machine learning & knowledge extraction

MDPI

Article

Audio-Based Screening of Respiratory Diseases Using Machine Learning: A Methodological Framework Evaluated on a Clinically Validated COVID-19 Cough Dataset

Arley Magnolia Aquino-García [1], Humberto Pérez-Espinosa [2,*], Javier Andreu-Perez [3] and Ansel Y. Rodríguez González [1,4,*]

[1] Tepic Technology Transfer Unit (UT3), Ensenada Scientific Research and Higher Education Center (CICESE), Andador 10, Between Streets 3 and 4, Tepic 63173, Nayarit, Mexico; aquinogarcia.arley@gmail.com
[2] National Institute of Astrophysics, Optics and Electronics (INAOE), Luis Enrique Erro No. 1, Santa Maria Tonantzintla, San Andres Cholula 72840, Puebla, Mexico
[3] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK; j.andreu-perez@essex.ac.uk
[4] Secretariat of Science, Humanities, Technology and Innovation (SECIHTI), Ciudad de México 03940, Mexico
* Correspondence: humbertop@inaoe.mx (H.P.E.); ansel@cicese.edu.mx (A.Y.R.G.)

Abstract

The development of AI-driven computational methods has enabled rapid and non-invasive analysis of respiratory sounds using acoustic data, particularly cough recordings. Although the COVID-19 pandemic accelerated research on cough-based acoustic analysis, many early studies were limited by insufficient data quality, lack of standardized protocols, and limited reproducibility due to data scarcity. In this study, we propose an audio analysis framework for cough-based respiratory disease screening research using COVID-19 as a clinically validated case dataset. All analyses were conducted on a single clinically acquired multicentric dataset collected under standardized conditions in certified laboratories in Mexico and Spain, comprising cough recordings from 1105 individuals. Model training and testing were performed exclusively within this dataset. The framework incorporates signal preprocessing and a comparative evaluation of segmentation strategies, showing that segmented cough analysis significantly outperforms full-signal analysis. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) for CNN2D models and the supervised Resample filter implemented in WEKA for classical machine learning models, both applied exclusively to the training subset to generate balanced training sets and prevent data leakage. Feature extraction and classification were carried out using Random Forest, Support Vector Machine (SVM), XGBoost, and a 2D Convolutional Neural Network (CNN2D), with hyperparameter optimization via AutoML. The proposed framework achieved a best balanced screening performance of 85.58% sensitivity and 86.65% specificity (Random Forest with GeMAPSvB01), while the highest-specificity configuration reached 93.90% specificity with 18.14% sensitivity (CNN2D with SMOTE and AutoML). These results demonstrate the methodological feasibility of the proposed framework under the evaluated conditions.

Keywords: acoustic cough analysis; class imbalance handling; respiratory disease detection; automatic classification; cough segmentation

Check for updates

Academic Editors: Tahir Cetin Akinci and Ömer Faruk Ertuğrul

Received: 16 January 2026
Revised: 14 March 2026
Accepted: 15 March 2026
Published: 20 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Advances in information and communications technologies (ICT) have enabled non-invasive, cost-effective systems for respiratory disease detection via automated cough analysis [1–4]. Before the COVID-19 pandemic, cough-based detection methodologies were primarily explored for chronic and non-communicable respiratory conditions, such as asthma, COPD, bronchiectasis, and pneumonia, focusing on controlled clinical scenarios, small-scale datasets, and disease-specific acoustic characterization [5]. While these early approaches demonstrated the feasibility of cough acoustic analysis as a diagnostic tool, they were often limited by reduced data availability, lack of standardized acquisition and labeling protocols, and insufficient attention to confounding factors such as comorbidities, demographics, and recording conditions.

Recent systematic reviews and clinical investigations have positioned AI-based cough-sound analysis not only as a diagnostic support tool but also as a scalable, non-invasive screening strategy with potential for early detection, triage, and population-level respiratory disease surveillance [6–10]. These studies emphasize that cough acoustics can complement laboratory testing by enabling rapid pre-screening in resource-limited and high-demand settings, where conventional diagnostic workflows may be slow and impractical.

The COVID-19 emergency in 2020 dramatically accelerated the development of cough analysis systems, shifting the focus from exploratory clinical studies toward rapid, large-scale screening tools leveraging artificial intelligence [11–15]. This accelerated transition revealed critical gaps inherited from pre-COVID-19 methodologies, including limited dataset validity, uncontrolled sources of bias, and weak reproducibility across studies. Many COVID-19 cough-based works emphasized isolated technical improvements without explicitly addressing whether the extracted acoustic markers were disease-specific or robust under real-world clinical variability.

This study is therefore guided by the central research question: Can robust and clinically meaningful acoustic signatures of respiratory disease be identified from standardized, clinically validated cough recordings when key methodological biases are systematically controlled? By explicitly connecting the rapid COVID-19-driven development of cough detection systems with unresolved limitations of pre-pandemic methodologies, this work frames cough-based COVID-19 detection as both a technical and methodological challenge.

Regarding the limitations of previous works, Coppock et al. [16] systematically identified seven critical methodological challenges that emerged prominently during the accelerated development of COVID-19 cough detection systems but are rooted in pre-COVID-19 research practices. These challenges highlight how earlier methodological gaps became amplified under pandemic-driven constraints:

- P1 Algorithms may not specifically detect COVID-19, reflecting pre-pandemic assumptions that cough acoustics alone are disease-discriminative.
- P2 Knowledge of disease status can bias cough production, a limitation rarely controlled in earlier cough studies.
- P3 Dataset validity and reliability are critical, particularly when rapid data collection compromises clinical verification.
- P4 Limited access to public datasets restricts reproducibility, a persistent issue prior to COVID-19.
- P5 Inclusion of non-cough sounds introduces confounding artifacts, exacerbated by uncontrolled recording environments.
- P6 Comorbidities and demographic variability complicate detection, and are often under-represented in pre-COVID datasets.
- P7 Lack of population diversity increases the risk of biased AI models.

To address these challenges, this study adopts five Study Provisions, each explicitly mapped to the limitations identified by Coppock et al. [16] and motivated by gaps in pre-COVID-19 detection methodologies:

SP1 Inclusion of both symptomatic and asymptomatic individuals directly mitigates P1 and P6 by reducing disease-label bias and improving clinical representativeness.
SP2 Collection of cough samples before diagnostic confirmation addresses P2 by minimizing behavioral bias associated with disease awareness.
SP3 Use of clinically validated samples responds to P3 by ensuring dataset reliability and diagnostic ground truth.
SP4 Explicit cough segmentation isolates true cough events, mitigating P5 by eliminating non-cough acoustic interference.
SP5 Use of a validated and publicly described database [13] contributes to addressing P4 and P7 by supporting reproducibility and population-level transparency.

Much of the literature about cough-based COVID-19 detection, including prior system-oriented approaches such as our earlier deep learning-based framework [13], treats preprocessing, segmentation, feature extraction, imbalance correction, and model selection as independent technical steps that are optimized in isolation and evaluated under heterogeneous protocols. This fragmented approach makes it difficult to determine whether reported performance gains arise from genuine disease-specific acoustic signatures or from methodological artifacts. Consequently, the central research gap is not the absence of more complex models, but the lack of a unified and bias-aware experimental framework that systematically evaluates how methodological design choices affect screening reliability and generalizability at the subject level.

Automated cough analysis using qRT-PCR-validated samples shows promise for COVID-19 screening when methodological rigor is enforced. Class imbalance was mitigated through undersampling [17] and oversampling [18], while AutoML techniques [19,20] optimized model parameters. By integrating lessons from pre-COVID-19 detection research with systematic bias control tailored to pandemic-scale data acquisition, this study demonstrates that robust cough-based diagnostic models can be derived from clinically validated and standardized acoustic data and evaluated under reproducible conditions.

Main Contributions

The main contributions of this study are summarized as follows:

- Formulation of a bias-aware experimental framework that explicitly links known methodological risks in cough-based diagnosis (P1–P7) with concrete mitigation strategies (SP1–SP5).
- Systematic and quantitative evaluation of cough-event segmentation, demonstrating its direct impact on sensitivity and specificity rather than treating segmentation as a fixed preprocessing step.
- Methodologically rigorous subject-level data partitioning to prevent data leakage and provide realistic performance estimates.
- Data-centric optimization through controlled class imbalance correction and AutoML-based hyperparameter tuning, enabling fair comparison across model families.
- Critical assessment of clinical screening feasibility, focusing on sensitivity and robustness instead of relying solely on aggregate accuracy metrics.

2. Related Works

Research on automatic COVID-19 detection has evolved across multiple methodological paradigms, ranging from imaging-based diagnostics to acoustic analysis and data-

centric machine learning strategies. These approaches differ not only in input modality but also in dataset acquisition protocols, validation rigor, feature representations, and evaluation methodologies. The following section organizes prior work into major categories to clarify methodological differences and identify recurring limitations that motivate the present study.

2.1. Imaging-Based COVID-19 Detection

Early machine learning approaches primarily relied on chest X-ray and computed tomography (CT) scans [21,22]. Several convolutional neural network architectures were specifically designed for COVID-19 detection from radiological images, including tailored frameworks such as COVID-Net [23]. Other studies demonstrated the feasibility of AI systems to distinguish COVID-19 from community-acquired pneumonia using CT scans under controlled clinical settings [24].

More recent analyses have emphasized both performance improvements and methodological challenges in imaging-based COVID-19 diagnosis. Systematic evaluations have highlighted common pitfalls related to dataset bias, lack of external validation, and potential overestimation of performance [25]. Additionally, recent reviews summarize advances and remaining limitations in CT-based deep learning systems [26].

Despite promising results, imaging-based systems require specialized equipment, clinical infrastructure, and trained personnel, limiting their scalability for rapid or remote population-level screening.

2.2. Acoustic Cough-Based Detection

Acoustic cough analysis has emerged as a non-invasive and accessible alternative, supported by the prevalence of cough as a symptom of respiratory infections [27]. Meister et al. [28] emphasized the feasibility of cough-based screening as a remotely deployable solution. Furthermore, the widespread availability of smartphones and AI-enabled portable devices strengthened the practicality of this approach [29].

Datasets developed for cough-based COVID-19 detection differ substantially in acquisition and diagnostic validation. Public datasets such as Coswara [30] rely primarily on self-reported COVID-19 status. Other datasets, such as Virufy [31], include qRT-PCR-confirmed labels collected under controlled conditions.

Table 1 summarizes representative audio datasets used for COVID-19 detection. Substantial variability can be observed in terms of dataset size, participant count, recording duration, and label validation protocols. Several large-scale datasets rely primarily on self-reported disease status, while clinically validated datasets tend to be comparatively smaller. Differences in sound modality (cough-only versus multimodal respiratory recordings) further contribute to methodological heterogeneity.

Table 1. Overview of publicly reported audio datasets for COVID-19 detection (n/a = not available).

Dataset	Samples (COVID)	Participants	Duration	Label Type	Sound Types
Virufy [31]	121 (48)	16	5 min	Clinically validated	Cough
COVID19-cough [32]	1324 (682)	n/a	58 min	Mixed/partially validated	Cough
Coswara [30]	2030 (343) †	n/a	50 h	Self-reported	Breathing, cough, voice
Tos COVID-19 [33]	5867 (2926)	2758	35 h	Self-reported, expert-labeled	Cough
COUGHVID [34]	27,550 (15,156)	21,543	35 h	Self-reported	Cough
COVID-19 Sounds [35]	53,449 (2106) *	36,116	552 h	Self-reported	Breathing, cough, voice

Such variability complicates direct comparison of reported performance metrics across studies, as classification outcomes may reflect differences in dataset composition, label reliability, or acquisition protocol rather than model architecture alone.

Reported performance metrics must therefore be interpreted within the context of dataset validation. For example, Andreu-Perez et al. [13] reported 96.43% sensitivity and 96.20% specificity using qRT-PCR-confirmed labels and standardized recording protocols combined with Empirical Mode Decomposition (EMD) and deep neural networks. However, performance metrics depend strongly on the evaluation protocol. When multiple cough segments are extracted per participant, subject-level partitioning is commonly recommended to ensure independence between training and test sets. For this reason, the present study enforces strict subject-level partitioning to obtain conservative estimates on unseen individuals. Sharma et al. [30] reported 66.74% accuracy using Random Forest classifiers on the self-reported Coswara dataset. Islam et al. [14] achieved 97.5% accuracy using deep neural networks trained on the clinically validated Virufy dataset. These differences highlight the influence of label reliability and acquisition protocol on reported outcomes.

Beyond dataset composition and validation protocols, cross-dataset generalization remains a critical challenge in cough-based COVID-19 detection. Studies evaluating model transferability across independent audio datasets have reported significant performance degradation when training and testing distributions differ [16]. Alternative deep architectures combining convolutional and recurrent layers have also been explored for cough-based detection [36], further illustrating architectural variability across studies.

2.3. Feature Representations and Deep Learning Architectures

Most acoustic-based systems rely on Mel-based spectral representations of cough signals. Mel spectrograms are commonly used, where the Mel scale applies a perceptually motivated nonlinear frequency transformation that approximates human auditory resolution. Mel-Frequency Cepstral Coefficients (MFCCs) are compact descriptors derived from the short-term power spectrum after mapping frequencies onto the Mel scale, capturing relevant spectral and timbral information.

Deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) models, and transformer-based architectures have been applied to these representations. Aytekin et al. [37] introduced a hierarchical spectrogram transformer trained on self-reported data. Other studies integrated acoustic features with clinical metadata [35,38,39]. While these works demonstrate the representational capacity of deep models, preprocessing and segmentation strategies are often described as part of the processing pipeline rather than comparatively evaluated.

2.4. Data-Centric Strategies and Imbalance Mitigation

Class imbalance remains a common challenge in COVID-19 cough datasets. Pahar et al. [40] addressed imbalance using SMOTE combined with ResNet50 and LSTM architectures, evaluated through cross-validation. Alsaif et al. [41] analyzed the effect of imbalance on minority class performance.

Beyond COVID-19 applications, systematic analyses have examined the broader impact of class imbalance in medical deep learning systems, emphasizing its influence on sensitivity and specificity trade-offs [42,43].

In addition, recent evaluations of medical AI models have highlighted the risk of performance overestimation due to data leakage and improper dataset partitioning strategies [44]. Careful distinction between subject-level and segment-level data partitioning is therefore essential to avoid optimistic bias and ensure realistic performance estimation.

2.5. Graph-Based and Alternative AI Paradigms

Beyond conventional acoustic modeling, graph-based machine learning approaches have been explored in broader COVID-19 research contexts. MEGA [45] applied graph analytics to infodemic risk modeling, while Graph Neural Networks (GNNs) have been used to model SARS-CoV-2 infection severity and structured biomedical relationships [46,47]. Additional studies have explored graph-based epidemic modeling and knowledge-graph-enhanced biomedical analytics [48,49].

Although these approaches do not directly address acoustic cough signal classification, they illustrate the broader trend toward structure-aware modeling in pandemic-related AI research.

2.6. Methodological Limitations and Research Gap

Across the literature, several recurring methodological considerations can be identified:

(i) variability in label validation protocols, including reliance on self-reported disease status in some datasets;
(ii) heterogeneous recording environments;
(iii) limited comparative evaluation of alternative segmentation strategies;
(iv) the need for strict subject-level independence during dataset partitioning;
(v) sensitivity of performance metrics to class imbalance handling strategies.

In clinically validated or partially validated approaches such as [13,40], segmentation is incorporated as part of the preprocessing pipeline; however, systematic comparison of segmentation strategies and explicit quantification of their impact on sensitivity and specificity are less frequently reported.

These observations motivate the present study, which emphasizes standardized clinical acquisition, strict subject-level independence, controlled imbalance correction, and systematic evaluation of multiple segmentation strategies within a unified experimental framework.

3. Proposed Screening Pipeline

We propose a screening pipeline for automatic respiratory disease detection using acoustic cough analysis, evaluated in this study on clinically validated COVID-19 cough recordings. The objective is to define a reproducible methodological framework that can be applied to new cough recordings under standardized acquisition conditions. Clinical deployment and claims of clinical effectiveness require external validation on independent cohorts and device/setting variability, which are outside the scope of this study. While this manuscript also includes a comparative evaluation of segmentation strategies and model configurations, those elements correspond to the research methodology used to justify specific design decisions and are described separately from the operational pipeline.

The screening pipeline consists of the following operational stages: (1) cough-event extraction using a predefined segmentation procedure; (2) temporal normalization to define a consistent analysis window; (3) feature extraction using 1D handcrafted descriptors or 3D time–frequency representations; (4) class imbalance handling applied exclusively within the training subset; (5) model training and inference; and (6) optional hyperparameter optimization during training. The comparative evaluation and selection of segmentation techniques are part of the research methodology of this study and are detailed in Section 3.3. These elements were used to justify the segmentation approach adopted in the proposed screening pipeline.

Several components of the proposed screening pipeline build upon established cough-based detection methodologies. Feature extraction using MFCC, Mel spectrograms, and acoustic descriptor sets such as GeMAPS has been widely adopted in COVID-19 cough classification studies [13,40,50]. Likewise, the use of classical machine learning algorithms

and convolutional neural networks is consistent with prior approaches in the field [13,40]. However, the objective of this work is not to introduce a disease-specific classifier, but to formalize a reproducible and generalizable methodological framework for respiratory disease screening based on cough acoustics. In this study, COVID-19 serves as a clinically validated case dataset to evaluate the framework. The contribution lies in structuring the full pipeline—from segmentation and temporal normalization to imbalance handling and subject-level partitioning—while explicitly quantifying how segmentation strategies affect downstream classification performance. As shown in Figure 1, this pipeline integrates all processing stages in a unified workflow. This positioning enables the proposed framework to be adapted to other respiratory diseases characterized by cough.

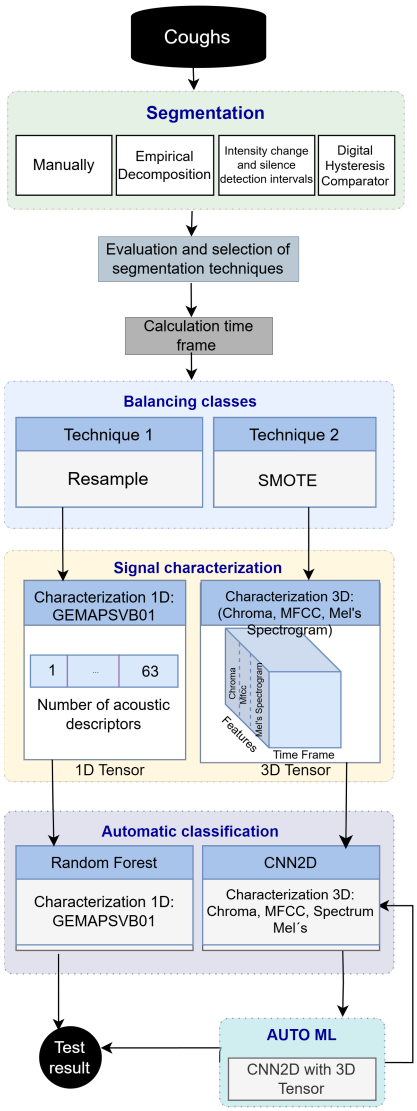


Figure 1. Method for automatic respiratory disease detection via cough. The pipeline includes: (a) database creation, (b) signal segmentation, (c) evaluation and selection of segmentation techniques, (d) time frame calculation, (e) class balancing, (f) signal characterization, and (g) classification and CNN2D parameter optimization with AutoML. Different colors indicate processing stages: green (segmentation), blue (class balancing), yellow (feature extraction), and purple (classification). Arrows represent the workflow sequence. 1D and 3D tensors refer to feature representations used as input to machine learning models. RF corresponds to Random Forest, and CNN2D to a two-dimensional convolutional neural network. SMOTE denotes Synthetic Minority Over-sampling Technique. AutoML refers to automated machine learning for hyperparameter optimization.

3.1. Data Used: Cough

The clinically validated cough dataset [13] was provided by CICESE, LANIIA, and the Spanish Hospital Costa del Sol. The data were collected between April and September 2020 during qRT-PCR testing under approved clinical and ethical protocols.

The research ethics and clinical protocols were approved by the corresponding local institutional bioethics committees in Mexico and Spain (Mexico approval code: **BIOETIC_HUM_2020_02**; Spain approval code: **APP_Covid19_03042020**). Data collection in both countries began in early April 2020, during the peak of the COVID-19 pandemic.

The dataset addresses five key concerns (SP1–SP5) defined in [16], including pre-diagnosis recordings and samples from both symptomatic and asymptomatic individuals. It comprises 1105 cough recordings (378 COVID-19-positive and 727 COVID-19-negative), with approximately 20% of positive cases corresponding to asymptomatic subjects.

Each cough recording corresponds to a distinct subject and was collected during a single session. Before segmentation, we had only one cough recording per individual.

Figure 2 shows the distribution. The dataset shows class imbalance, with negative samples predominating [51], which may bias models. Balancing strategies are applied to mitigate this.

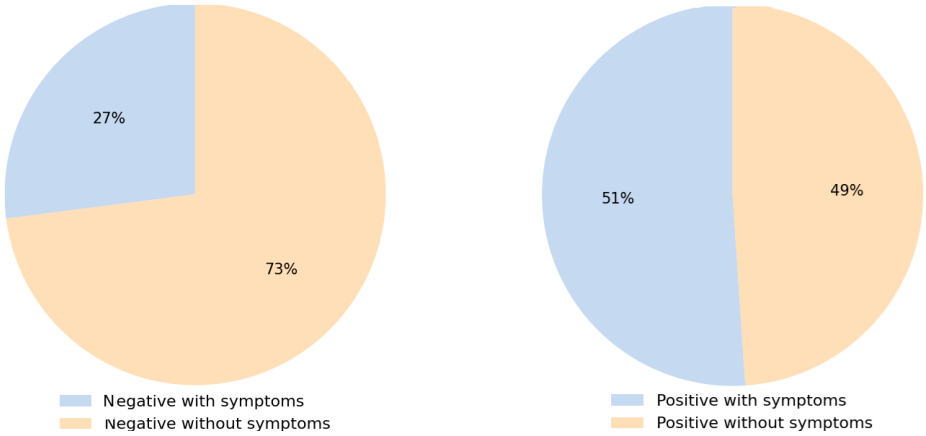


Figure 2. Proportion of symptomatic and asymptomatic individuals by COVID-19 positive and negative groups.

Recordings were collected in controlled clinical settings and semi-controlled outdoor environments with minimal ambient noise. Each sample includes a voice prompt from medical staff, requiring signal segmentation to isolate cough events. Dataset characteristics are summarized in Table 2.

Table 2. Summary of the dataset used in this study.

Attribute	Details
Total Participants	1105
COVID-19 Positive	378 (20% asymptomatic, 80% symptomatic)
COVID-19 Negative	727 (73% asymptomatic, 27% symptomatic)
Gender Positive	53% Male, 45% Female, 2% Not specified
Gender Negative	53% Male, 43% Female, 4% Not specified
Age: COVID-19 Positive	2.65% 1–20, 42.45% 21–40, 45.36% 41–60, 9.54% 61–80
Age: COVID-19 Negative	2.73% 1–20, 53.38% 21–40, 36.81% 41–60, 7.07% 61–80

Table 2. *Cont.*

Attribute	Details
Recording environment	Noise-free in a clinic room and outdoor, isolated from urban noise with nature sounds
Collection Period	4 April 2020, to 21 September 2020
Location	Malaga, Spain, and Tepic, Mexico
Clinical Validation	qRT-PCR testing for COVID-19 confirmation

3.2. Preprocessing: Segmentation

Audio segmentation identifies acoustic changes to isolate events like coughs [52]. This study evaluates three automatic methods against a manual baseline, where coughs are annotated by inspecting each recording individually.

Empirical Mode Decomposition (EMD) [53]: identifies modes reflecting coughing periods [13,54] and is applied in wheeze detection and segmentation [55–61]. It decomposes signals into intrinsic mode functions (IMFs).

Intensity Change and Silence Detection Intervals: detects acoustic changes via parameter adjustments for cough/silence detection with Praat [62], generating TextGrid files.

Digital Hysteresis Comparator: detects acoustic changes using upper/lower energy thresholds [63]. Oscillations below the minimum are silence, others are cough. Intervals of 230–550 ms are applied [64], with code from the detect-segment-cough repository [34].

3.3. Segmentation Techniques Evaluation

This subsection describes the evaluation protocol used in this study to compare alternative segmentation strategies and select the approach adopted in the proposed screening pipeline. This comparative assessment constitutes part of the research methodology rather than a stage of the deployable screening framework.

We propose a method to evaluate segmentation and identify the most effective approach. The evaluation procedure is detailed in Algorithm 1.

Algorithm 1 Segmentation Evaluation

Step 1: Sample Selection. Randomly select representative audio samples from both positive and negative classes.
Step 2: Manual Segmentation. Annotate cough and silence segments based on amplitude spectrograms; store annotations.
Step 3: Automatic Segmentation. Apply various automated segmentation methods to the same samples.
Step 4: Comparison. Align manual and automatic annotations to identify TP, TN, FP, and FN, including partial matches.
Step 5: Evaluation. Compute TP, TN, FP, and FN per sample to derive performance metrics such as accuracy and F-measure.

Segment classification is based on:

- **True Positive (TP):** Temporal overlap (non-zero intersection) between manual and automatic segments.
- **True Negative (TN):** Agreement on silence.
- **False Positive (FP):** Automatic detection where manual annotation indicates silence.
- **False Negative (FN):** Failure to detect manually annotated cough.

The method with the highest accuracy and F-measure is selected as optimal. Figure 3 displays the Praat interface for visualizing and editing TextGrid annotations. Panel (a) shows the manual segmentation used as the reference for evaluation. Panel (b) presents

the output of an automatic method, specifically the Digital Hysteresis Comparator implemented following the COUGHVID segmentation approach [34], which adopts a cough duration interval of 230–550 ms based on prior literature [64], for comparison.

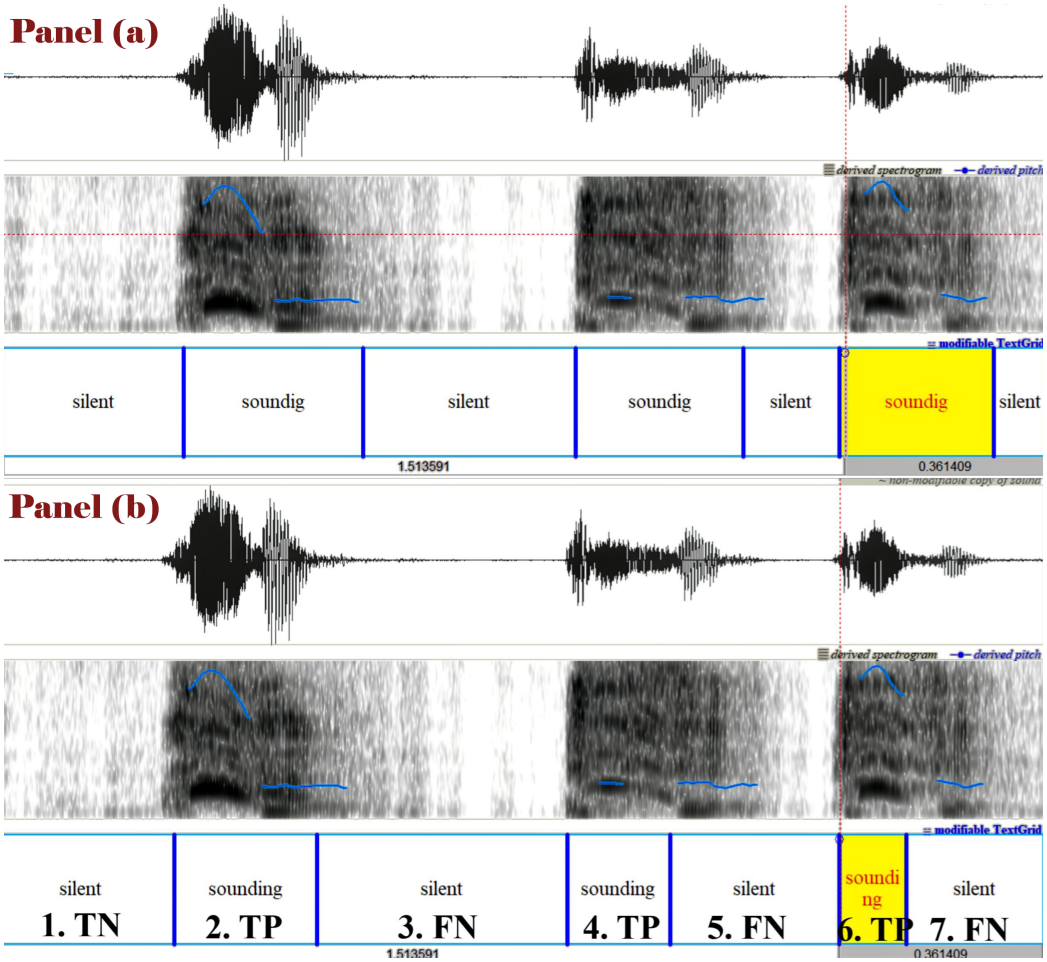


Figure 3. Praat interface showing `.wav` file and TextGrid for silence and cough segmentation: (**Top**) manual, (**Bottom**) automatic segmentation. The waveform (top panel) and spectrogram (bottom panel) are displayed along with the pitch contour (blue line). Labeled intervals ("silent" and "sounding") correspond to non-cough and cough-related acoustic events, respectively. The red dashed vertical line indicates the temporal cursor position. Blue boundaries represent segment limits defined in the TextGrid. Highlighted segments (yellow) indicate detected cough events. The labels correspond to classification outcomes for each segment. The numbered segments (1–7) correspond to: 1 (TN), 2 (TP), 3 (FN), 4 (TP), 5 (FN), 6 (TP), and 7 (FN).

The candidate segmentation techniques were selected to represent three distinct methodological families commonly used in acoustic event detection: adaptive signal decomposition (EMD), intensity-based thresholding (Intensity Change and Silence Detection), and energy-based hysteresis thresholding (Digital Hysteresis Comparator). These approaches differ in computational complexity, parameter sensitivity, and robustness to noise, allowing a balanced comparison between model-driven and threshold-based strategies. The selected parameter ranges were derived from prior cough physiology studies [64] and from publicly available implementations [34], ensuring physiological plausibility and methodological reproducibility rather than dataset-specific tuning.

3.4. Temporal Framing of Cough Signals

Machine learning models require feature vectors with consistent dimensions. As cough segments vary, a fixed time window standardizes sample length, aligned with the dataset

and segmentation strategy. Minimizing duration variability ensures reliability. The procedure for selecting the optimal window size is described in Algorithm 2.

Algorithm 2 Calculation of Time Frame

Step 1. Calculate audio duration, detect maximum and minimum audio duration, as well as sampling frequency.
Step 2. Determine if the data follows a normal distribution.
Step 3. Establish the 5% confidence interval of significance.
Step 4. Calculate the window with (1).
Step 5. The identified cough segments are cut out.

The time frames are calculated with (1). Where max is the maximum duration in seconds for the segments of a data set, and fs corresponds to the sampling rate of the segments.

$$TimeFrame = \left\lfloor \frac{max \cdot f_s}{h} \right\rfloor \quad (1)$$

where:

1. **max**: Maximum duration of audio segments (s) extracted from the database, defining the upper limit for frame calculation.
2. f_s: Sampling frequency (Hz), used to convert duration (max) into sample count.
3. h: Hop length (samples per frame), critical for segmenting signals during feature extraction.
4. **TimeFrame**: Total number of frames required to represent the maximum selected segment duration under hop length h.

A 4400 Hz sampling frequency was used. Based on the empirical duration distribution (Figure 4), the upper bound defined by the 97.5th percentile was $max = 0.9386312653$ s. Using a hop length of $h = 12$ samples, Equation (1) yields 344 frames to cover the longest representative segments. This window was set as the reference, truncating segments exceeding 0.9386 s for consistent feature extraction.

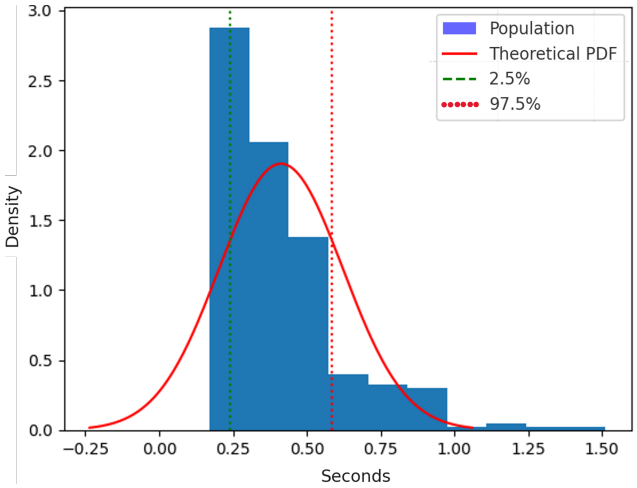


Figure 4. Segment duration histogram with Digital Hysteresis Comparator.

Figure 4 shows the cough duration histogram from the CICESE database. The green and red dotted lines mark the 2.5th and 97.5th percentiles, excluding segments outside 0.2–0.9386 s. The fixed 344-frame window serves as the temporal reference.

The red curve, a theoretical PDF of a normal distribution, highlights intrinsic variability in cough signals.

The Digital Hysteresis Comparator segments cough events of 0.20–2.4 s, reflecting the acoustic structure of coughs where brief silences are tolerated within thresholds. Although the segmenter may detect longer events, segments were temporally normalized by truncating durations above 0.9386 s to ensure fixed-size feature extraction.

Typical coughs last 230–550 ms [64], but longer events appear in the CICESE database due to signal traits (Figure 4). Segment durations were standardized using a 4400 Hz sampling rate and a 344-frame window (corresponding to 0.9386 s).

These findings align with studies reporting maximum durations near 2.5 s [65–68]. Including this duration threshold in the time frame calculation improved dataset-specific tuning and event representation.

Training and Test Sets

The dataset was split into 70% training and 30% testing sets using Scikit-learn's `train_test_split` with a fixed random state (42) for reproducibility. To prevent data leakage, all segments from a patient were allocated to a single subset. Because multiple cough events were extracted from each subject's single recording after segmentation, subject-level partitioning ensured that no derived segments from the same individual appeared in both subsets. The split was performed at the subject level with random assignment while preserving the natural class proportions of the dataset (Table 2). During training, class balancing was applied only to the training subset to mitigate class imbalance, while the test subset remained unchanged to support unbiased evaluation.

A 70/30 split was selected as a compromise between providing sufficient data for model training while preserving an independent hold-out subset for unbiased evaluation. Subject-level partitioning was enforced to prevent segments from the same individual appearing in both subsets, thereby minimizing optimistic bias and data leakage. Cross-validation was not employed for the primary evaluation because the objective was to simulate a realistic deployment scenario in which models are applied to unseen individuals rather than to randomly mixed segments. An internal validation split within the training data was used only for hyperparameter selection and did not replace the final evaluation on the held-out test subset.

Because segmentation produces multiple cough events per participant, the unit of partitioning was the participant (not the segment). Therefore, any comparison with studies that report results under different partitioning schemes should be interpreted cautiously, as performance metrics can be optimistically biased when participant-level independence is not enforced.

3.5. Signal Characterization

Audio signal characterization extracts dominant, discriminative features representing its properties [69], which machine learning algorithms use for learning and prediction. This work applies 1D and 3D characterizations for their proven effectiveness in acoustic cough analysis [13,50,70,71]. Below, we outline the 1D characterization techniques.

3.5.1. Characterization: 1D Tensor

Signal characterization represents acoustic signals with specific informational content. The techniques applied are:

- **Extensive Combination of Low-Level Acoustic Descriptors for Emotions (*Emolarge*)** [72,73].
- **Combination of Low-Level Acoustic Descriptors for Paralinguistic Phenomena (*Is*10_*paraling*)** [74,75].
- **Geneva Minimum Acoustic Parameter Set (*Gemapsvb*01)** [76].

These descriptor sets were selected because they represent standardized and widely validated acoustic parameter configurations originally developed for paralinguistic and affective computing tasks, where subtle temporal–spectral variations must be captured. Their predefined structure enhances reproducibility and comparability across studies. Rather than designing ad hoc feature sets, we adopted established configurations to reduce methodological arbitrariness and to evaluate their transferability to respiratory disease screening.

Feature standardization was applied during preprocessing using the `StandardScaler` function from the scikit-learn library. This transformation performs z-score normalization by centering each feature with respect to the mean and scaling it by the standard deviation computed from the training subset prior to model training. The standardized feature representations were then used to train the classical machine learning models evaluated in this study (SVM, XGBoost, and Random Forest).

3.5.2. Characterization: 3D Tensor

Audio characterization converts raw signals into numerical features for machine learning (ML) models. We used three common techniques—MFCC, Chroma, and Mel Spectrogram—each extracting 33 coefficients, forming a 3D tensor of shape (Window size $\times$ 33 $\times$ 3).

These methods were selected for their frequent use in cough classification studies [50] and strong performance in our preliminary tests.

- **MFCC** [77].
- **Chroma** [78].
- **Mel Spectrogram** [79].

3.6. Class Balancing

Class imbalance, common in medical datasets, arises from unequal class representation, typically with fewer disease-positive samples [80], as in the CICESE dataset with more COVID-19-negative cases.

To mitigate this, SMOTE (Synthetic Minority Oversampling Technique) generated synthetic minority samples on 3D feature segments, as validated for CNN2D [40,81–85]. We also applied the Resample technique [86] on 1D data, creating random sub-samples, effective with Random Forest, SVM, and XGBoost [87–89].

It is important to clarify that SMOTE and WEKA Resample were not applied simultaneously within the same model. SMOTE was exclusively used for balancing 3D time–frequency feature segments in CNN2D training, while the supervised Resample filter was applied to 1D acoustic descriptor representations used in classical machine learning models (SVM, XGBoost, and Random Forest). In all cases, balancing was performed only on the training subset to preserve test set independence.

3.7. Classification Algorithms

The evaluated algorithms include Random Forest [90], Support Vector Machine (SVM) [91], and XGBoost [92]. CNNs with 3D tensor inputs were also tested, following prior work on COVID-19 cough detection [13]. Random Forest served as a baseline. The next section details these algorithms.

Convolutional Neural Networks are designed for multidimensional data, such as RGB images with three 2D channels [93]. Our CNN2D processes 3D time–frequency tensors through four convolutional blocks using 2 $\times$ 2 kernels with `valid` padding, progressively reducing spatial dimensions. Spectrograms are padded (Algorithm 2) to ensure uniform input size.

Each convolutional block consists of a Conv2D layer with an increasing number of filters (16, 32, 64, and 128), followed by MaxPooling2D (2×2) for dimensionality reduction and 20% Dropout for regularization. This progressive increase in filter depth enables hierarchical learning of time–frequency representations from cough spectrograms, evolving from low-level spectral patterns toward more abstract acoustic structures.

After the final convolutional block, the feature maps are flattened (3648 units) and passed to a fully connected dense layer with 2 units and softmax activation for binary classification.

The kernel sizes and network depth were selected through automated hyperparameter optimization using AutoML, corresponding to the configuration that achieved the best validation performance among the evaluated candidates. The resulting architecture represents a balance between expressive capacity and overfitting control, particularly considering the size of the clinically validated dataset.

All CNN models were trained from scratch without pre-trained weight initialization. Training and testing splits were defined at the participant level, ensuring strict subject independence, and no participant appeared in more than one subset.

The final CNN2D model contains a total of 50,554 trainable parameters, as summarized in Table 3.

Table 3. CNN2D architecture based on model summary.

Block	Layer	Output Shape	Trainable Params
Block 1	Conv2D + ReLU	(32, 343, 16)	208
	MaxPooling2D	(16, 171, 16)	0
	Dropout (0.2)	(16, 171, 16)	0
Block 2	Conv2D + ReLU	(15, 170, 32)	16,448
	MaxPooling2D	(7, 85, 32)	0
	Dropout (0.2)	(7, 85, 32)	0
Block 3	Conv2D + ReLU	(6, 84, 64)	16,448
	MaxPooling2D	(3, 42, 64)	0
	Dropout (0.2)	(3, 42, 64)	0
Block 4	Conv2D + ReLU	(2, 41, 128)	16,448
	MaxPooling2D	(1, 20, 128)	0
	Dropout (0.2)	(1, 20, 128)	0
Flatten	Flatten	(3648)	0
Out	Dense (Softmax)	(2)	1002
Total			50,554

3.8. Parameter Optimization

AutoML automates hyperparameter optimization. Here, HyperOpt tuned CNN parameters [94–96], using the Tree-structured Parzen Estimator (TPE) algorithm [97], a Bayesian approach that models the objective function and iteratively updates parameter distributions. Users define the search space, and TPE refines it to find optimal configurations (Algorithm 3).

In this study, hyperparameter optimization was conducted using HyperOpt with the Tree-structured Parzen Estimator (TPE) algorithm. The dataset was first partitioned at the subject level into 70% training and 30% testing subsets to prevent data leakage, and the test subset was kept strictly untouched during model selection.

Within the 70% training subset, an internal validation split (80% training/20% validation) was created for hyperparameter selection. For each sampled hyperparameter configuration, a CNN2D model was trained on the internal training portion and evaluated

using ROC-AUC computed on the internal validation subset. The optimization objective was defined as loss $= -$ROC-AUC .

Algorithm 3 AutoML

Step 1: Train a model with an initial set of hyperparameters.
Step 2: Divide observed objective function values into "good" and "bad" groups.
Step 3: Calculate $P(x \mid good)/P(x \mid bad)$ for each hyperparameter.
Step 4: Identify hyperparameters that maximize model performance based on probabilities.
Step 5: Tune the model with the selected hyperparameters by training and evaluating performance.
Step 6: Repeat steps 2 to 5 until a stopping criterion is met (e.g., max_evals parameter in $fmin$ function). Limit the number of hyperparameter evaluations to ensure efficient execution.

The search was executed for a fixed number of evaluations ($max_evals = 10$). After selecting the optimal configuration, the model was retrained on the full 70% training subset and evaluated once on the untouched 30% test subset to obtain the final reported metrics.

The hyperparameter search space was defined as follows: dropout rate $\in \{0.25, 0.50, 0.75\}$, batch size $\in \{12, 28, 128\}$, optimizer $\in \{$Adadelta, Adam, RMSprop$\}$, and the maximum number of epochs set to 70.

For clarity and reproducibility, the complete processing pipeline was executed in the following order: (1) automatic cough segmentation using the selected Digital Hysteresis Comparator; (2) temporal framing and duration filtering as described in Algorithm 3; (3) subject-level partitioning into training (70%) and testing (30%) subsets to prevent data leakage; (4) feature extraction (1D acoustic descriptors or 3D time–frequency representations); (5) class imbalance mitigation applied exclusively to the training subset (SMOTE for CNN2D and Resample for classical machine learning models); and (6) model training and evaluation.

4. Results

Results are structured into three experiments. Experiment 1 compares three segmentation techniques based on accuracy and F-measure, where accuracy reflects correctly identified cough segments, and F-measure is the harmonic mean of precision and recall. Experiment 2 optimizes CNN2D training parameters. Experiment 3 evaluates classification performance using CNN2D and Random Forest, reporting sensitivity (true positive rate) and specificity (true negative rate). Each experiment is described in detail below.

4.1. Experiment 1: Segmentation Evaluation and Selection

Automatic segmentation was used to isolate cough events from the recordings prior to feature extraction and model training. Three automatic methods and manual segmentation were evaluated using Algorithm 1.

Thirty recordings (approximately 90 s each) containing patient verbalizations and voluntary coughs ($\sim$13 segments per recording, $\sim$0.19 s each) were selected for segmentation evaluation. Manual cough–silence annotation was performed on these 30 recordings to provide the reference labels for comparison.

Automatic methods were validated against manual segmentation using accuracy and F-measure (Table 4). Due to the high annotation cost, the validation set was limited to 30 samples for rigorous but feasible evaluation.

Manual segmentation requires approximately 96 min per recording, corresponding to approximately 24 h for 30 recordings. Performing this annotation on an initial pool of 120 recordings would require approximately 96 h of manual work. Therefore, a represen-

tative subset of 30 recordings was selected for the comparative evaluation of automatic segmentation methods against the manual reference. This subset included both COVID-19 positive and negative cases to preserve class diversity.

Table 4. Segmentation techniques and their performance (95% CIs).

Technique	Accuracy (95% CI)	F-Measure (95% CI)	Total Execution Time (h)
Intensity Change and Silence Detection Intervals	85.64% [85.17%, 86.11%]	84.95% [84.45%, 85.45%]	8.00
Digital Hysteresis Comparator	88.64% [88.10%, 89.18%]	89.00% [88.43%, 89.57%]	4.00
Empirical Mode Decomposition (EMD)	86.64% [86.14%, 87.14%]	86.84% [86.30%, 87.38%]	6.90
Manually	100.00% [100.00%, 100.00%]	100.00% [100.00%, 100.00%]	24.00

Based on Table 4, a one-way ANOVA showed significant differences in segmentation performance for Accuracy ($F = 54.54$, $p = 0.0178$) and F-measure ($F = 31.20$, $p = 0.0306$). Tukey HSD tests found no significant pairwise differences, likely due to limited sample size or low variability.

Manual segmentation achieved 100% Accuracy and F-measure but required 24 h. Among automated methods, the Digital Hysteresis Comparator provided the best trade-off, with 88.64% Accuracy and 89.00% F-measure in 4 h, while EMD and Intensity Change performed slightly lower. These results confirm the Digital Hysteresis Comparator as the most efficient automated approach, with manual segmentation as the reference.

4.2. Experiment 2: Parameter Optimization

Hyperparameter optimization was performed using AutoML to train the CNN2D model with a 344-time frame, as defined in the proposed method. The objective was to enhance training performance and mitigate overfitting in learning acoustic features. The optimized parameters for CNN2D, using CICESE COVID-19 cough data segmented by the Digital Hysteresis Comparator, are listed in Table 5.

Table 5. Optimized hyperparameters obtained using AutoML for the CNN2D architecture. found with AutoML for CNN2D.

Algorithm	CNN2D
Batch size	12
Epochs	68
Optimizer	Adam
Activation	ReLU
Regularization L1_L2	l1 = 0.0, l2 = 9.9999997 $\times 10^{-6}$
Classification layer activation function	Softmax

Based on the optimal hyperparameters identified via AutoML (Table 5), the CNN2D model was trained for 68 epochs using the Adam optimizer [98], with ReLU activations [99] in hidden layers and softmax [100] in the output for classification.

$L1_{L2}$ regularization, combining L1 (Lasso) and L2 (Ridge) penalties [101], was employed to enhance generalization by eliminating irrelevant features and preventing large weights.

The final model, retrained with these parameters, achieved a computational cost of 28,895,052 FLOPS per forward pass (TensorFlow Profiler). The architecture comprises

four convolutional layers with increasing filters, MaxPooling for dimensionality reduction, and Dropout with L2 regularization to control overfitting, ensuring a balance between performance and efficiency.

4.3. Experiment 3: Automatic Classification

Automatic classification performance was evaluated using sensitivity and specificity metrics for CNN2D models with 3D acoustic features (Chroma, MFCC, Mel Spectrogram) and Random Forest with 1D features (*Gemapsvb*01, *Is*_10*Paraling*, *Emolarge*), as described in the proposed method.

The computational complexity of extracting features from 0.19 s cough segments [102,103] was also analyzed. The 3D approach yields $33 \times 344 \times 3$ tensors, capturing rich acoustic patterns with complexity $O(6.97 \times 10^7)$, combining Mel Spectrogram, MFCC, and Chroma [104,105]. Despite its computational cost, this method suits high-performance applications.

Conversely, *Gemapsvb*01 characterization provides a lightweight alternative, with complexity of $O(17 \times 8000)$ for 16 kHz audio, ideal for resource-constrained settings. It is important to note that 1D acoustic descriptors (GeMAPSvB01, Emolarge, Is10_Paraling) were extracted at 16 kHz, while the 3D time–frequency representations used for CNN2D training were processed at 4.4 kHz following temporal normalization.

Table 6 summarizes algorithms, feature types, class balancing, metrics, and AutoML usage. Models were trained on COVID-19 coughs from the CICESE dataset, segmented with the Digital Hysteresis Comparator, which outperformed other automatic methods in precision and F-measure. A "-" symbol in Table 6 denotes non-applied techniques.

As shown in Table 6, Random Forest with resampling and 1D characterization (*Gemapsvb*01) achieved the highest sensitivity (85.58%) and the highest F1-score (86.05%), surpassing SVM and XGBoost across balanced performance metrics. In addition to sensitivity and specificity, complementary indicators including accuracy, precision (PPV), negative predictive value (NPV), and F1-score were computed to provide a comprehensive assessment of classifier behavior under class imbalance conditions. These additional metrics confirm that Random Forest provides the most stable trade-off between false positives and false negatives.

CNN2D with 3D characterization (MFCC, Chroma, Mel spectrogram), balanced via SMOTE and optimized with AutoML, attained the highest specificity (93.90%), but exhibited reduced sensitivity (18.14%). Although its overall accuracy remained comparatively high (77.33%), the reduced F1-score (26.00%) reflects limited positive case detection capability. Therefore, this configuration represents a specificity-oriented operating point rather than a balanced screening model.

SMOTE improved specificity but substantially reduced sensitivity, reflecting a shift toward a specificity-dominant decision boundary. This behavior is consistent with potential synthetic sample overlap and noise effects described in [106]. Despite its computational cost [107], SMOTE outperformed VAEs and wavelet-based augmentation, which were excluded due to lower performance and increased noise artifacts.

ANOVA revealed statistically significant differences in sensitivity ($F = 17.12$, $p = 0.0003$), while differences in specificity were not statistically significant ($F = 1.87$, $p = 0.1983$). Tukey HSD identified Random Forest as the most clinically balanced configuration (85.58% sensitivity, 86.65% specificity). In contrast, CNN2D configurations demonstrated asymmetric behavior, favoring either sensitivity or specificity depending on balancing and optimization strategy. The inclusion of F1-score and predictive values further reinforces this interpretation, confirming Random Forest as the most stable operating point within the evaluated framework.

Table 6. Automatic classification results with 95% Confidence Intervals (CIs). Bold values indicate the best performance. * Without segmentation.

Characterization	Algorithm	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	Precision (PPV)	NPV	F1	Balancing	Auto ML
Emolarge	SVM	10.03% (8.19%, 11.87%)	92.82% (91.66%, 93.98%)	63.70% (61.95%, 65.45%)	43.10% (36.80%, 49.40%)	65.54% (63.74%, 67.34%)	16.94%	Resample	-
Is10_Paraling	SVM	90.18% (88.87%, 91.49%)	12.76% (11.71%, 13.81%)	39.36% (38.10%, 40.62%)	35.11% (33.81%, 36.41%)	71.28% (67.87%, 74.69%)	50.69%	Resample	-
Gemapsvb01	SVM	36.02% (34.51%, 37.53%)	70.19% (68.19%, 72.19%)	47.75% (46.47%, 49.03%)	36.48% (34.96%, 38.00%)	69.77% (67.75%, 71.79%)	48.15%	Resample	-
Emolarge	XGBoost	42.03% (37.55%, 46.51%)	72.48% (68.63%, 76.33%)	57.26% (55.88%, 58.64%)	60.47%	55.55%	49.49%	Resample	-
Is10_Paraling	XGBoost	47.94% (43.28%, 52.60%)	70.54% (66.61%, 74.47%)	59.24% (57.87%, 60.61%)	61.96%	57.52%	54.00%	Resample	-
Gemapsvb01	XGBoost	44.76% (40.16%, 49.36%)	29.38% (25.14%, 33.62%)	37.07% (35.70%, 38.44%)	38.77%	34.73%	41.53%	Resample	-
Emolarge	Random Forest	77.68% (76.53%, 78.83%)	80.77% (79.65%, 81.89%)	79.23% (78.42%, 80.04%)	80.15% (79.03%, 81.27%)	78.34% (77.19%, 79.49%)	78.91%	Resample	-
Is10_Paraling	Random Forest	79.94% (78.53%, 81.35%)	86.29% (85.16%, 87.42%)	83.41% (82.51%, 84.30%)	83.58% (82.25%, 84.91%)	83.12% (81.95%, 84.29%)	81.67%	Resample	-
Gemapsvb01	**Random Forest**	**85.58% (82.56%, 88.60%)**	**86.65% (83.82%, 89.48%)**	86.10% (85.14%, 87.06%)	85.37%	86.85%	86.05%	**Resample**	**-**
* MFCC, Chroma, Mel	CNN2D	34.44% (24.63%, 44.25%)	77.77% (74.35%, 81.19%)	58.94% (52.23%, 65.65%)	54.39% (41.45%, 67.33%)	60.67% (52.84%, 68.50%)	42.47%	-	-
MFCC, Chroma, Mel	CNN2D	50.00% (45.59%, 54.41%)	57.35% (55.26%, 59.44%)	43.01%	71.88%	50.37%	-	Yes	
MFCC, Chroma, Mel	CNN2D	53.92% (49.56%, 58.28%)	87.98% (85.23%, 90.73%)	76.54% (74.99%, 78.09%)	69.42% (66.14%, 72.70%)	79.07% (77.36%, 80.78%)	60.69%	-	-
MFCC, Chroma, Mel	CNN2D	34.10% (29.97%, 38.23%)	78.01% (74.59%, 81.43%)	62.37% (60.33%, 64.41%)	46.18% (42.08%, 50.28%)	68.15% (65.86%, 70.44%)	38.87%	SMOTE	-
MFCC, Chroma, Mel	CNN2D	18.14% (14.61%, 21.67%)	**93.90% (91.64%, 96.16%)**	77.33% (74.90%, 79.76%)	45.45% (35.65%, 55.25%)	80.39% (77.98%, 82.80%)	26.00%	SMOTE	Yes

The configurations reported in Table 6 also allow partial comparison of different pipeline components, including segmentation, class balancing, and AutoML-based hyperparameter optimization. Although these comparisons illustrate how individual methodological choices influence model behavior, they do not constitute a controlled ablation study isolating each component independently.

Figure 5 presents t-SNE [108] clustering, illustrating effective class separation by CNN2D representations despite the observed imbalance in classification sensitivity.

Figure 6 shows Grad-CAM visualizations for CNN2D. Positive COVID-19 coughs (a) show diffuse activations, while negative coughs (b) display focused responses, particularly in mid-to-high frequency regions, indicating that the model attends to acoustically relevant spectral components. Although CNN2D demonstrates meaningful feature localization, its reduced sensitivity limits its effectiveness as a standalone screening configuration.

Figure 7 presents the confusion matrix for CNN2D with 3D characterization using SMOTE and AutoML, showing strong negative case accuracy (832 true negatives). False positives and false negatives highlight the model's specificity-oriented behavior and remaining limitations in positive case detection.

Figure 8 shows that the confusion matrix for Random Forest model achieved 4272 correct predictions (2136 true positives and 2136 true negatives), with 329 false negatives and 366 false positives. This distribution confirms its balanced detection capability across both classes.

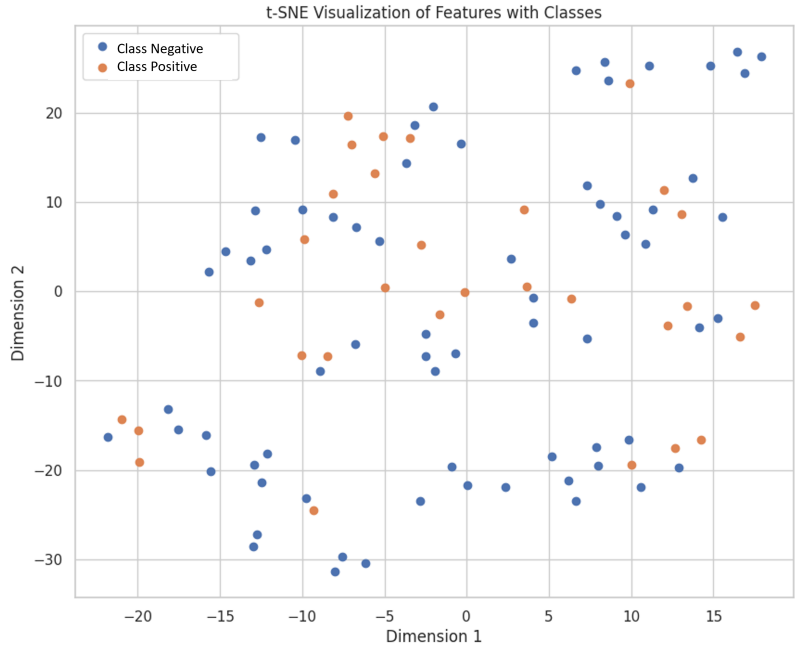
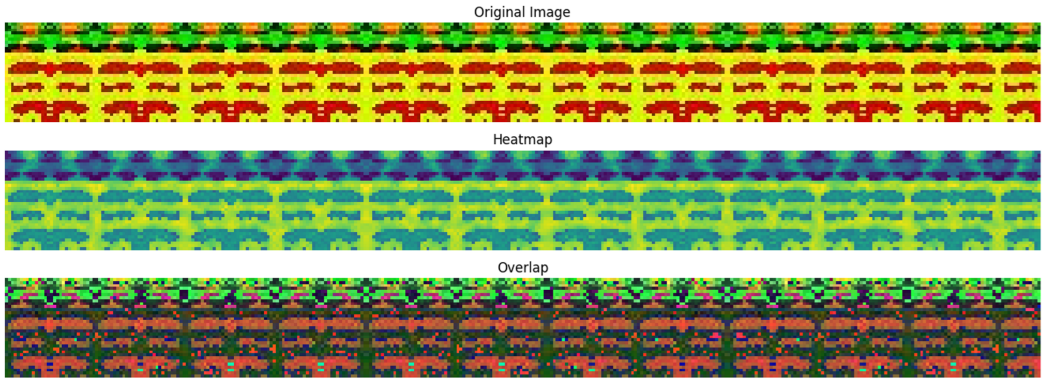
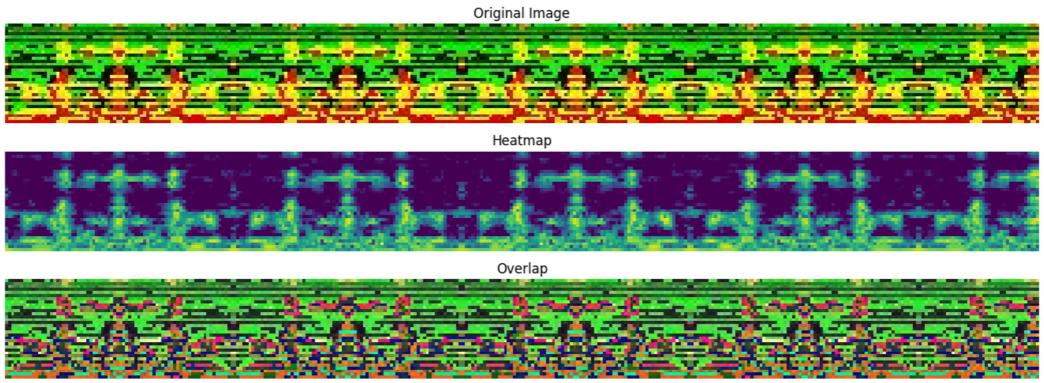


Figure 5. t-SNE plot for CNN2D classification.



(**a**) CNN2D positive COVID-19 class



(**b**) CNN2D negative COVID-19 class

Figure 6. Activation maps based on Grad-CAM for segmented coughs classified with CNN2D positive and negative COVID-19 classes.

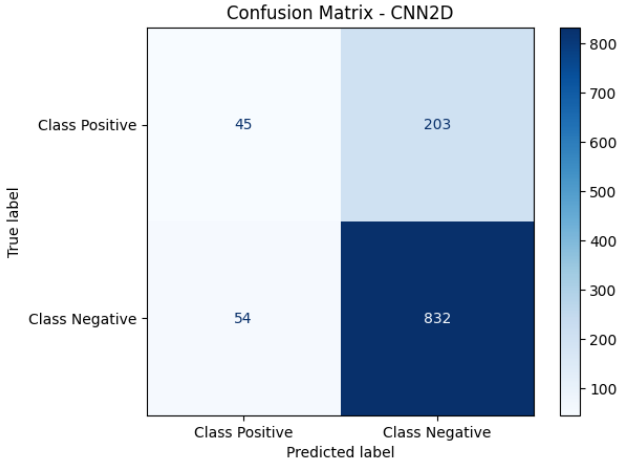


Figure 7. Confusion matrix for CNN2D classification of 248 positive and 886 negative cough segments.

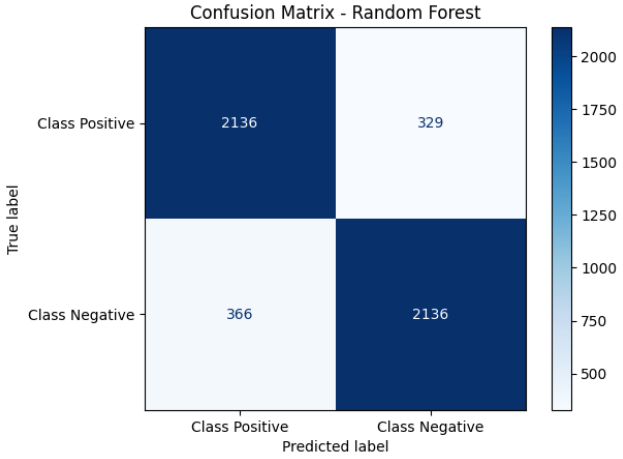


Figure 8. Confusion matrix for Random Forest classification of 2465 positive and 2502 negative cough segments.

4.4. Comparison with Previously Reported Results

This subsection compares the classification performance obtained in this study with previously published works reported in the Section 2 that address COVID-19 detection from cough sounds. The comparison focuses primarily on sensitivity and specificity due to their relevance for screening-oriented applications; however, the present study additionally reports accuracy, predictive values, and F1-score to ensure multi-metric transparency.

Table 7 summarizes the reported performance of representative prior works alongside the best-performing configurations obtained in this study. While numerical values provide useful reference points, direct comparison across studies must be interpreted cautiously due to differences in dataset composition, validation procedures, segmentation strategies, threshold selection, and model complexity.

In particular, evaluation protocols differ in whether data partitioning is performed at the participant level or at the segment level. When multiple segments are derived from a single participant recording, segment-level partitioning may place segments from the same individual into both training and test subsets, leading to potentially optimistic estimates due to participant-specific acoustic patterns. In this study, strict subject-level partitioning was enforced to prevent this form of leakage and to approximate a realistic screening scenario on unseen individuals.

Table 7. Comparison between previously reported works and the best results obtained in this study. Reported values correspond to those presented in the original publications; direct numerical comparison should be interpreted cautiously due to differences in datasets and evaluation protocols. Bold values highlight the results achieved in this study.

Study	Classification Model	Sensitivity (%)	Specificity (%)
Andreu-Perez et al. [13]	Deep Neural Network (EMD-based)	96.43	96.20
Pahar et al. [40]	ResNet50 (deep CNN)	98.50	94.20
This study	**Random Forest (GeMAPSvB01)**	**85.58**	86.65
This study	**CNN2D (AutoML + SMOTE)**	18.14	**93.90**

Previous studies, such as Andreu-Perez et al. [13] and Pahar et al. [40], report very high sensitivity and specificity using deep learning architectures, including ResNet-based models. These results highlight the potential of deep models for cough-based COVID-19 detection, but segmentation is typically treated as a fixed preprocessing step and model optimization is primarily architecture-driven.

In contrast, the present study places explicit emphasis on cough segmentation as a core methodological component. Multiple segmentation techniques are systematically evaluated, and their impact on downstream classification performance is quantitatively analyzed. The results demonstrate that appropriate segmentation substantially improves both sensitivity and specificity, even when using conventional machine learning algorithms.

Within this structured framework, Random Forest with GeMAPSvB01 features and resampling achieved a sensitivity of 85.58%, specificity of 86.65%, and F1-score of 86.05%, indicating clinically relevant and balanced performance without reliance on highly complex deep architectures. Furthermore, the CNN2D model with 3D acoustic features, optimized via AutoML and balanced using SMOTE, achieved the highest specificity (93.90%) within this study, demonstrating that deep models can be selectively employed when false-positive reduction is prioritized.

Overall, the contribution of this work lies not in surpassing previously reported peak performance values, but in providing a reproducible and systematically evaluated methodological framework for cough-based disease detection. By integrating clinically validated data, explicit segmentation evaluation, subject-level independence, imbalance mitigation, and multi-metric reporting, the proposed approach offers practical guidance for the structured design and evaluation of acoustic screening systems and provides methodological guidance for cough-based acoustic screening systems, while the present evaluation is limited to COVID-19 data.

Although this study employs the same clinically validated dataset described in [13], the objectives of the two studies differ. The prior work focused on optimizing a deep neural architecture to maximize classification performance on that dataset. In contrast, the present study focuses on defining and evaluating a structured methodological framework that includes segmentation analysis, subject-level partitioning, imbalance mitigation, and multi-metric reporting.

Accordingly, the goal of this work is not to maximize dataset-specific performance but to provide a transparent and reproducible experimental pipeline for cough-based screening research.

5. Discussion

Experiment 1 (Algorithm 2) confirmed the Digital Hysteresis Comparator as the most effective automatic segmentation method for COVID-19 coughs, achieving the highest ac-

curacy and F-measure. Its 4 h execution time contrasts sharply with the manual annotation effort (approximately 48 min per recording), making it suitable for scalable applications.

In comparison with prior studies, segmentation is frequently treated as a fixed preprocessing stage rather than as a methodological variable subject to evaluation [13,40]. While previous works have reported strong classification performance using predefined segmentation pipelines, they rarely quantify how segmentation strategy influences downstream model behavior. The present findings indicate that segmentation quality directly affects both sensitivity and specificity, suggesting that segmentation design should be considered a central methodological component rather than a secondary preprocessing choice.

Experiments 2 and 3 focused on optimizing CNN2D training through AutoML. With 28,895,052 FLOPS, the model maintains moderate computational complexity for mid-range GPUs or optimized CPUs. Dropout and L2 regularization improved generalization with limited additional cost, and reducing filter counts could further enhance efficiency.

Different configurations of segmentation, parameter optimization, and class balancing were associated with variations in classification performance. The CNN2D baseline configuration (33.44% sensitivity, 77.77% specificity) showed higher performance values (53.92% sensitivity and 87.98% specificity) when trained using cough segments obtained through the Digital Hysteresis Comparator. When SMOTE and AutoML were incorporated into the integrated pipeline configuration evaluated in this study, specificity reached 93.90%.

These results reflect the combined effect of the selected preprocessing, balancing, and optimization strategies within the complete pipeline configuration evaluated in this study. However, the present work does not include a formal ablation analysis to isolate the individual contribution of each component.

Random Forest with *Gemapsvb*01 achieved 85.58% sensitivity and 86.65% specificity, outperforming *Emolarge* and *Is*_10*Paraling* despite their larger feature sets. Importantly, it also achieved the highest F1-score (86.05%), confirming balanced discrimination across both classes. In contrast, CNN2D configurations optimized for specificity exhibited substantially lower F1-scores, reflecting asymmetric positive case detection. SVM and XGBoost showed lower overall performance with identical feature representations. Feature scaling may affect margin-based classifiers such as SVM; therefore, future work will evaluate standardized preprocessing pipelines under the same subject-level evaluation protocol. When contrasted with previously reported deep learning models that achieve higher peak sensitivity values [13,40], the results obtained here highlight a different contribution. Rather than prioritizing maximum classification accuracy within a specific dataset, this study emphasizes methodological transparency, subject-level independence, segmentation evaluation, and systematic analysis of imbalance mitigation strategies. The performance of Random Forest with standardized acoustic descriptors demonstrates that clinically relevant discrimination can be achieved without exclusive reliance on highly complex deep architectures.

Differences in reported performance across studies should therefore be interpreted in light of methodological design choices rather than solely as indicators of model superiority.

Particular attention must be given to the low sensitivity observed in the CNN2D baseline configuration (33.44%). In screening contexts, false negatives may delay diagnosis and facilitate continued disease transmission, which underscores the need for high sensitivity. Although CNN2D achieved competitive specificity in some configurations, its reduced sensitivity limits its suitability for screening purposes, where minimizing missed cases is essential.

This behavior may be related to representational and data-related factors, including the characteristics of 2D time–frequency inputs and dataset size; however, a controlled analysis isolating these factors was beyond the scope of this study.

These findings highlight that overall accuracy alone is insufficient for evaluating screening systems, and metric prioritization must align with the intended application.

For this reason, the present work reports multiple evaluation metrics and emphasizes balanced performance indicators rather than focusing exclusively on peak accuracy values.

Therefore, despite competitive specificity, CNN2D was not considered the most appropriate configuration for screening-oriented deployment in this study.

Overall, the pipeline configuration that combines the Digital Hysteresis Comparator, Algorithm 3 (time-frame selection), and CNN2D with 3D features, SMOTE, and AutoML achieved the highest specificity in this study.

In real-world respiratory disease screening scenarios, the relative importance of sensitivity and specificity depends on the intended clinical objective. For early triage applications—such as preliminary screening in primary care or large-scale surveillance—higher sensitivity is typically prioritized to minimize false negatives. In this context, the Random Forest configuration with GeMAPSvB01 features, achieving 85.58% sensitivity and 86.65% specificity, represents a balanced compromise between detection capability and false-positive control.

Conversely, in automated large-scale monitoring systems where false alarms may generate unnecessary diagnostic follow-up or resource burden, higher specificity becomes particularly valuable. Under such circumstances, the CNN2D configuration optimized with SMOTE and AutoML, achieving 93.90% specificity, may be preferable despite its reduced sensitivity.

Although the CNN2D configuration achieved high specificity (93.90%), its low sensitivity (18.14%) indicates that it would miss a substantial proportion of true positive cases. Such an operating point is not suitable for standalone screening applications, where minimizing false negatives is critical. Therefore, this configuration should not be interpreted as a primary screening solution but rather as an illustration of the sensitivity–specificity trade-off, potentially useful within multi-stage or confirmatory frameworks.

It should be noted that the reported sensitivity corresponds to the default classification threshold applied to the softmax outputs. No threshold tuning or operating point optimization was performed. Alternative threshold selection strategies, such as maximizing Youden's index or prioritizing sensitivity-constrained decision rules, may substantially alter the sensitivity–specificity balance. Therefore, the reported low sensitivity reflects the default operating configuration rather than the full discriminative capacity of the model.

These results indicate that model selection should be aligned with deployment priorities rather than guided solely by overall accuracy. A sequential screening strategy could therefore be considered, where a high-sensitivity model is applied initially and a high-specificity model subsequently refines classification decisions.

Importantly, this study positions COVID-19 as a clinically validated case dataset to evaluate a broader screening framework for cough-based respiratory disease detection. The contribution lies in structuring a reproducible pipeline that integrates segmentation evaluation, temporal normalization, subject-level partitioning, and multi-metric assessment.

Although both studies employ the same clinically validated dataset, direct performance comparison must consider methodological differences. Reference [13] optimized a specific deep neural architecture for maximum classification accuracy. In contrast, the present work prioritizes structural evaluation of the screening pipeline, including segmentation variability, subject-level independence, imbalance handling, and transparent metric reporting. These design choices may yield more conservative performance estimates but provide greater methodological robustness and transferability.

Accordingly, the objective of the present study is not to outperform previous models on the same dataset, but to analyze how methodological design decisions influence the reliability and reproducibility of cough-based screening systems.

The dataset employed in this study is multicentric, including recordings collected in Mexico and Spain under independent ethical approvals and standardized diagnostic procedures. The cohort encompasses symptomatic and asymptomatic individuals, a broad age range (10–80 years), and balanced sex distributions. However, no subgroup-specific performance analysis was conducted.Therefore, the study does not provide empirical evidence of bias mitigation across demographic subgroups; rather, it reports methodological safeguards (e.g., clinically validated labels, pre-diagnosis acquisition, subject-level partitioning) intended to reduce known sources of bias. Therefore, demographic invariance of classifier performance cannot be inferred from the current results, and future work should evaluate subgroup robustness to assess fairness and generalization. Although data in this study were collected under semi-controlled conditions, practical deployments may involve variability in microphone quality, mobile devices, sampling rates, and background noise. Such variability can affect cough segmentation accuracy and the stability of acoustic features, thereby altering signal-to-noise ratios and spectral characteristics. These considerations motivate future work on noise-robust preprocessing, device-invariant feature representations, and domain adaptation strategies to improve generalization in real-world screening scenarios.

Although confusion matrices are reported, no stratified error analysis was conducted across demographic variables or recording site. Although confusion matrices are reported, no stratified error analysis was conducted across demographic variables or recording site. Consequently, the distribution of false positives and false negatives across demographic or clinical subgroups was not analyzed in this study.

Because the ROC-AUC used for hyperparameter selection was computed on an internal validation split within the training data, future studies may further strengthen the evaluation by incorporating nested cross-validation or repeated subject-level splits to reduce variability while preserving strict test-set independence.

A formal ablation study isolating the contribution of segmentation, class balancing, and hyperparameter optimization was not conducted in this work. Future research may evaluate the relative impact of each component through controlled experimental comparisons to further refine the proposed framework.

Handcrafted features remain advantageous due to their interpretability, computational efficiency, and reproducibility, particularly in clinical and resource-constrained settings. However, they rely on predefined assumptions and may overlook complex nonlinear patterns in pathological cough signals. Learned representations may improve robustness and generalization but require greater computational resources and may reduce transparency. Hybrid approaches integrating interpretable descriptors with data-driven representations may offer a balanced path forward.

6. Conclusions

This work proposes a methodological framework for cough-based respiratory disease screening research using acoustic analysis, evaluated on a clinically validated COVID-19 dataset as a case study.

Random Forest with 1D characterization and CNN2D with AutoML-optimized 3D features represent complementary operating points within the proposed framework. Random Forest with *Gemapsvb*01 and resampling achieved 85.58% sensitivity, 86.65% specificity, and the highest F1-score (86.05%), confirming balanced discrimination under multi-metric evaluation. In contrast, CNN2D optimized with SMOTE and AutoML achieved the highest

specificity (93.90%) but reduced sensitivity (18.14%), illustrating the sensitivity–specificity trade-off inherent in model optimization.

Unsegmented audio or retained silence segments reduce classification performance by introducing irrelevant information and weakening discriminative feature extraction, reinforcing the importance of structured segmentation strategies.

The present study evaluated the proposed methodological framework exclusively on COVID-19 cough data. Its applicability to other cough-related respiratory conditions remains to be empirically validated using independent datasets. Future research may explore the extension of the framework to multiple conditions, alongside decision-threshold optimization and subgroup-specific robustness analysis to assess fairness and generalization.

Using a clinically validated multicentric dataset, this study advances structured segmentation analysis and transparent multi-metric evaluation of cough-based respiratory screening systems. The approach integrates signal processing and machine learning within a reproducible framework, offering transferable guidance for acoustic disease detection research.

Future research about the role of feature engineering should analyze how different acoustic descriptors contribute to cough pathology discrimination across segmentation levels. Approaches such as feature attribution, feature selection, and representation learning could help characterize interactions among descriptor sets and reveal how discriminative acoustic cues emerge under different temporal framing strategies. Such analyses may support the development of medical tools that combine interpretable handcrafted features with data-driven representations, ultimately improving both methodological transparency and screening performance.

These results provide a subject-level performance estimate under a leakage-aware evaluation protocol and support the methodological feasibility of the proposed framework as a research pipeline rather than evidence of clinical readiness for deployment.

Author Contributions: Conceptualization, H.P.-E. and J.A.-P.; methodology, H.P.-E. and J.A.-P.; software, A.M.A.-G.; validation, H.P.-E., J.A.-P. and A.Y.R.G.; investigation, A.M.A.-G.; writing—original draft preparation, A.M.A.-G.; writing—review and editing, H.P.-E., J.A.-P. and A.Y.R.G.; supervision, H.P.-E. and J.A.-P.; project administration, H.P.-E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Council on Humanities, Sciences and Technologies (CONAHCYT), Mexico, through a graduate study fellowship (CVU #1135474).

Institutional Review Board Statement: Institutional bioethics committees in Mexico and Spain approved the clinical and research ethics protocols used in this study (protocol BIOETIC_HUM_2020_02, Mexico; protocol APP_Covid_19_03042020, Spain). The study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study are available upon reasonable request from the corresponding author.

Acknowledgments: We thank the National Council of Humanities, Sciences and Technologies (CONAHCYT) for the financial support of the scholarship for Master's students with CVU no. 1135474 for providing us with the opportunity and the necessary tools to do this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AutoML	Automated Machine Learning
CI	Confidence Interval
CNN	Convolutional Neural Network
CNN2D	Two-Dimensional Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CT	Computed Tomography
DNN	Deep Neural Network
EMD	Empirical Mode Decomposition
FN	False Negative
FP	False Positive
GEMAPS	Geneva Minimalistic Acoustic Parameter Set
IMF	Intrinsic Mode Function
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
PCR	Polymerase Chain Reaction
qRT-PCR	Quantitative Reverse Transcription Polymerase Chain Reaction
ReLU	Rectified Linear Unit
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimator

References

1. Laguarta, J.; Hueto, F.; Subirana, B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J. Eng. Med. Biol.* **2020**, *1*, 275–281. [CrossRef] [PubMed]
2. Deshpande, G.; Schuller, B. An overview on audio, signal, speech, & language processing for COVID-19. *arXiv* **2020**, arXiv:2005.08579. [CrossRef]
3. Bagad, P.; Dalmia, A.; Doshi, J.; Nagrani, A.; Bhamare, P.; Mahale, A.; Rane, S.; Agarwal, N.; Panicker, R. Cough against covid: Evidence of covid-19 signature in cough sounds. *arXiv* **2020**, arXiv:2009.08790. [CrossRef]
4. Brown, C.; Chauhan, J.; Grammenos, A.; Han, J.; Hasthanasombat, A.; Spathis, D.; Xia, T.; Cicuta, P.; Mascolo, C. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *arXiv* **2020**, arXiv:2006.05919. [CrossRef]
5. Ghrabli, S.; Elgendi, M.; Menon, C. Identifying unique spectral fingerprints in cough sounds for diagnosing respiratory ailments. *Sci. Rep.* **2024**, *14*, 593. [CrossRef]
6. Kapetanidis, P.; Kalioras, F.; Tsakonas, C.; Tzamalis, P.; Kontogiannis, G.; Karamanidou, T.; Stavropoulos, T.G.; Nikoletseas, S. Respiratory Diseases Diagnosis Using Audio Analysis and Artificial Intelligence: A Systematic Review. *Sensors* **2024**, *24*, 1173. [CrossRef]
7. Hegde, S.; Sreeram, S.; Alter, I.L.; Shor, C.; Valdez, T.A.; Meister, K.D.; Rameau, A. Cough Sounds in Screening and Diagnostics: A Scoping Review. *Laryngoscope* **2024**, *134*, 1023–1031. [CrossRef]
8. Sharan, R.V.; Qian, K.; Yamamoto, Y. Automated Cough Sound Analysis for Detecting Childhood Pneumonia. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 193–203. [CrossRef]
9. Kim, J.; Choi, Y.S.; Lee, Y.J.; Yeo, S.G.; Kim, K.W.; Kim, M.S.; Rahmati, M.; Yon, D.K.; Lee, J. Limitations of the Cough Sound-Based COVID-19 Diagnosis Artificial Intelligence Model and Its Future Direction: Longitudinal Observation Study. *J. Med. Internet Res.* **2024**, *26*, e51640. [CrossRef]
10. Wen, G.; Wang, C.; Zhao, W.; Meng, J.; Xu, Y.; Wang, R.; Zeng, Z. Assessing chronic obstructive pulmonary disease risk based on exhalation and cough sounds. *Biomed. Eng. OnLine* **2025**, *24*, 82. [CrossRef]
11. Tena, A.; Claria, F.; Solsona, F. Automated detection of COVID-19 cough. *Biomed. Signal Process. Control* **2022**, *71*, 103175. [CrossRef] [PubMed]

12. Lella, K.K.; Pja, A. Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath. *Alex. Eng. J.* **2022**, *61*, 1319–1334. [CrossRef]
13. Andreu-Perez, J.; Pérez-Espinosa, H.; Timonet, E.; Kiani, M.; Girón-Pérez, M.I.; Benitez-Trinidad, A.B.; Jarchi, D.; Rosales-Pérez, A.; Gatzoulis, N.; Reyes-Galaviz, O.F.; et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. *IEEE Trans. Serv. Comput.* **2021**, *15*, 1220–1232. [CrossRef]
14. Islam, R.; Abdel-Raheem, E.; Tarique, M. A study of using cough sounds and deep neural networks for the early detection of COVID-19. *Biomed. Eng. Adv.* **2022**, *3*, 100025. [CrossRef]
15. Chowdhury, N.K.; Kabir, M.A.; Rahman, M.M.; Islam, S.M.S. Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Comput. Biol. Med.* **2022**, *145*, 105405. [CrossRef]
16. Coppock, H.; Jones, L.; Kiskin, I.; Schuller, B. COVID-19 detection from audio: Seven grains of salt. *Lancet Digit. Health* **2021**, *3*, e537–e538. [PubMed]
17. Nguyen, K.; Drossos, K.; Virtanen, T. Temporal sub-sampling of audio feature sequences for automated audio captioning. *arXiv* **2020**, arXiv:2007.02676. [CrossRef]
18. Bhattacharjee, M.; Ghosh, K.; Banerjee, A.; Chatterjee, S. Multilabel sentiment prediction by addressing imbalanced class problem using oversampling. In *Proceedings of the Advances in Smart Communication Technology and Information Processing: OPTRONIX 2020*; Springer: Singapore, 2021; pp. 239–249. [CrossRef]
19. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [CrossRef]
20. Alsharef, A.; Aggarwal, K.; Sonia.; Kumar, M.; Mishra, A. Review of ML and AutoML solutions to forecast time-series data. *Arch. Comput. Methods Eng.* **2022**, *29*, 5297–5311. [CrossRef]
21. Öztürk, Ş.; Özkaya, U.; Barstuğan, M. Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. *Int. J. Imaging Syst. Technol.* **2021**, *31*, 5–15. [CrossRef]
22. Kc, K.; Yin, Z.; Wu, M.; Wu, Z. Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images. *Signal Image Video Process.* **2021**, *15*, 959–966. [CrossRef]
23. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef] [PubMed]
24. Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* **2020**, *296*, E65–E71. [CrossRef] [PubMed]
25. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and diagnose COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [CrossRef]
26. Yang, J.; Zhang, Y.; Shi, Y. Deep learning for COVID-19 detection using chest CT: Recent advances and challenges. *Artif. Intell. Med.* **2023**, *139*, 102532.
27. Latif, S.; Usman, M.; Manzoor, S.; Iqbal, W.; Qadir, J.; Tyson, G.; Castro, I.; Razi, A.; Boulos, M.N.K.; Weller, A.; et al. Leveraging data science to combat COVID-19: A comprehensive review. *IEEE Trans. Artif. Intell.* **2020**, *1*, 85–103. [CrossRef]
28. Meister, J.A.; Nguyen, K.A.; Luo, Z. Audio feature ranking for sound-based COVID-19 patient detection. In *Proceedings of the Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 146–158. [CrossRef]
29. Nestor, B.; Hunter, J.; Kainkaryam, R.; Drysdale, E.; Inglis, J.B.; Shapiro, A.; Nagaraj, S.; Ghassemi, M.; Foschini, L.; Goldenberg, A. Machine learning COVID-19 detection from wearables. *Lancet Digit. Health* **2023**, *5*, e182–e184. [CrossRef]
30. Bhattacharya, D.; Sharma, N.K.; Dutta, D.; Chetupalli, S.R.; Mote, P.; Ganapathy, S.; Chandrakiran, C.; Nori, S.; Suhail, K.; Gonuguntla, S.; et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. *Sci. Data* **2023**, *10*, 397. [CrossRef]
31. Chaudhari, G.; Jiang, X.; Fakhry, A.; Han, A.; Xiao, J.; Shen, S.; Khanzada, A. Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. *arXiv* **2020**, arXiv:2011.13320. [CrossRef]
32. Ponomarchuk, A.; Burenko, I.; Malkin, E.; Nazarov, I.; Kokh, V.; Avetisian, M.; Zhukov, L. Project Achoo: A Practical Model and Application for COVID-19 Detection From Recordings of Breath, Voice, and Cough. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 175–187. [CrossRef]
33. Pizzo, D.T.; Esteban, S. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples. *arXiv* **2021**, arXiv:2104.13247. [CrossRef]
34. Orlandic, L.; Teijeiro, T.; Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci. Data* **2021**, *8*, 156. [CrossRef]
35. Xia, T.; Spathis, D.; Brown, C.; Ch, J.; Grammenos, A.; Han, J.; Hasthanasombat, A.; Bondareva, E.; Dang, T.; Floto, A.; et al. COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021 .

36. Muguli, A.; Pinto, J.; Natarajan, S. Cough-based COVID-19 detection using CNN-LSTM architecture. *Biomed. Signal Process. Control* **2022**, *68*, 102626.
37. Aytekin, I.; Dalmaz, O.; Gonc, K.; Ankishan, H.; Saritas, E.U.; Bagci, U.; Celik, H.; Çukur, T. COVID-19 Detection From Respiratory Sounds With Hierarchical Spectrogram Transformers. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1273–1284. [CrossRef]
38. Wu, Y.; Rocha, B.M.; Kaimakamis, E.; Cheimariotis, G.A.; Petmezas, G.; Chatzis, E.; Kilintzis, V.; Stefanopoulos, L.; Pessoa, D.; Marques, A.; et al. A deep learning method for predicting the COVID-19 ICU patient outcome fusing X-rays, respiratory sounds, and ICU parameters. *Expert Syst. Appl.* **2024**, *235*, 121089. [CrossRef]
39. Kim, S.; Baek, J.Y.; Lee, S.P. COVID-19 Detection Model with Acoustic Features from Cough Sound and Its Application. *Appl. Sci.* **2023**, *13*, 2378. [CrossRef]
40. Pahar, M.; Klopper, M.; Warren, R.; Niesler, T. COVID-19 cough classification using machine learning and global smartphone recordings. *Comput. Biol. Med.* **2021**, *135*, 104572. [CrossRef] [PubMed]
41. Alsaif, S.A.; Hidri, A. Impact of data balancing during training for best predictions. *Informatica* **2021**, *45*, 3479. [CrossRef]
42. Radiuk, P.M. Impact of training set imbalance on deep learning performance. *Int. J. Intell. Syst. Appl.* **2020**, *12*, 12–20.
43. Fernández, A.; García, S. Class imbalance in medical deep learning: Impact and mitigation strategies. *Inf. Fusion* **2023**, *94*, 33–52.
44. van de Mortel, L.A.; van Wingen, G.A. Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance. *Mol. Psychiatry* **2025**, *30*, 6070–6071. [CrossRef]
45. Hang, C.N.; Yu, P.D.; Chen, S.; Tan, C.W.; Chen, G. MEGA: Machine Learning-Enhanced Graph Analytics for Infodemic Risk Management. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 6100–6111. [CrossRef] [PubMed]
46. Sehanobish, A.; Ravindra, N.; van Dijk, D. Gaining Insight into SARS-CoV-2 Infection and COVID-19 Severity Using Self-supervised Edge Features and Graph Neural Networks. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 4864–4873. [CrossRef]
47. Zhang, Y.; Wang, X.; Li, P. Graph neural network for COVID-19 severity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence (AAAI): Washington, DC, USA, 2021; Volume 35, pp. 12345–12352.
48. Ke, G.; Meng, Q.; Finley, T.; Wang, T. Modeling COVID-19 epidemic using graph neural networks. *IEEE Access* **2020**, *8*, 123456–123467.
49. Chen, Y.; Wang, X. Knowledge graph-enhanced modeling for COVID-19 biomedical analytics. *Briefings Bioinform.* **2024**. [CrossRef]
50. García, A.M.A. Análisis Acústico de la tos para Detección de COVID con énfasis en el Balanceo de Clases. Master's Thesis, CICESE, Tepic, Mexico, 2023.
51. Rezvani, S.; Wang, X. A broad review on class imbalance learning techniques. *Appl. Soft Comput.* **2023**, *143*, 110415. [CrossRef]
52. Theodorou, T.; Mporas, I.; Fakotakis, N. An overview of automatic audio segmentation. *Int. J. Inf. Technol. Comput. Sci. (IJITCS)* **2014**, *6*, 1. [CrossRef]
53. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
54. Solak, F.Z. COVID-19 detection using variational mode decomposition of cough sounds. *Konya J. Eng. Sci.* **2023**, *11*, 354–369. [CrossRef]
55. Taplidou, S.A.; Hadjileontiadis, L.J. Wheeze detection based on time-frequency analysis of breath sounds. *Comput. Biol. Med.* **2007**, *37*, 1073–1083. [CrossRef]
56. Chien, J.C.; Wu, H.D.; Chong, F.C.; Li, C.I. Wheeze Detection Using Cepstral Analysis in Gaussian Mixture Models. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE: Piscataway, NJ, USA, 2007; pp. 3168–3171. [CrossRef]
57. Riella, R.; Nohama, P.; Maia, J. Method for automatic detection of wheezing in lung sounds. *Braz. J. Med Biol. Res.* **2009**, *42*, 674–684. [CrossRef]
58. Zhang, J.; Ser, W.; Yu, J.; Zhang, T. A Novel Wheeze Detection Method for Wearable Monitoring Systems. In *2009 International Symposium on Intelligent Ubiquitous Computing and Education*; IEEE: Piscataway, NJ, USA, 2009; pp. 331–334. [CrossRef]
59. Oletic, D.; Bilas, V. Asthmatic Wheeze Detection From Compressively Sensed Respiratory Sound Spectra. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1406–1414. [CrossRef]
60. De La Torre Cruz, J.; Cañadas Quesada, F.J.; Carabias Orti, J.J.; Vera Candeas, P.; Ruiz Reyes, N. Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds. *Expert Syst. Appl.* **2020**, *147*, 113212. [CrossRef]
61. Rocha, B.M.; Pessoa, D.; Marques, A.; de Carvalho, P.; Paiva, R.P. Automatic Wheeze Segmentation Using Harmonic-Percussive Source Separation and Empirical Mode Decomposition. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 1926–1934. [CrossRef] [PubMed]
62. Correa Duarte, J.A. *Manual de Análisis Acústico del Habla con Praat*; Number 49; Instituto Caro y Cuervo, Imprenta Patriótica: Bogotá, Colombia; 2014.

63. Miyara, F. Introducción a la Acústica. Publicación Interna de la Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR Rosario (Arg.) 2003. Available online: https://d1wqtxts1xzle7.cloudfront.net/34173522/acustica-libre.pdf (accessed on 11 March 2026).
64. Chang, A.B. The physiology of cough. *Paediatr. Respir. Rev.* **2006**, *7*, 2–8. [CrossRef]
65. Zhang, C. An overview of cough sounds analysis. In *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*; Atlantis Press: Dordrecht, The Netherlands, 2017; pp. 703–709.
66. Hiew, Y.H.; Smith, J.; Earis, J.; Cheetham, B.; Woodcock, A. DSP algorithm for cough identification and counting. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*; IEEE: Piscataway, NJ, USA, 2002; Volume 4, pp. IV–3888–IV–3891. [CrossRef]
67. Singh, V.P.; Rohith, J.; Mittal, V.K. Preliminary analysis of cough sounds. In *2015 Annual IEEE India Conference (INDICON)*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6. [CrossRef]
68. Piirilä, P.; Sovijärvi, A.R. Differences in Acoustic and Dynamic Characteristics of Spontaneous Cough in Pulmonary Diseases. *Chest* **1989**, *96*, 46–53. [CrossRef]
69. Sharma, G.; Umapathy, K.; Krishnan, S. Trends in audio signal feature extraction methods. *Appl. Acoust.* **2020**, *158*, 107020. [CrossRef]
70. Despotovic, V.; Ismael, M.; Cornil, M.; Mc Call, R.; Fagherazzi, G. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Comput. Biol. Med.* **2021**, *138*, 104944. [CrossRef]
71. Al-Tahri, M.; Al-Tamimi, N.; Al-Harbi, S.; Abduallah, A.; Alaskar, H.; Sbai, Z. Deceptive Detection based on Audio Spectrum Analysis using Deep Learning. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4. [CrossRef]
72. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In *2010 IEEE 18th Signal Processing and Communications Applications Conference*; IEEE: Piscataway, NJ, USA, 2009. [CrossRef]
73. Pérez-Rosas, V.; Mihalcea, R.; Resnicow, K.; Singh, S.; An, L. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1426–1435. [CrossRef]
74. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 2794–2797. [CrossRef]
75. Yazdani, A.; Simchi, H.; Shekofteh, Y. Emotion recognition in persian speech using deep neural networks. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*; IEEE: Piscataway, NJ, USA, 2021; pp. 374–378. [CrossRef]
76. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
77. Gupta, S.; Jaafar, J.; Ahmad, W.W.; Bansal, A. Feature extraction using MFCC. *Signal Image Process. Int. J.* **2013**, *4*, 101–108. [CrossRef]
78. Shah, A.; Kattel, M.; Nepal, A.; Shrestha, D. Chroma feature extraction. Chroma Feature Extraction using Fourier Transform. 2019. Available online: https://www.researchgate.net/profile/Ayush-Shah-6/publication/363487456_Chroma_Feature_Extractionpdf/data/631f9a1770cc936cd301efc1/Chroma-Feature-Extraction.pdf (accessed on 11 March 2026).
79. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
80. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **2019**, *90*, 103089. [CrossRef] [PubMed]
81. Özdemir, A.; Polat, K.; Alhudhaif, A. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. *Expert Syst. Appl.* **2021**, *178*, 114986. [CrossRef]
82. Wardoyo, R.; Wirawan, I.M.A.; Pradipta, I.G.A. Oversampling approach using radius-SMOTE for imbalance electroencephalography datasets. *Emerg. Sci. J.* **2022**, *6*, 382–398. [CrossRef]
83. Islam, M.T.; Islam, M.R.; Uddin, M.P.; Ulhaq, A. A deep learning-based hyperspectral object classification approach via imbalanced training samples handling. *Remote Sens.* **2023**, *15*, 3532. [CrossRef]
84. Dağılma, Y.; Özbay, E. Processing and Classification of Covid-19 Audio Data with Optimization-Based Learning Techniques. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
85. Lee, J.N.; Lee, J.Y. An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Appl. Sci.* **2023**, *13*, 3571. [CrossRef]
86. Bouckaert, R.R.; Frank, E.; Hall, M.; Kirkby, R.; Reutemann, P.; Seewald, A.; Scuse, D. *WEKA Manual for Version 3-9-1*; University of Waikato: Hamilton, New Zealand, 2016; pp. 1–341.

87. Naga RamaDevi, G.; Janga Reddy, M.; Baswaraj, D. Improve the Efficiency of the Classifiers Using Resample Technique on Image Segmentation Dataset. In *Proceedings of the ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 932–941. [CrossRef]
88. Aryanti, R.; Arifin, Y.T.; Khairunas, S.; Misriati, T.; Dalis, S.; Baidawi, T.; Safitri, R.A.; Marlina, S. The use of resampling techniques to overcome imbalance of data on the classification algorithm. *AIP Conf. Proc.* **2023**, *2714*, 020053 [CrossRef]
89. David, S.K.; Saeb, A.; Al Rubeaan, K. Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics. *Comput. Eng. Intell. Syst.* **2013**, *4*, 28–38.
90. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
91. Salcedo-Sanz, S.; Rojo-Álvarez, J.L.; Martínez-Ramón, M.; Camps-Valls, G. Support vector machines in engineering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [CrossRef]
92. Chen, T. Xgboost: Extreme Gradient Boosting; R Package Version 0.4-2. 2015 ; Volume 1. Available online: http://r.meteo.uni.wroc.pl/web/packages/xgboost/vignettes/xgboost.pdf (accessed on 11 March 2026).
93. Demir, F.; Abdullah, D.A.; Sengur, A. A new deep CNN model for environmental sound classification. *IEEE Access* **2020**, *8*, 66529–66537. [CrossRef]
94. Goyal, T.; Huang, P.; Sutton, F.; Maag, B.; Sommer, P. SMiLe–Automated End-to-end Sensing and Machine Learning Co-Design. In *Proceedings of the 2022 International Conference On Embedded Wireless Systems and Networks*; JMLR.org: Cambridge, UK, 2022; pp. 12–23.
95. Anwar, T. COVID19 Diagnosis using AutoML from 3D CT scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*; IEEE: Piscataway, NJ, USA, 2021; pp. 503–507.
96. Yazdani-Abyaneh, A.H.; Krunz, M. Automatic Machine Learning for Multi-Receiver CNN Technology Classifiers. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 39–44. [CrossRef]
97. Bergstra, J.; Yamins, D.; Cox, D.D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *SciPy* **2013**, *13*, 20
98. Singarimbun, R.N.; Nababan, E.B.; Sitompul, O.S. Adaptive Moment Estimation To Minimize Square Error In Backpropagation Algorithm. In *Proceedings of the 2019 International Conference of Computer Science and Information Technology (ICoSNIKOM)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7. [CrossRef]
99. Bai, Y. RELU-function and derived function review. In *Proceedings of the SHS Web of Conferences*; EDP Sciences: Les Ulis, France, 2022; Volume 144, p. 02006. [CrossRef]
100. Dunne, R.A.; Campbell, N.A. On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In *Proceedings of the 8th Australian Conference on Neural Networks (ACONN)*; Commonwealth Scientific and Industrial Research Organisation (CSIRO): Canberra, Australia, 1997; Volume 181, p. 185.
101. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; pp. 216–261. Available online: http://www.deeplearningbook.org (accessed on 11 March 2026).
102. Sainath, T.N.; Narayanan, A.; Weiss, R.J.; Variani, E.; Wilson, K.W.; Bacchiani, M.; Shafran, I. Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction. In *Proceedings of the Interspeech*; International Speech Communication Association (ISCA): San Francisco, CA, USA, 2016 ; pp. 1971–1975. [CrossRef]
103. Deshpande, G.; Viraraghavan, V.S.; Duggirala, M.; Patel, S. Detecting emotional valence using time-domain analysis of speech signals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: Piscataway, NJ, USA, 2019; pp. 3605–3608. [CrossRef]
104. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms*, 3rd. ed.; The MIT Press: Cambridge, MA, USA; London, UK, 2009; Volume 2, pp. 187–195.
105. Knuth, D.E. *The Art of Computer Programming*; Pearson Education: London, UK, 1997; Volume 3.
106. Liu, J. Importance-SMOTE: A synthetic minority oversampling method for noisy imbalanced data. *Soft Comput.* **2022**, *26*, 1141–1163. [CrossRef]
107. Rahman, D.A.; Lestari, D.P. COVID-19 Classification Using Cough Sounds. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6. [CrossRef]
108. Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill* **2016**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.