neural Decision Forest and $\mathbb{P}_{Th}[y|x]$ output for sample x produced by Deep Neural Decision Tree.

Table 5. Combinations of several training strategies.

Strategy	Feature Selection	Hyper-parameters	Up-sampling	Threshold
#	(RFECV)	Selection Method	(SMOTE)	Moving
Strategy 1	✗	Default	✗	✗
Strategy 2	✗	Default	✗	✓
Strategy 3	✓	Default	✗	✓
Strategy 4	✓	Bayesian Optimization	✗	✓
Strategy 5	✓	Bayesian Optimization	✓	✓

4.6 Threshold moving

In binary classification, the threshold used to determine class labels based on predicted probabilities is crucial for the model's performance. To improve class label assignment, we employ the threshold moving technique, as relying on the default threshold of 0.50 often leads to suboptimal results. With this method, the ROC-AUC score is taken into account while determining the optimal threshold for a binary classifier. In the field of medicine, the ROC-AUC score is a commonly used assessment metric, particularly for evaluating the effectiveness of diagnostic procedures [73]. ROC-AUC score-based threshold optimization is achieved by cross-validation tests. We calculate ROC-AUC scores over a threshold value range of 0.1 to 1, using 0.001 increments. The best threshold is then determined by selecting the one that produced the highest ROC-AUC score.

4.7 Training strategies

We introduce five strategies to evaluate the effectiveness of different components of our proposed method: strategy 1, strategy 2, strategy 3, strategy 4, and strategy 5. Table 5 shows the combinations of different training strategies used in each strategy. Strategy 1 exclusively relies on our trained classifiers. In strategy 2, we only use the threshold moving technique (to select the optimal threshold based on ROC-AUC score) with a trained classifier. In strategy 3, we use both the threshold moving technique and the feature dimension reduction technique (to determine the key features using the RFECV method and Extra-Trees classifier) with a trained classifier. In strategy 4, we use the threshold moving technique, optimal feature selection using the RFECV method and the Extra-Trees classifier, and Bayesian Optimization (to select the best hyper-parameters). In strategy 5, we employ the threshold moving technique, optimize feature selection using the RFECV method, utilize Bayesian Optimization, and apply SMOTE to balance the data of the minority class in an imbalanced dataset.

Strategies 2 through 5, excluding strategy 1, use the threshold moving technique to determine the optimal threshold based on ROC-AUC score. We also use 10-fold stratified cross-validation to assess the performance of our trained classifiers in all strategies. However, benchmark datasets conspicuously indicate a limited representation of the positive COVID-19 category, potentially causing an adverse effect on the ML classifier's performance. To enhance the classifier's performance, we have integrated SMOTE during training in strategy 5 to balance the dataset. The difference among strategy 1, strategy 2, strategy 3, strategy 4, and strategy 5 is that strategy 1 and strategy 2 do not use feature selection method (RFECV) during the training phase, while strategy 3, strategy 4 and strategy 5 do. In addition, strategy 1, strategy 2 and strategy 3 use default hyper-parameters during the training phase, while strategy 4 and strategy 5 use Bayesian Optimization to select the best hyper-parameters. Moreover, in strategy 5, SMOTE is the sole technique used for up-sampling, while the other strategies do not incorporate any up-sampling methods.

5 RESULTS AND DISCUSSION

In this section, we present the results of our experiments on identifying COVID-19 by analyzing cough sounds. We first outline the evaluation metrics used to assess the performance of our proposed methods. Next, we present the classification performance of four strategies, which helps us to select the best strategy based on its performance. Then, we describe the optimal feature selection techniques used to select the most important features. Finally, we evaluate the effectiveness of our proposed method using state-of-the-art methods on different datasets.

5.1 Evaluation Metrics

In our experimental evaluation, we employ 10-fold stratified cross-validation to evaluate the performance using six standard evaluation metrics: ROC-AUC, Accuracy (Acc.), Precision, Recall/Sensitivity, Specificity (Spec.), and F1 score. The definitions of Precision, Recall/Sensitivity, Accuracy, Specificity, and F1 score are provided below:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Here, True Positive, False Positive, True Negative, and False Negative are represented as TP, FP, TN and FN, respectively. Receiver Operating Characteristic (ROC) is determined by the TPR (true positive rate) as a fraction of the FPR (false positive rate). Area Under the Curve (AUC) signifies the area under the ROC curve.

5.2 Classification performance of five strategies

All five strategies employ the use of DNDT and DNDF classifiers, and strategy 2 through strategy 5 incorporate the threshold moving technique. Strategy 3, strategy 4, and strategy 5 implement RFECV method, employing Extra-Trees classifier to identify the most relevant features. Strategy 4 and strategy 5 further integrate Bayesian Optimization to fine-tune hyperparameters for the trained classifiers. Additionally, in strategy 5, SMOTE is applied during the training phase to address data imbalance by generating synthetic samples for both positive and negative instances of cough related to COVID-19. The AUC-based classification performance of all five strategies is outlined in Table 6.

Across all datasets, Strategy 2 consistently outperforms Strategy 1 in terms of AUC, underscoring the effectiveness of the threshold moving technique. Specifically, when using the DNDT classifier, Strategy 2 achieves superior AUC scores of 0.72, 0.94, 0.62, 0.72, 0.94, and 0.93 for the Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa datasets, respectively. Similarly, with the DNDF classifier, Strategy 2 surpasses Strategy 1 with higher AUC scores of 0.72, 0.93, 0.65, 0.70, 0.93, and 0.93 for the same datasets.

From a variety of datasets, Strategy 3 demonstrates superior overall performance compared to both Strategy 1 and Strategy 2. In particular, in the Cambridge symptomatic dataset, Strategy 3, utilizing the DNDT classifier, outshines Strategy 1 and Strategy 2, achieving a higher AUC score of 0.95. Similarly, in the Coswara dataset, Strategy 3, employing the DNDT classifier, outperforms the other two strategies with a higher AUC of 0.63.

In the comparative analysis of Strategies 1, 2, 3, and 4, it is apparent that Strategy 4 consistently excels over the other three across diverse datasets. Particularly in the Cambridge asymptomatic dataset, Strategy 4, incorporating both DNDT and DNDF classifiers, outperforms Strategies 1, 2, and 3 with higher AUC scores of 0.79 and 0.83, respectively. Furthermore, when employing the DNDT classifier, Strategy 4 surpasses the other three on the Virufy dataset, achieving a superior AUC value of 0.95. Finally, in the case of Virufy combined with the NoCoCoDa dataset, the DNDF classifier within Strategy 4 outperforms the other strategies, achieving a superior AUC score of 0.98.

When evaluating Strategies 1, 2, 3, 4, and 5, it is apparent that Strategy 5 consistently outperforms the other four across nearly all datasets, demonstrating its effectiveness in COVID-19 classification. In the Cambridge asymptomatic dataset, Strategy 5, which utilizes both DNDT and DNDF classifiers, outperforms Strategies 1, 2, 3, and 4, achieving superior AUC scores of 0.89 for both classifiers. Furthermore, Strategy 5's DNDF classifier outperforms the others with an AUC score of 0.95 in the Cambridge symptomatic dataset. In the Coswara dataset, Strategy 5, employing both DNDT and DNDF classifiers, excels, achieving higher AUC scores of 0.68 and 0.73, respectively. Additionally, Strategy 5 with the DNDT classifier exhibits superior performance on the Virufy dataset compared to the other four, boasting an AUC score of 0.97. When Virufy is combined with the NoCoCoDa dataset, Strategy 5's DNDT and DNDF classifiers outshine the other four strategies in terms of AUC, with values of 0.98 and 0.99. Using strategy 5, which outperforms the other four strategies, our trained classifiers are utilized to categorize COVID-19 from cough sounds. The comprehensive classification performance of Strategy 5 across all evaluation metrics and datasets is detailed in Table 7.

Figure 2 displays the confusion matrices for approach 5, which use DNDT classifier to determine how well it detects COVID-19 across various datasets. In that sequence, Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa are the datasets represented by the matrices labeled (2a) – (2f). Furthermore, confusion matrices for strategy 5 that employ DNDF classifier are displayed in Figure 3, providing details on how well it performs for COVID-19 detection on the same range of datasets. The matrices labeled (3a) – (3f) again correspond to the previously described datasets: Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa. DNDT and DNDF classifiers in strategy 5 accurately predict 126 out of 141 COVID-19 positive cough samples and 262 out of 298 cough samples from healthy individuals in Cambridge asymptomatic dataset. DNDT classifier using method 5 in Cambridge symptomatic dataset correctly identifies 31 out of 32 cough samples from healthy people and 50 out of 54 COVID-19 positive cough samples. Additionally, 88.89% of COVID-19 positive cough samples and 100% of cough samples from healthy people are successfully predicted by DNDF classifier using strategy 5. Coswara dataset shows that DNDT classifier with strategy 5 makes accurate predictions, properly categorizing 647 out of 1,134 cough samples from healthy individuals and 145 out of 185 COVID-19 positive cough samples.

Moreover, 85.19% of cough samples from healthy people and 61.08% of COVID-19 positive cough samples are accurately predicted by DNDF classifier using strategy 5. Using strategy 5, DNDT classifier in COUGHVID dataset produces accurate predictions, properly recognizing 434 out of 680 cough samples that are positive for COVID-19 and 415 out of 680 cough samples that are from healthy persons. By accurately predicting 500 out of 680 COVID-19 positive cough samples and 441 out of 680 cough samples from healthy people, DNDF classifier with strategy 5 also shows effectiveness. DNDT classifier using strategy 5 performs well in Virufy dataset, correctly predicting 100% of cough samples from COVID-19 positive patients and 94.52% of cough samples from healthy persons. Furthermore, DNDF classifier using strategy 5 makes accurate predictions, correctly detecting 100% of cough samples from healthy individuals and 85.42% of COVID-19 positive cough samples. Both DNDT and DNDF classifiers in strategy 5's integration of Virufy dataset with NoCoCoDa demonstrate better accuracy, correctly predicting 95.04% and 97.52% of COVID-19 positive cough samples, respectively. Additionally, 100% of the cough samples from healthy people are accurately identified by both classifiers in Strategy 5.

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

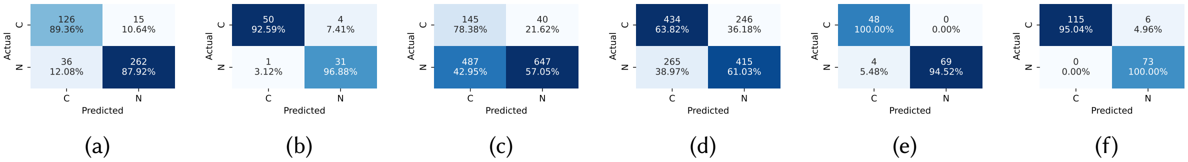
Table 8 provides a comparative analysis of our innovative approaches for COVID-19 diagnosis from cough samples in contrast to contemporary methods. Our methodologies encompass several pivotal components, including feature selection via Recursive Feature Elimination with Cross-Validation (RFECV), hyperparameter tuning

Table 6. Comparison among all strategies.

Dataset	Category	Classification method	AUC				
			Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Cambridge	Asymptomatic	DNDT	0.55	0.72	0.72	0.79	0.89
		DNDF	0.51	0.72	0.70	0.83	0.89
	Symptomatic	DNDT	0.77	0.94	0.95	0.91	0.95
		DNDF	0.69	0.93	0.93	0.93	0.95
Coswara	-	DNDT	0.51	0.62	0.63	0.59	0.68
		DNDF	0.50	0.65	0.62	0.60	0.73
COUGHVID	-	DNDT	0.65	0.72	0.72	0.59	0.62
		DNDF	0.63	0.70	0.70	0.70	0.69
Virufy	-	DNDT	0.85	0.94	0.89	0.95	0.97
		DNDF	0.82	0.93	0.88	0.87	0.93
Virufy + NoCoCoDa	-	DNDT	0.86	0.93	0.93	0.91	0.98
		DNDF	0.79	0.93	0.93	0.98	0.99

Table 7. Classification performance of strategy 5 (RFECV+BO+SMOTE+TM).

Dataset	Category	Method	Acc.	AUC	Precision	Recall	F1 Score	Spec.
Cambridge	Asymptomatic	DNDT	0.88	0.89	0.81	0.89	0.84	0.88
		DNDF	0.88	0.89	0.80	0.90	0.84	0.88
	Symptomatic	DNDT	0.94	0.95	0.98	0.93	0.95	0.97
		DNDF	0.93	0.95	1	0.90	0.93	1
Coswara	-	DNDT	0.60	0.68	0.27	0.78	0.38	0.57
		DNDF	0.82	0.73	0.41	0.62	0.48	0.85
COUGHVID	-	DNDT	0.62	0.62	0.65	0.64	0.62	0.61
		DNDF	0.69	0.69	0.70	0.74	0.69	0.65
Virufy	-	DNDT	0.97	0.97	0.94	1	0.97	0.95
		DNDF	0.94	0.93	1	0.86	0.92	1
Virufy + NoCoCoDa	-	DNDT	0.97	0.98	1	0.95	0.97	1
		DNDF	0.99	0.99	1	0.98	0.99	1



(a) (b) (c) (d) (e) (f)

Fig. 2. The confusion matrices for strategy 5, employing the Deep Neural Decision Tree (DNDT) classifier with 10-fold cross-validation, are presented in (2a) – (2f), representing the matrices across the Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa datasets, respectively. "C" represents COVID-19 Positive, and "N" represents Non COVID-19 cough instances.

through Bayesian Optimization (BO), data augmentation using Synthetic Minority Over-sampling Technique (SMOTE) during training, and dynamic threshold adjustment guided by ROC-AUC score (TM). It is worth noting that the comparative assessment is conducted with a deliberate focus on three specific prior studies, namely, Brown et al. [18], Soltanian and Borna [101] and Chowdhury et al. [26]. The rationale behind this limitation is

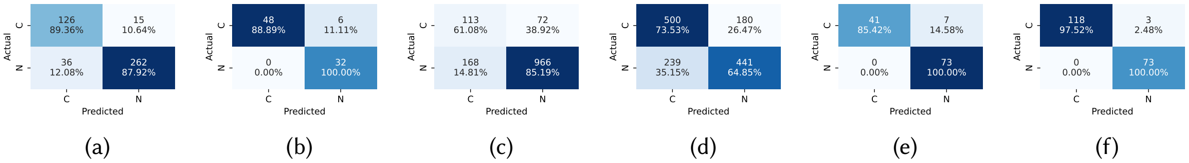


Fig. 3. The confusion matrices for strategy 5, employing the Deep Neural Decision Forest (DNDF) classifier with 10-fold cross-validation, are presented in (3a) – (3f), representing the matrices across the Cambridge asymptomatic, Cambridge symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa datasets, respectively. "C" represents COVID-19 Positive, and "N" represents Non COVID-19 cough instances.

rooted in the availability of detailed implementation specifics and datasets for these select works, which ensures a more accurate and comprehensive comparative analysis.

Our proposed method demonstrates constant superiority over state-of-the-art techniques on a wide range of datasets, confirming its strong performance in the field of COVID-19 diagnosis. DNDT and DNDF classifiers perform well when we focus on the Cambridge Asymptomatic dataset, outperforming other classifiers in evaluation metrics. Together, these classifiers accomplish the amazing accomplishment of earning the highest AUC score, demonstrating their outstanding discriminative power with an astounding AUC of 0.89. Furthermore, DNDT classifier stands out for achieving the highest precision among all methods, attaining a precision score of 0.81. Conversely, the DNDF classifier distinguishes itself with the highest recall, setting a remarkable precedent with a recall score of 0.90. These results on the Cambridge Asymptomatic dataset underscore the comprehensiveness and balance of our proposed method's performance. In the Cambridge Symptomatic dataset, Aytekin et al. [12] secures the highest AUC score at 0.98. Additionally, Dentamaro et al. [30], Aytekin et al. [12], and the DNDT classifiers all scoring an impressive recall score of 0.93. Notable distinctions include the fact that Chowdhury et al. [26] and the DNDF classifiers both attain the maximum precision score of 1, highlighting their accuracy in identifying COVID-19 instance.

DNDF classifier performs on the Coswara dataset, achieving an impressive AUC value of 0.73. In the meantime, DNDT classifier performs well, outperforming other methods in recall with a noteworthy score of 0.78. Notably, with a score of 0.76, Chowdhury et al. [26] obtains the maximum precision on this dataset. The DNDF classifier surpasses alternative methods on the COUGHVID dataset, demonstrating precision and recall values of 0.70 and 0.74, respectively. It is noteworthy that Pavel and Ciocoiu [85] attains the highest AUC on this dataset with a score of 0.76. DNDT classifier performs better on the Virufy dataset in terms of both AUC and recall, with outstanding outcomes of 0.97 and 1, respectively. In contrast, Soltanian and Borna [101], Islam et al. [53], and the DNDF classifiers achieve a perfect precision score of 1. Both Chowdhury et al. [26] and DNDF classifiers are featured in the Virufy dataset combined with NoCoCoDa; they achieve the highest recall with an impressive score of 0.98. Moreover, both the DNDT and DNDF classifiers attain a perfect precision score of 1. In a different dataset examined in Melek [67], and DNDF classifiers both achieve perfect scores of 0.99 for AUC performance. When compared to DNDT classifier, DNDF classifier performs better in terms of AUC and precision across the Coswara and COUGHVID datasets. On the Virufy dataset, however, DNDT classifier outperforms the DNDF classifier, especially in terms of AUC and recall. It's interesting to note that DNDF classifier regains its dominance in the setting of the Virufy dataset merged with NoCoCoDa, surpassing DNDT classifier in terms of both AUC and recall.

These commendable AUC, precision, and recall values collectively underscore the compelling effectiveness of our proposed method in the challenging task of classifying COVID-19 cases based on cough sound data. These

Table 8. A comparison between our proposed method and state-of-the-art methods.

Dataset	Category	Method	AUC	Precision	Recall
Cambridge	Asymptomatic	Brown et al. [18]	0.80	0.72	0.69
		Dentamaro et al. [30]	0.83	0.80	0.80
		Chowdhury et al. [26]	0.88	0.75	0.81
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	**0.89**	**0.81**	0.89
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	**0.89**	0.80	**0.90**
	Symptomatic	Brown et al. [18]	0.87	0.70	0.90
		Chowdhury et al. [25]	-	0.87	0.82
		Dentamaro et al. [30]	0.93	0.89	**0.93**
		Chowdhury et al. [26]	0.95	**1**	0.91
		Aytekin et al. [12]	**0.98**	0.94	**0.93**
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	0.95	0.98	**0.93**
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	0.95	**1**	0.90
Coswara	-	Chowdhury et al. [26]	0.66	**0.76**	0.47
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	0.68	0.27	**0.78**
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	**0.73**	0.41	0.62
COUGHVID	-	Pavel and Ciocoiu [85]	**0.76**	0.69	0.68
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	0.62	0.65	0.64
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	0.69	**0.70**	**0.74**
Virufy	-	Soltanian and Borna [101]	-	**1**	0.95
		Islam et al. [53]	-	**1**	0.95
		Chowdhury et al. [26]	0.94	0.89	0.98
		Sobahi et al. [99]	-	0.99	0.97
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	**0.97**	0.94	**1**
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	0.93	**1**	0.86
Virufy + NoCoCoDa	-	Melek [67]	**0.99**	0.99	0.97
		Chowdhury et al. [26]	0.98	0.99	**0.98**
		Proposed (DNDT+RFECV+BO+SMOTE+TM)	0.98	**1**	0.95
		Proposed (DNDF+RFECV+BO+SMOTE+TM)	**0.99**	**1**	**0.98**

- **Bold** values represent the maximum values.

results not only validate the robustness of our approach but also underscore its potential to make a significant contribution to advancing COVID-19 diagnostic capabilities.

5.4 Cross-dataset evaluation

In the course of our research, we conduct a comprehensive cross-dataset study aimed at assessing the effectiveness of our proposed method in diagnosing COVID-19 based on cough sounds. Our study encompass a diverse range of COVID-19 cough datasets, which included the Cambridge (Asymptomatic), Cambridge (Symptomatic), Coswara, COUGHVID, Virufy, and Virufy with NoCoCoDa datasets. This comprehensive approach allow us to evaluate the robustness and generalizability of our method. To ensure the integrity of our study, we follow a systematic approach. We train our proposed method on one dataset, allowing it to learn from the unique characteristics of that dataset. Subsequently, we rigorously validate the model's performance on the remaining datasets, one dataset at a time. This validation process, performs independently for each dataset, is a crucial step in assessing the method's adaptability and its ability to distinguish COVID-19 cough sounds under varying conditions. RFECV is intentionally taken out of our methodology, which is one noteworthy component. This choice is taken due to the fact that different datasets have drastically different ideal feature properties that are necessary for a reliable COVID-19 diagnosis. To improve the overall performance of the technique, we use BO to fine-tune hyperparameters. In addition, we increased the data and addressed class imbalance, a frequent issue in medical

Table 9. Classification performance of our proposed method in cross-dataset study.

Trained Dataset	Testing Dataset	AUC	Precision	Recall	F1 Score
Cambridge (Asymptomatic)	Cambridge (Symptomatic)	0.56	0.76	0.24	0.37
	Coswara	0.52	0.18	0.14	0.15
	COUGHVID	0.53	0.60	0.15	0.25
	Virufy	0.63	0.48	0.90	0.62
	Virufy + NoCoCoDa	0.56	0.78	0.21	0.33
Cambridge (Symptomatic)	Cambridge (Asymptomatic)	0.62	0.42	0.68	0.52
	Coswara	0.55	0.17	0.56	0.26
	COUGHVID	0.51	0.51	0.67	0.58
	Virufy	0.57	0.57	0.27	0.37
	Virufy + NoCoCoDa	0.65	0.79	0.51	0.62
Coswara	Cambridge (Asymptomatic)	0.59	0.39	0.65	0.49
	Cambridge (Symptomatic)	0.57	0.67	0.93	0.78
	COUGHVID	0.53	0.52	0.88	0.65
	Virufy	0.51	0.45	0.10	0.17
	Virufy + NoCoCoDa	0.62	0.71	0.77	0.74
COUGHVID	Cambridge (Asymptomatic)	0.50	1	0.01	0.01
	Cambridge (Symptomatic)	0.55	0.77	0.19	0.30
	Coswara	0.51	0.15	0.23	0.18
	Virufy	0.69	0.60	0.69	0.64
	Virufy + NoCoCoDa	0.66	0.79	0.56	0.66
Virufy	Cambridge (Asymptomatic)	0.59	0.38	0.82	0.51
	Cambridge (Symptomatic)	0.62	0.70	0.80	0.75
	Coswara	0.53	0.15	0.77	0.25
	COUGHVID	0.53	0.53	0.47	0.50
	Virufy + NoCoCoDa	0.83	0.88	0.86	0.87
Virufy + NoCoCoDa	Cambridge (Asymptomatic)	0.64	0.44	0.72	0.55
	Cambridge (Symptomatic)	0.62	0.76	0.48	0.59
	Coswara	0.52	0.15	0.90	0.25
	COUGHVID	0.51	0.56	0.11	0.19
	Virufy	0.76	0.70	0.73	0.71

datasets, by utilizing SMOTE during the training phase. Deciding on the best classification threshold is made possible by TM technique. We examine DNDT classifier in this context.

The cross-dataset study's (DNDT+BO+SMOTE+TM) classification performance is displayed in Table 9. The combination of the NoCoCoDa and Virufy datasets performs better than other testing datasets when Virufy dataset is used for training, obtaining an F1 score of 0.87 and an AUC of 0.83, respectively. When COUGHVID is used as the training dataset, the Cambridge (Asymptomatic) dataset outperforms the other testing datasets in terms of precision, scoring a perfect 1. Furthermore, with Cambridge (Symptomatic) dataset using Coswara as the training dataset, it performs exceptionally well in recall, outperforming other testing datasets with a score of 0.93. Within the training dataset of Cambridge (Asymptomatic), remarkable performance metrics come to the forefront when assessing the performance of our method across various testing datasets. The testing datasets from Virufy + NoCoCoDa and Cambridge (Symptomatic) stand out in particular for their noteworthy accomplishments in terms of AUC and precision. In particular, the testing dataset for Virufy exhibits outstanding performance in several dimensions, such as AUC, recall, and F1 score. The testing datasets for Virufy + NoCoCoDa and Cambridge (Symptomatic) both yield an equal AUC score of 0.56, supported by precision values of 0.78 and 0.76, respectively. Furthermore, Virufy testing dataset continues to provide excellent results, with an AUC of 0.63. Recall and F1 values of 0.90 and 0.62, respectively, support this outstanding performance and highlight the method's competence in differentiating COVID-19 cough sounds when trained on the Cambridge (Asymptomatic) dataset.

The evaluation of Cambridge (Symptomatic) training dataset across diverse testing datasets, including COUGHVID and Virufy + NoCoCoDa, reveals robust and noteworthy performance across multiple key metrics. Specifically, these datasets achieve impressive scores in AUC, precision, recall, and F1 score, highlighting their proficiency in distinguishing COVID-19 cough sounds. The testing datasets for COUGHVID and Virufy + NoCoCoDa show good performance for the Cambridge (Symptomatic) training dataset. They obtain, respectively, precision values of 0.51 and 0.79, recall scores of 0.67 and 0.51, AUC scores of 0.51 and 0.65, and F1 scores of 0.58 and 0.62. The Cambridge (Asymptomatic) testing dataset also demonstrates commendable performance, particularly excelling in AUC with a score of 0.62, and achieving a notable recall score of 0.68. Furthermore, Virufy testing dataset emerges as a formidable competitor, particularly demonstrating excellence in AUC with a score of 0.57 and exhibiting commendable precision with a score of 0.57. These results collectively underscore the effectiveness and versatility of the method when applied to different datasets, further emphasizing its potential in COVID-19 cough sound classification.

The Coswara training dataset yields impressive outcomes when subjected to evaluation by the Cambridge (Symptomatic), COUGHVID, and Virufy + NoCoCoDa testing datasets. These testing datasets showcase robust performance across essential metrics, including AUC, precision, recall, and F1 scores. Consequently, they attain noteworthy performance indicators, including values of 0.57, 0.67, 0.93, and 0.78; 0.53, 0.52, 0.88, and 0.65; and 0.62, 0.71, 0.77, and 0.74, respectively. Furthermore, the Cambridge (Asymptomatic) testing dataset reinforces the method's effectiveness, particularly in terms of AUC (0.59) and recall (0.65). Furthermore, the assessment of the COUGHVID training dataset demonstrates significant performance in various metrics, encompassing AUC, precision, recall, and F1 score. Notably, the Virufy and Virufy + NoCoCoDa testing datasets stand out as robust achievers, securing scores of 0.69, 0.60, 0.69, 0.64, and 0.66, 0.79, 0.56, and 0.66, respectively.

In the Virufy training dataset assessment, the Cambridge (Symptomatic) and Virufy + NoCoCoDa testing datasets perform well, securing AUC, precision, recall, and F1 scores of 0.62, 0.70, 0.80, and 0.75, and 0.83, 0.88, 0.86, and 0.87, respectively. Moreover, the Cambridge (Asymptomatic) testing dataset shows notable performance, particularly in AUC (0.59) and recall (0.82). The Virufy testing dataset performs well in the context of the Virufy integrated with NoCoCoDa training dataset, with scores of 0.76, 0.70, 0.73, and 0.71 for AUC, precision, recall, and F1 score, respectively. Furthermore, the Cambridge (Asymptomatic) testing dataset shows commendable results, particularly in terms of AUC (0.64) and recall (0.72). Additionally, the Cambridge (Symptomatic) testing dataset performs well, especially in terms of accuracy (0.76) and AUC (0.62).

Using Kernel Density Estimation (KDE) curves, Figure 4 illustrates the distribution of MFCC features across COVID-19 positive and negative samples across several datasets. The KDE curve's y-axis shows the estimated probability density at particular places along the x-axis, while the x-axis indicates the data points that correspond to MFCC features. The peak probability density in the distribution is represented by the highest point on the y-axis. In particular, peak probability densities of roughly 0.088, 0.023, 0.019, 0.014, 0.006, and 0.004 are revealed by the KDE curve for MFCC features in COVID-19 positive cough samples for the COUGHVID, Coswara, Cambridge (Asymptomatic), Virufy merged with NoCoCoDa, Cambridge (Symptomatic), and Virufy datasets, respectively. Likewise, peak probability densities are also noted for non-COVID-19 cough samples, and they are around 0.146, 0.094, 0.035, 0.014, 0.010, and 0.004 for the Coswara, COUGHVID, Cambridge (Asymptomatic), Virufy combined with NoCoCoDa, Virufy, and Cambridge (Symptomatic) datasets, respectively. The properties of the Cambridge (Asymptomatic), Cambridge (Symptomatic), Virufy, and Virufy + NoCoCoDa datasets are found to be identical after examining the highest probable densities and shapes in Figure 4. In addition, it is noted that the training and testing datasets have similar features when assessing classification accuracy and looking at feature distribution.

In Section 5.3, Table 8 emphasizes the superior performance of the DNDT and DNDF classifiers in distinguishing between COVID-19-positive and healthy individuals within individual datasets. However, as shown in Table

9 in this section, the classification performance of the DNDT classifier falls short in most cases, except in the Virufy and Virufy merged with NoCoCoDa datasets combination during the cross-dataset study. The improved results in this combination can be attributed to the inclusion of Virufy datasets in the subset merged with the NoCoCoDa dataset. While the DNDT classifier excels in individual datasets, its performance is less satisfactory in cross-dataset studies with differing training and testing datasets. Additionally, Akman et al. [3] notes the superior performance of their classifier on trained datasets but acknowledges challenges in classifying datasets differing from the training set. This aligns with the concerns highlighted by Coppock et al. [28] regarding pervasive bias in current COVID-19 audio datasets, allowing machine learning models to infer COVID-19 status not just from unique audio biomarkers but also from other dataset correlations, including comorbidity, geographical, ethnic factors, and background noise.

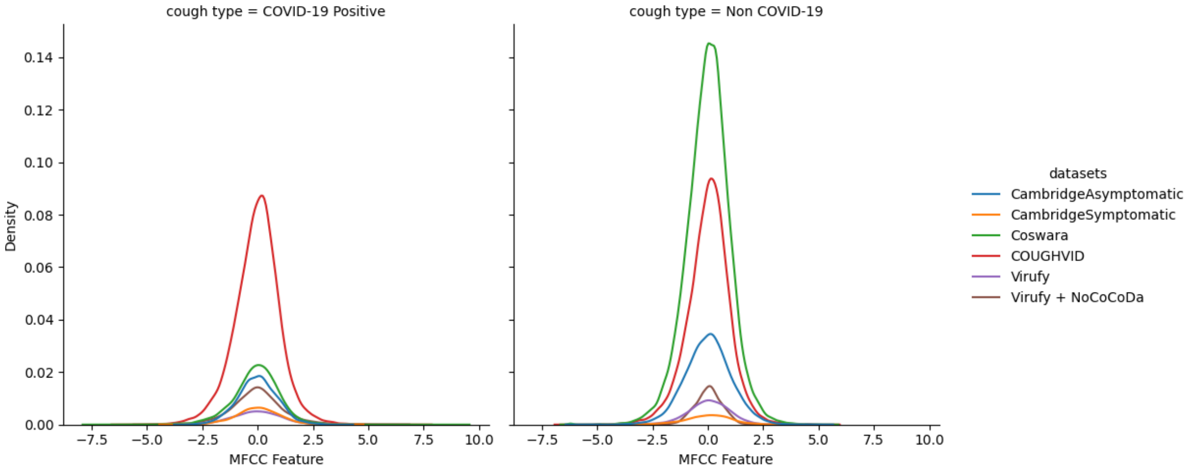


Fig. 4. The distribution of MFCC features across COVID-19 positive and negative samples across several datasets.

5.5 Classification performance of proposed method for the combined dataset

Given the intricate nature of feature characteristics, it was a strategic choice to forgo applying the RFECV technique for the Combined dataset. It is an acknowledgment that there can be significant variability in data, even within the same area. In one dataset, the best features for precise COVID-19 diagnosis could not be the same in another. This focus on flexibility and data-specific optimization highlights a dedication to accurate categorization in the context of heterogeneous datasets. In the meanwhile, the cornerstone of our model refining approach has been BO. By providing a data-driven and intelligent method for hyperparameter tuning, BO allows the model to dynamically adjust its configurations. This improves predicted accuracy and establishes the model as a reliable and adaptable tool for COVID-19 diagnosis. It is evidence of the incorporation of cutting-edge methods to guarantee optimal performance.

Furthermore, during the training phase, we have integrated SMOTE. By fostering a more representative training dataset, this data augmentation technique lessens the effects of class imbalance and improves the model's capacity to identify patterns in both major and minor classes. Additionally, TM technique has been utilized, which modifies the decision threshold in real-time according to the ROC-AUC scores. With the help of this dynamic adaptation, the model's sensitivity and specificity are precisely balanced, resulting in predictions that are precise and well-informed.

Table 10. A comparison between our proposed methods for the combined dataset.

Dataset	Method	Acc.	AUC	Precision	Recall	F1-Score	Specificity
Combined	DNDT+BO+SMOTE+TM	**0.76**	**0.75**	**0.65**	**0.69**	**0.66**	**0.80**
	DNDF+BO+SMOTE+TM	0.68	0.66	0.55	0.61	0.57	0.72

- **Bold** values represent the maximum values.

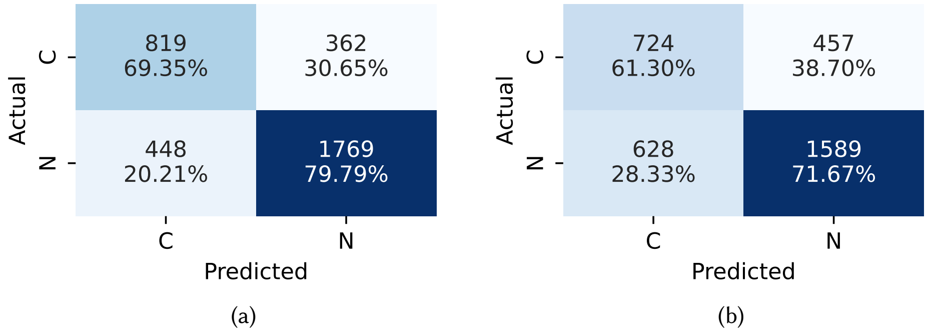


Fig. 5. The confusion matrices for the Combined dataset, employing the DNDT and DNDF classifiers through 10-fold cross-validation, are depicted in (5a) and (5b), respectively. In these matrices, "C" signifies COVID-19 Positive, while "N" signifies Non COVID-19 cough instances.

Table 10 presents a comparative analysis that offers a quantitative assessment of our suggested techniques for COVID-19 identification from cough samples, emphasizing the Combined dataset in particular. The DNDT+BO+SMOTE+TM method consistently demonstrates superior performance compared to the DNDF+BO+SMOTE+TM method across all evaluation metrics. Specifically, the DNDT+BO+SMOTE+TM method exhibits higher values for Accuracy (0.76), ROC-AUC score (0.75), Precision (0.65), Recall (0.69), F1 score (0.66), and Specificity (0.80). Figure 5 depicts the confusion matrices for the Combined dataset, showcasing the effectiveness of our proposed methodologies for accurate COVID-19 classification from cough sounds using DNDT and DNDF classifiers. The matrices for DNDT and DNDF classifiers, obtained through 10-fold cross-validation, are presented in (5a) and (5b), respectively. The DNDT classifier accurately identifies 819 out of 1,181 COVID-19 positive cough samples and 1,769 out of 2,217 cough samples from healthy individuals. Similarly, the DNDF classifier successfully predicts 724 out of 1,181 COVID-19 positive cough samples and 1,589 out of 2,217 cough samples from healthy individuals.

6 CONCLUSION AND FUTURE WORK

In conclusion, this paper presents a novel method employing deep neural decision trees and forests for the classification of COVID-19 based on cough sound analysis. Our method entails feature extraction, feature selection through RFECV, hyperparameter optimization via Bayesian Optimization, SMOTE for data augmentation during training, and the establishment of an optimal threshold using threshold moving. Extensive performance evaluation is conducted across five diverse datasets, including Cambridge, Coswara, COUGHVID, Virufy, and the combined Virufy with NoCoCoDa dataset. We present two ML-based methods for COVID-19 classification using cough sound data: DNDT+RFECV+BO+SMOTE+TM and DNDF+RFECV+BO+SMOTE+TM. On all datasets, our strategies consistently outperform the state-of-the-art techniques. They particularly obtain noteworthy AUC scores of 0.89, 0.95, 0.73, 0.69, 0.97, and 0.99 for the Cambridge Asymptomatic, Cambridge Symptomatic, Coswara, COUGHVID, Virufy, and Virufy merged with NoCoCoDa datasets, respectively, and impressive Recall ratings of 0.90, 0.93, 0.78, 0.74, 1, and 0.98. These findings highlight the efficiency of our proposed methods for correctly classifying COVID-19 from cough sound data, with potentially useful implications for the early detection and tracking of the illness.

In the scope of our comprehensive cross-dataset analysis encompassing five diverse datasets, our observations highlight noteworthy distinctions in model performance. Specifically, the Virufy merged with NoCoCoDa testing dataset achieves the highest AUC and F1 score, reaching 0.83 and 0.87, respectively, when the training dataset is Virufy. Furthermore, the Cambridge (Symptomatic) dataset showcases outstanding performance in recall, surpassing other testing datasets with an impressive score of 0.93, particularly when Coswara is the training dataset. Our in-depth analysis of classification efficiency and feature characteristics underscores that the Cambridge (Asymptomatic), Cambridge (Symptomatic), Virufy, and Virufy merged with NoCoCoDa datasets exhibit notable similarities in their classification attributes. Furthermore, we amalgamate the datasets from Cambridge (Asymptomatic), Cambridge (Symptomatic), Coswara, COUGHVID, and Virufy merged with NoCoCoDa to create a unified dataset for COVID-19 classification. Within the context of this consolidated dataset, our DNDT+BO+SMOTE+TM method outperforms the DNDF+BO+SMOTE+TM method. Demonstrating an Accuracy of 0.76, an AUC of 0.75, a Precision of 0.65, a Recall of 0.69, an F1-score of 0.66, and a Specificity score of 0.80, the DNDT+BO+SMOTE+TM method attains noteworthy metrics, underscoring its effectiveness in this expanded dataset scenario.

In our future endeavors, we aim to shift our focus towards the practical implementation of this research-centric method for initial COVID-19 detection in real-world settings. These findings should undergo rigorous validation with accurate COVID-19 labeling, which can be achieved, for example, through PCR testing. Furthermore, the integration of explainable AI techniques becomes indispensable to shed light on and enhance our understanding of the model's decision-making process—a pivotal aspect for the clinical deployment of AI-based COVID-19 detection systems.

REFERENCES

[1] Omar M Abdeldayem, Areeg M Dabbish, Mahmoud M Habashy, Mohamed K Mostafa, Mohamed Elhefnawy, Lobna Amin, Eslam G Al-Sakkari, Ahmed Ragab, and Eldon R Rene. 2022. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. *Science of The Total Environment* 803 (2022), 149834.

[2] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. 2020. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (covid-19) in China: A report of 1014 cases. *Radiology* (2020).

[3] Alican Akman, Harry Coppock, Alexander Gaskell, Panagiotis Tzirakis, Lyn Jones, and Björn W Schuller. 2022. Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges. *Frontiers in Digital Health* 4 (2022), 789980.

[4] Mahmood Al-khassaweneh and Raed Bani Abdelrahman. 2013. A signal processing approach for the diagnosis of asthma from cough sounds. *Journal of Medical Engineering & Technology* 37, 3 (2013), 165–171. PMID: 23631519.

[5] José Gómez Aleixandre, Mohamed Elgendi, and Carlo Menon. 2022. The Use of Audio Signals for Detecting COVID-19: A Systematic Review. *Sensors* 22, 21 (2022), 8114.

[6] Dr-Mohammed Aly and Nouf Alotaibi. 2022. A novel deep learning model to detect COVID-19 based on wavelet features extracted from Mel-scale spectrogram of patients' cough and breathing sounds. *Informatics in Medicine Unlocked* 32 (08 2022), 101049.

[7] Javier Andreu-Perez, Humberto Pérez-Espinosa, Eva Timonet, Mehrin Kiani, Manuel I. Girón-Pérez, Alma B. Benitez-Trinidad, Delaram Jarchi, Alejandro Rosales-Pérez, Nick Gatzoulis, Orion F. Reyes-Galaviz, Alejandro Torres-García, Carlos A. Reyes-García, Zulfiqar Ali, and Francisco Rivas. 2022. A Generic Deep Learning Based Cough Analysis System From Clinically Validated Samples for Point-of-Need Covid-19 Test and Severity Levels. *IEEE Transactions on Services Computing* 15, 3 (2022), 1220–1232.

[8] Arup Anupam, N Jagan Mohan, Sudarsan Sahoo, and Sudipta Chakraborty. 2021. 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)*. IEEE, 1391–1397.

[9] Alice E. Ashby, Julia A. Meister, Khuong An Nguyen, Zhiyuan Luo, and Werner Gentzke. 2022. Cough-based COVID-19 detection with audio quality clustering and confidence measure based learning. In *Proceedings of the Eleventh Symposium on Conformal and*

Probabilistic Prediction with Applications (Proceedings of Machine Learning Research, Vol. 179), Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson (Eds.). PMLR, 129–148.

[10] Muhammad Awais, Abhishek Bhuva, Dipen Bhuva, Saman Fatima, and Touseef Sadiq. 2023. Optimized DEC: An Effective Cough Detection Framework using Optimal Weighted Features-aided Deep Ensemble Classifier for COVID-19. *Biomedical Signal Processing and Control* (2023), 105026.

[11] G Ayappan and S Anila. 2023. Mayfly Optimization with Deep Belief Network-Based Automated COVID-19 Cough Classification Using Biological Audio Signals. *Cybernetics and Systems* (2023), 1–20.

[12] Idil Aytekin, Onat Dalmaz, Kaan Gonc, Haydar Ankishan, Emine U Saritas, Ulas Bagci, Haydar Celik, and Tolga Cukur. 2023. COVID-19 Detection from Respiratory Sounds with Hierarchical Spectrogram Transformers. arXiv:2207.09529 [cs.SD]

[13] Monelli Ayyavaraiah and Bondu Venkateswarlu. 2023. Adaptive Boosting Based Supervised Learning Approach for Covid-19 Prediction from Cough Audio Signals. *International Journal of Intelligent Systems and Applications in Engineering* 11, 6s (2023), 38–51.

[14] Piyush Bagad, Aman Dalmia, Jigar Doshi, Arsha Nagrani, Parag Bhamare, Amrita Mahale, Saurabh Rane, Neeraj Agarwal, and Rahul Panicker. 2020. Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds. *ArXiv* abs/2009.08790 (2020).

[15] Elmehdi Benmalek, Jamal El Mhamdi, Abdelilah Jilbab, and Atman Jbari. 2023. A cough-based Covid-19 detection with gammatone and mel-frequency cepstral coefficients. *Diagnostyka* 24, 2 (2023).

[16] Elmehdi Benmalek, Jamal El Mhamdi, Abdelilah Jilbab, and Atman Jbari. 2022. A cough-based COVID-19 detection system using PCA and machine learning classifiers. *Applied Computer Science* 18, 4 (2022).

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra (Eds.). 18 – 24.

[18] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. 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* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3474–3484.

[19] Mario Cesarelli, Marcello Di Giammarco, Giacomo Iadarola, Fabio Martinelli, Francesco Mercaldo, Antonella Santone, and Michele Tavone. 2022. COVID-19 Detection from Cough Recording by means of Explainable Deep Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1702–1707.

[20] Jiangeng Chang, Yucheng Ruan, Cui Shaoze, John Soong Tshon Yit, and Mengling Feng. 2022. UFRC: A Unified Framework for Reliable COVID-19 Detection on Crowdsourced Cough Audio. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 3418–3421.

[21] Yi Chang, Xin Jing, Zhao Ren, and Björn W Schuller. 2022. CovNet: A transfer learning framework for automatic COVID-19 detection from crowd-sourced cough sounds. *Frontiers in Digital Health* 3 (2022), 799067.

[22] Hanieh Chatrzarrin, Amaya Arcelus, Rafik Goubran, and Frank Knoefel. 2011. Feature extraction for the differentiation of dry and wet cough sounds. In *2011 IEEE International Symposium on Medical Measurements and Applications*. 162–166.

[23] Gunvant R. Chaudhari, Xinyi Jiang, Ahmed E. Fakhry, Asriel Han, Jaclyn Xiao, Sabrina Shen, and Amil Khanzada. 2020. Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough. *ArXiv* abs/2011.13320 (2020).

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[25] Muhammad EH Chowdhury, Nabil Ibtehaz, Tawsifur Rahman, Yosra Magdi Salih Mekki, Yazan Qibalwey, Sakib Mahmud, Maymouna Ezeddin, Susu Zughaier, and Sumaya Ali SA Al-Maadeed. 2021. QUCoughScope: an artificially intelligent mobile application to detect asymptomatic COVID-19 patients using cough and breathing sounds. *arXiv preprint arXiv:2103.12063* (2021).

[26] Nihad Chowdhury, Ashad Kabir, Md. Muhtadir Rahman, and Sheikh Mohammed Shariful Islam. 2022. Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Computers in Biology and Medicine* 145 (03 2022), 105405.

[27] Madison Cohen-McFarlane, Rafik A. Goubran, and Frank Knoefel. 2020. Novel Coronavirus Cough Database: NoCoCoDa. *IEEE Access* 8 (2020), 154087–154094.

[28] Harry Coppock, Lyn Jones, Ivan Kiskin, and Björn Schuller. 2021. COVID-19 detection from audio: seven grains of salt. *The Lancet Digital Health* 3, 9 (2021), e537–e538.

[29] Ganeshkumar Deivasikamani, Rohith C Manoj, et al. 2022. Covid Cough Classification using KNN Classification Algorithm. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 232–237.

[30] Vincenzo Dentamaro, Paolo Giglio, Donato Impedovo, Luigi Moretti, and Giuseppe Pirlo. 2022. AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath. *Pattern Recognition* 127 (2022), 108656.

[31] Gauri Deshpande, Anton Batliner, and Björn W Schuller. 2022. AI-Based human audio processing for COVID-19: A comprehensive overview. *Pattern recognition* 122 (2022), 108289.

[32] Gauri Deshpande and Björn Schuller. 2021. The DiCOVA 2021 Challenge: an encoder-decoder approach for COVID-19 recognition from coughing audio. 931–935.

[33] Vladimir Despotović, Muhannad Ismael, Maël Cornil, Rod McCall, and Guy Fagherazzi. 2021. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Computers in Biology and Medicine* 138 (2021), 104944 – 104944.

[34] Katharina Eggensperger, Matthias Feurer, Frank Hutter, James Bergstra, Jasper Snoek, Holger Hoos, Kevin Leyton-Brown, et al. 2013. Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In *NIPS workshop on Bayesian Optimization in Theory and Practice*, Vol. 10.

[35] Yunus Emre Erdoğan and Ali Narin. 2021. COVID-19 detection with traditional and deep features on cough acoustic signals. *Computers in Biology and Medicine* 136 (2021), 104765.

[36] Michael Esposito, Sunil Rao, Vivek Narayanaswamy, and Andreas Spanias. 2021. Covid-19 detection using audio spectral features and machine learning. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 1146–1150.

[37] Ahmed E. Fakhry, Xinyi Jiang, Jaclyn Xiao, Gunvant R. Chaudhari, Asriel Han, and Amil Khanzada. 2021. Virufy: A Multi-Branch Deep Learning Network for Automated Detection of COVID-19. In *Interspeech*.

[38] Ke Feng, Fengyu He, Jessica Steinmann, and Ilteris Demirkiran. 2021. Deep-learning Based Approach to Identify Covid-19. In *SoutheastCon 2021*. 1–4.

[39] Rahul Gomes, Connor Kamrowski, Jordan Langlois, Papia Rozario, Ian Dircks, Keegan Grottodden, Matthew Martinez, Wei Zhong Tee, Kyle Sargeant, Corbin LaFleur, et al. 2022. A comprehensive review of machine learning used to combat COVID-19. *Diagnostics* 12, 8 (2022), 1853.

[40] Rahul Gupta, Theodora Chaspari, Jangwon Kim, Naveen Kumar, Daniel Bone, and Shrikanth Narayanan. 2016. Pathological speech processing: State-of-the-art, current challenges, and future directions. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6470–6474.

[41] Rinki Gupta, Thota Ashavanthini Krishna, and Mohd Adeeb. 2022. Cough-based COVID-19 Detection with Multi-head Deep Neural Network. In *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Vol. 1. IEEE, 1–6.

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

[43] Esin Darici Haritaoglu, Nicholas Rasmussen, Daniel CH Tan, Jaclyn Xiao, Gunvant Chaudhari, Akanksha Rajput, Praveen Govindan, Christian Canham, Wei Chen, Minami Yamaura, et al. 2022. Using deep learning with large aggregated datasets for COVID-19 classification from cough. *arXiv preprint arXiv:2201.01669* (2022).

[44] Christopher Harte, Mark Sandler, and Martin Gasser. 2006. Detecting Harmonic Change in Musical Audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia* (Santa Barbara, California, USA) *(AMCMM '06)*. Association for Computing Machinery, New York, NY, USA, 21–26.

[45] John Harvill, Yash R Wani, Mark Hasegawa-Johnson, Narendra Ahuja, David Beiser, and David Chestek. 2021. Classification of COVID-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association, 4261–4265.

[46] Md Mahadi Hasan, Muhammad Usama Islam, Muhammad Jafar Sadeq, Wai-Keung Fung, and Jasim Uddin. 2023. Review on the evaluation and development of artificial intelligence for COVID-19 containment. *Sensors* 23, 1 (2023), 527.

[47] Abdelfatah Hassan, Ismail Shahin, and Mohamed Bader Alsabek. 2020. COVID-19 Detection System using Recurrent Neural Networks. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. 1–5.

[48] Ezz El-Din Hemdan, Walid El-Shafai, and Amged Mahmoud. 2022. CR19: a framework for preliminary detection of COVID-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications. *Journal of Ambient Intelligence and Humanized Computing* (02 2022).

[49] Aneeqa Ijaz, Muhammad Nabeel, Usama Masood, Tahir Mahmood, Mydah Sajid Hashmi, Iryna Posokhova, Ali Rizwan, and Ali Imran. 2022. Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey. *Informatics in Medicine Unlocked* 29 (2022), 100832.

[50] Ali Imran, Iryna Posokhova, Haneya N. Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N. John, MD Iftikhar Hussain, and Muhammad Nabeel. 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked* 20 (2020), 100378.

[51] Mesayu Elida Irawati and Hasballah Zakaria. 2021. Classification model for COVID-19 detection through recording of cough using XGboost classifier algorithm. In *2021 International Symposium on Electronics and Smart Devices (ISESD)*. IEEE, 1–5.

[52] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. 2021. Early detection of COVID-19 patients using chromagram features of cough sound recordings with machine learning algorithms. In *2021 International Conference on Microelectronics (ICM)*. IEEE, 82–85.

[53] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. 2022. A study of using cough sounds and deep neural networks for the early detection of Covid-19. *Biomedical Engineering Advances* 3 (2022), 100025.

[54] Dan-Ning Jiang, Lie Lu, HongJiang Zhang, Jianhua Tao, and Lianhong Cai. 2002. Music type classification by spectral contrast feature. *Proceedings. IEEE International Conference on Multimedia and Expo* 1 (2002), 113–116 vol.1.

[55] Varada Vivek Khanna, Krishnaraj Chadaga, Niranjana Sampathila, Srikanth Prabhu, Rajagopala Chadaga, and Shashikiran Umakanth. 2022. Diagnosing COVID-19 using artificial intelligence: A comprehensive review. *Network Modeling Analysis in Health Informatics*

and Bioinformatics 11, 1 (2022), 25.

[56] Hossein Khorramdelazad, Mohammad Hossein Kazemi, Alireza Najafi, Maryam Keykhaee, Reza Zolfaghari Emameh, and Reza Falak. 2021. Immunopathological similarities between COVID-19 and influenza: Investigating the consequences of Co-infection. *Microbial Pathogenesis* 152 (2021), 104554.

[57] Sera Kim, Ji-Young Baek, and Seok-Pil Lee. 2023. COVID-19 Detection Model with Acoustic Features from Cough Sound and Its Application. *Applied Sciences* 13 (02 2023), 2378.

[58] Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulò. 2015. Deep Neural Decision Forests. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1467–1475.

[59] Piyush Kumawat, Utkarsh, Aditya Chikhale, and Ramesh K Bhukya. 2022. COVID-19 Detection From Audio Signals Using LR-MLP-RF-GMM Classifiers. In *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. 1–6.

[60] Jordi Laguarta, Ferran Hueto, and Brian Subirana. 2020. COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020), 275–281.

[61] Jordi Laguarta, Ferran Hueto, and Brian Subirana. 2020. COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020), 275–281.

[62] Kranthi Kumar Lella and Alphonse PJA. 2021. A literature review on COVID-19 disease diagnosis from respiratory sound data. *arXiv preprint arXiv:2112.07670* (2021).

[63] Kranthi Kumar Lella and Alphonse Pja. 2022. 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, 2 (2022), 1319–1334.

[64] Yan Li and Liming Xia. 2020. Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management. *American Journal of Roentgenology* 214, 6 (2020), 1280–1286. PMID: 32130038.

[65] Anjali Malviya, Rahul Dixit, Anupam Shukla, and Nagendra Kushwaha. 2023. Long Short-Term Memory-based Deep Learning Model for COVID-19 Detection using Coughing Sound. *SN Computer Science* 4, 5 (2023), 505.

[66] Kavish Rupesh Mehta, Punid Ramesh Natesan, and Sumit Kumar Jindal. 2023. Proposed Experimental Design of a Portable COVID-19 Screening Device Using Cough Audio Samples. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*. Springer, 39–50.

[67] Mesut Melek. 2021. Diagnosis of COVID-19 and Non-COVID-19 Patients by Classifying Only a Single Cough Sound. *Neural Comput. Appl.* 33, 24 (dec 2021), 17621–17632.

[68] Jie Meng, Peng Zhang, Jianhua Wang, Aohui Wang, and Long Zhang. 2022. Detection of COVID-19 by Cough Sound: A Method Based on DSC+ BiLSTM. In *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE, 1–5.

[69] Puneet Misra and Arun Singh Yadav. 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol* 11, 3 (2020), 659–665.

[70] Emad A Mohammed, Mohammad Keyhani, Amir Sanati-Nezhad, S Hossein Hejazi, and Behrouz H Far. 2021. An ensemble learning approach to digital corona virus preliminary screening from cough sounds. *Scientific Reports* 11, 1 (2021), 15404.

[71] Pauline Mouawad, Tammuz Dubnov, and Shlomo Dubnov. 2021. Robust detection of COVID-19 in cough sounds: using recurrence dynamics and variable Markov model. *SN Computer Science* 2, 1 (2021), 34.

[72] Ananya Muguli, Lancelot Mark Pinto, R. Nirmala, Neeraj Kumar Sharma, Prashant Krishnan, Prasanta Kumar Ghosh, Rohit Kumar, Shreyas Ramoji, Shrirama Bhat, Srikanth Raj Chetupalli, Sriram Ganapathy, and Viral Nanda. 2021. DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. In *Interspeech*.

[73] Francis Nahm. 2022. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology* 75 (01 2022).

[74] Kazem Askari Nasab, Jamal Mirzaei, Alireza Zali, Sarfenaz Gholizadeh, and Meisam Akhlaghdoust. 2023. Coronavirus diagnosis using cough sounds: Artificial intelligence approaches. *Frontiers in Artificial Intelligence* 6 (2023).

[75] Dinh Son Nguyen, Khoa Tran Dang, and Huyen Trang Ton Nu. 2022. COVCOUGH: An Artificial Intelligence Application to Detect COVID-19 Patients through Smartphone-recorded Coughs. In *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 1518–1522.

[76] Long H Nguyen, Nhat Truong Pham, Liu Tai Nguyen, Thanh Tin Nguyen, Hai Nguyen, Ngoc Duy Nguyen, Thanh Thi Nguyen, Sy Dzung Nguyen, Asim Bhatti, Chee Peng Lim, et al. 2023. Fruit-cov: An efficient vision-based framework for speedy detection and diagnosis of sars-cov-2 infections through recorded cough sounds. *Expert Systems with Applications* 213 (2023), 119212.

[77] Lara Orlandic, Tomás Teijeiro, and David Atienza Alonso. 2020. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data* 8 (2020).

[78] Lara Orlandic, Tomás Teijeiro, and David Atienza. 2022. 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 (2022), 107743.

[79] P Padmalatha, Gowrisree Rudraraju, Narayana Rao Sripada, Baswaraj Mamidgi, Charishma Gottipulla, Charan Jalukuru, ShubhaDeepti Palreddy, Nikhil kumar Reddy Bhoge, Priyanka Firmal, Venkat Yechuri, PV Sudhakar, B Devimadhavi, S Srinivas, KKL Prasad, and Niranjan Joshi. 2022. Screening COVID-19 by Swaasa AI Platform using cough sounds: A cross-sectional study. medRxiv.

[80] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. 2021. COVID-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine* 135 (2021), 104572.

[81] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. 2022. COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. *Computers in Biology and Medicine* 141 (2022), 105153.

[82] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas R. Niesler. 2020. COVID-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine* 135 (2020), 104572 – 104572.

[83] Ankit Pal and Malaikannan Sankarasubbu. 2021. Pay Attention to the Cough: Early Diagnosis of COVID-19 Using Interpretable Symptoms Embeddings with Cough Sound Signal Processing. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (Virtual Event, Republic of Korea) *(SAC '21)*. Association for Computing Machinery, New York, NY, USA, 620–628.

[84] Irina Pavel and Iulian B Ciocoiu. 2022. Evaluation of Bag-of-Words Classifiers for COVID-19 Detection from Cough Recordings. In *2022 E-Health and Bioengineering Conference (EHB)*. IEEE, 1–4.

[85] Irina Pavel and Iulian B. Ciocoiu. 2023. COVID-19 Detection from Cough Recordings Using Bag-of-Words Classifiers. *Sensors* 23, 11 (2023).

[86] D. Pizzo, Sara Esteban, and M. Scetta. 2021. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples. *ArXiv* abs/2104.13247 (2021).

[87] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. 2023. A cough-based algorithm for automatic diagnosis of pertussis. *PLOS ONE* 18, 3 (2023), e0162128.

[88] Tawsifur Rahman, Nabil Ibtehaz, Amith Khandakar, Md. Sakib Hossain, Yosra Magdi, Maymouna Ezeddin, Enamul Bhuiyan, Mohamed Ayari, Anas Tahir, Yazan Qiblawey, Sakib Mahmud, Susu Zughaier, Tariq Abbas, Somaya Al-ma'adeed, and Muhammad Chowdhury. 2022. QUCoughScope: An Intelligent Application to Detect COVID-19 Patients Using Cough and Breath Sounds. *Diagnostics* 12 (04 2022), 920.

[89] Sunil Rao, Vivek Narayanaswamy, Michael Esposito, Jayaraman Thiagarajan, and Andreas Spanias. 2021. Deep Learning with hyper-parameter tuning for COVID-19 Cough Detection. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 1–5.

[90] Sunil Rao, Vivek Narayanaswamy, Michael Esposito, Jayaraman J Thiagarajan, and Andreas Spanias. 2021. COVID-19 detection using cough sound analysis and deep learning algorithms. *Intelligent Decision Technologies* 15, 4 (2021), 655–665.

[91] Alanazi Rayan, Sager holyl alruwaili, Alaa S. Alaerjan, Saad Alanazi, Ahmed I. Taloba, Osama R. Shahin, and Mostafa Salem. 2023. Utilizing CNN-LSTM techniques for the enhancement of medical systems. *Alexandria Engineering Journal* 72 (2023), 323–338.

[92] Zhao Ren, Yi Chang, Katrin D Bartl-Pokorny, Florian B Pokorny, and Björn W Schuller. 2022. The acoustic dissection of cough: diving into machine listening-based COVID-19 analysis and detection. *Journal of Voice* (2022).

[93] KC Santosh, Nicholas Rasmussen, Muntasir Mamun, and Sunil Aryal. 2022. A systematic review on cough sound analysis for Covid-19 diagnosis and screening: is my cough sound COVID-19? *PeerJ Computer Science* 8 (2022), e958.

[94] Björn Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panagiotis Tzirakis, Chloe Brown, Chauhan Jagmohan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, and Casper Kaandorp. 2021. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. 431–435.

[95] Neeraj Kumar Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, R. Nirmala, Prasanta Kumar Ghosh, and Sriram Ganapathy. 2020. Coswara - A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. *ArXiv* abs/2005.10548 (2020).

[96] Jiakun Shen, Xueshuai Zhang, Wenchao Wang, Zhihua Huang, Pengyuan Zhang, and Yonghong Yan. 2021. Cough-Based COVID-19 Detection with Multi-Band Long-Short Term Memory and Convolutional Neural Networks. In *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences* (Beijing, China) *(ISAIMS '21)*. Association for Computing Machinery, New York, NY, USA, 209–215.

[97] Jiakun Shen, Xueshuai Zhang, Pengyuan Zhang, Yonghong Yan, Shaoxing Zhang, Zhihua Huang, Yanfen Tang, Yu Wang, Fujie Zhang, and Aijun Sun. 2023. Piecewise Position Encoding in Convolutional Neural Network for Cough-Based Covid-19 Detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[98] Hamdi Skander, Abdelouahab Moussaoui, Mourad Oussalah, and Mohamed Saidi. 2022. *Autoencoders and Ensemble-Based Solution for COVID-19 Diagnosis from Cough Sound*. 279–291.

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

[100] Isabella Södergren, Maryam Pahlavan Nodeh, Prakash Chandra Chhipa, Konstantina Nikolaidou, and György Kovács. 2021. Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines. In *Interspeech*.

[101] Mohammad Soltanian and Keivan Borna. 2021. Covid-19 Recognition from Cough Sounds Using Lightweight Separable-Quadratic Convolutional Network. *Biomedical Signal Processing and Control* 72 (11 2021), 103333.

[102] Myoung-Jin Son and Seok-Pil Lee. 2022. COVID-19 Diagnosis from Crowdsourced Cough Sound Data. *Applied Sciences* 12 (02 2022), 1795.

[103] Brian Subirana, Ferran Hueto, Prithvi Rajasekaran, Jordi Laguarta, Susana Puig, Josep Malvehy, Oriol Mitjà, Antoni Trilla, Carlos Iván Moreno, Jos'e Francisco Munoz Valle, Ana Esther Mercado Gonz'alez, Barbara Vizmanos, and Sanjay E. Sarma. 2020. Hi Sigma, do I have the Coronavirus?: Call for a New Artificial Intelligence Approach to Support Health Care Professionals Dealing With The COVID-19 Pandemic. *ArXiv* abs/2004.06510 (2020).

[104] Vinayak Swarnkar, Udantha R. Abeyratne, Anne B. Chang, Yusuf A. Amrulloh, Amalia Setyati, and Rina Triasih. 2013. Automatic identification of wet and dry cough in pediatric patients with respiratory diseases. *Annals of Biomedical Engineering* 41, 5 (2013), 1016–1028.

[105] Mohammed Tawfik, Sunil Nimbhore, Nasser M Al-Zidi, Zeyad AT Ahmed, and Ali Mansour Almadani. 2022. Multi-features extraction for automating COVID-19 detection from cough sound using deep neural networks. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 944–950.

[106] Alberto Tena, Francesc Claria, and Francesc Solsona. 2022. Automated detection of COVID-19 cough. *Biomedical Signal Processing and Control* 71 (2022), 103175.

[107] Kien Trang, Hoang An Nguyen, Long TonThat, Hung Ngoc Do, and Bao Quoc Vuong. 2022. COVID-19 Disease Classification by Cough Records Analysis using Machine Learning. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE, 457–462.

[108] Sezer Ulukaya, Ahmet Alp Sarıca, Oğuzhan Erdem, and Ali Karaali. 2023. MSCCov19Net: multi-branch deep learning model for COVID-19 detection from cough sounds. *Medical & Biological Engineering & Computing* (2023), 1 – 11.

[109] Anand Vinod, Neethu Mohan, Sachin Kumar S, and Soman K P. 2023. Covid Cough Identification using Machine Learning and Deep Learning Networks. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*. 1–4.

[110] Wei Wang, Qijie Shang, and Haoyuan Lu. 2023. Automatic COVID-19 Detection from Cough Sounds Using Multi-Headed Convolutional Neural Networks. *Applied Sciences* 13, 12 (2023).

[111] Anne Wullenweber, Alican Akman, and Björn W Schuller. 2022. CoughLIME: Sonified explanations for the predictions of COVID-19 cough classifiers. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1342–1345.

[112] Tong Xia, Dimitris Spathis, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, and Cecilia Mascolo. 2021. COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening. In *NeurIPS Datasets and Benchmarks*.

[113] Ai Tang Xiao, Yi Xin Tong, and Sheng Zhang. 2020. False negative of RT-PCR and prolonged nucleic acid conversion in COVID-19: rather than recurrence. *Journal of Medical Virology* 92, 9 (2020), 1504–1507.

[114] Hao Xue and Flora D. Salim. 2021. Exploring Self-Supervised Representation Ensembles for COVID-19 Cough Classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021).

[115] Tianhao Yan, Hao Meng, Shuo Liu, Emilia Parada-Cabaleiro, Zhao Ren, and Björn W Schuller. 2022. Convoluational transformer with adaptive position embedding for Covid-19 detection from cough sounds. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9092–9096.

[116] Ouissam Zealouk, Hassan Satori, Mohamed Hamidi, Naouar Laaidi, Amine Salek, and Khalid Satori. 2021. Analysis of COVID-19 resulting cough using formants and automatic speech recognition system. *Journal of Voice* (2021).

[117] Rami Zewail, Tameem Bakr, and Ahmed Abdullatif. 2022. Resource-Aware Identification Of COVID-19 Cough Sounds Using Wavelet Scattering Embeddings. In *2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. 365–370.

[118] Xiyu Zhang, Michael Pettinati, Ali Jalali, Kuldeep Singh Rajput, and Nandakumar Selvaraj. 2021. Novel covid-19 screening using cough recordings of a mobile patient monitoring system. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2353–2357.

[119] Xueshuai Zhang, Jiakun Shen, Jun Zhou, Pengyuan Zhang, Yonghong Yan, Zhihua Huang, Yanfen Tang, Yu Wang, Fujie Zhang, Shaoxing Zhang, et al. 2022. Robust cough feature extraction and classification method for covid-19 cough detection based on vocalization characteristics. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*. 2168–2172.

Received 9 May 2024