Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1142~1150
e-ISSN: 2548-6861

Comparative Evaluation of MFCC and Mel-spectrogram Features for CNN-Based Respiratory Abnormality Detection

Simboni Simboni Tege[1*, **, **], Kafunda Katalay Pierre[2**], Oshasha Oshasha Fiston[***, ****], Sylvestre Frey[4**], Albert Ntumba Nkongolo[5**], Biaba Kuya Jirince [6**, *****]**
* Department of Management Information Systems, Higher Pedagogical Institute of Isiro, Isiro, D.R. Congo
** Mention of Mathematcs, Statistics and Computer Science, University of Kinshasa, Kinshasa, D.R. Congo
*** General Commissariat for Atomic Energy, Regional Center for Nuclear Studies of Kinshasa, Kinshasa, D.R. Congo
**** CRIA-Center for Research in Applied Computing, Kinshasa, D.R. Congo
*****International School, Vietnam National University, Hanoi,Vietnam.
tege.simboni1@gmail.com [1], kafundakatalay@gmail.com [2], fiston.oshasha.oshasha@cgea-rdc.com [3], sylvestre.frey@unikin.ac.cd [4], albertntumba1994@gmail.com [5], jirincebiaba@gmail.com [6]

Article Info

Article history:

Received 2026-01-26
Revised 2026-02-17
Accepted 2026-03-02

Keyword:

Convolutional neural networks,
Respiratory sounds,
MFCC,
Mel-spectrogram,
Medical diagnosis.

ABSTRACT

Automated respiratory sound analysis addresses critical limitations in traditional clinical auscultation, particularly high inter-observer variability and limited specialist access in resource-constrained settings. This study rigorously compares Mel-Frequency Cepstral Coefficients (MFCC) and Mel-spectrogram representations for classifying respiratory abnormalities using convolutional neural networks. Using the ICBHI 2017 dataset (920 recordings, 6,898 cycles from 126 patients), we implemented identical CNN architectures differing only in input features. Class imbalance was addressed through Synthetic Minority Over-sampling Technique applied exclusively to training data. The MFCC model achieved 83% accuracy with superior sensitivity for normal sounds (97% recall), while Mel-spectrograms reached 82% accuracy with higher precision (95%). MFCC demonstrated better crackle detection (76% vs 73% recall) and wheeze precision (75% vs 71%), attributed to enhanced transient spectral capture through discrete cosine transformation. Both models showed strong discrimination (AUC > 0.90). MFCC offers computational efficiency advantages for screening applications, while Mel-spectrograms provide interpretability for diagnostic contexts. This controlled comparison provides evidence-based guidance for computer-aided respiratory diagnostic system design, particularly relevant for resource-limited healthcare environments.

I. INTRODUCTION

Respiratory auscultation remains fundamental to pulmonary disease diagnosis since Laennec's stethoscope invention in 1816. However, traditional methods face significant challenges including inter-observer variability exceeding 40% among clinicians and limited specialist access, particularly affecting primary care and resource-limited settings [1], [2]. These limitations compromise diagnostic reliability and healthcare accessibility, especially in regions with scarce pulmonology expertise [3].

The global respiratory disease burden continues escalating, with conditions like chronic obstructive pulmonary disease, asthma, pneumonia, and tuberculosis affecting millions worldwide. In sub-Saharan Africa, including the Democratic Republic of Congo, respiratory infections represent leading mortality causes among vulnerable populations. Electronic stethoscopes combined with artificial intelligence offer transformative potential by enabling objective, reproducible respiratory sound analysis that augments clinical decision-making and democratizes expert-level assessment [4], [5].

Normal respiratory sounds exhibit broadband frequency spectra (100-1000 Hz) from turbulent airflow [6], [7]. Pathological adventitious sounds manifest as discrete acoustic events. Crackles are discontinuous explosive sounds (<20 ms duration, 200-2000 Hz) from sudden airway opening,

characteristic of pneumonia and pulmonary fibrosis [8]. Wheezes are continuous musical sounds (>250 ms, 100-1000 Hz) from narrowed airway oscillations, indicating asthma and chronic obstructive pulmonary disease [9].

Effective classification requires acoustic features capturing pathological characteristics while maintaining computational efficiency. Two dominant approaches exist: Mel-frequency spectrograms preserve detailed time-frequency information through short-time Fourier transform with Mel-scale binning, facilitating expert visualization [10]. Mel-Frequency Cepstral Coefficients apply discrete cosine transform to log Mel-spectrograms, emphasizing spectral envelope in compressed format while achieving decorrelation, originally developed for speech recognition but applicable to biomedical sound tasks [11].

Deep learning advances, particularly convolutional neural networks applied to spectrogram representations, have catalyzed progress in automated respiratory analysis [12], [13]. Studies using the ICBHI 2017 dataset report 60-85% accuracies [14], [15]. However, systematic comparisons controlling architectural and training confounds remain limited. Most investigations use dissimilar architectures or preprocessing pipelines, preventing direct performance attribution to feature methodology. Class imbalance issues are often inadequately addressed, potentially biasing results.

This study addresses these gaps through rigorous controlled comparison of MFCC and Mel-spectrogram features for CNN-based classification. Our contributions include: implementing identical architectures differing only in input representation; systematic SMOTE application preventing data leakage; comprehensive multi-metric evaluation; detailed class-specific performance analysis; and computational efficiency discussion for resource-constrained deployment. The remainder describes methodology, presents results, discusses findings and implications, and concludes with recommendations.

II. METHOD

Our study follows the architecture shown below. Respiratory sounds are first acquired using an electronic stethoscope, then preprocessed and converted into Mel-spectrogram or MFCC representations. These features are fed into a Convolutional Neural Network (CNN) that automatically learns relevant patterns to classify each sound into one of four categories: Normal, Wheezes, Crackles, or Combined abnormalities.

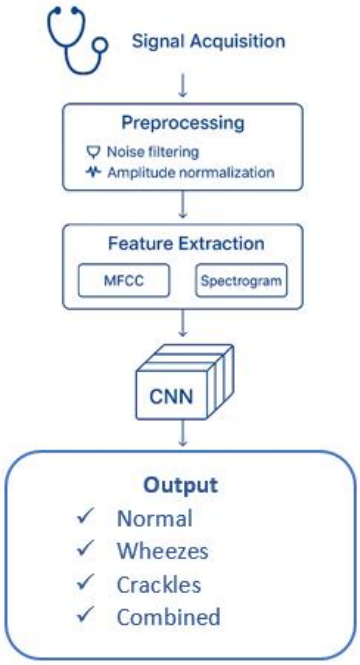

Figure 1: System Architecture

A. Dataset and Preprocessing

This study utilizes the ICBHI 2017 Respiratory Sound Database comprising 920 audio recordings from 126 subjects collected at multiple chest locations [16]. The dataset contains 6,898 respiratory cycles with expert annotations marking boundaries and indicating crackles and wheezes presence. Recordings were acquired using electronic stethoscopes at various sampling rates (4-44.1 kHz). Class distribution shows significant imbalance: 3,642 normal (52.8%), 1,864 crackles (27.0%), 886 wheezes (12.9%), and 506 combined abnormalities (7.3%).

Preprocessing followed a systematic pipeline implemented in Python using Librosa [17] and SciPy. For each recording, the audio signal was first loaded at its native sampling rate (4–44.1 kHz), then resampled to a uniform 44,000 Hz using librosa.load(). Individual respiratory cycles were extracted based on expert annotations (start_time, end_time in seconds) by computing sample indices: start_sample = int(start_time × 44000), end_sample = int(end_time × 44000). Each extracted segment was then resampled to exactly 44,000 samples (1 second duration) using scipy.signal.resample(), which applies Fourier-based interpolation to preserve spectral characteristics while ensuring uniform CNN input dimensions regardless of original cycle duration. No additional bandpass filtering was applied beyond the implicit anti-aliasing during resampling, as the full frequency spectrum was preserved to avoid discarding clinically relevant information. Implicit amplitude normalization occurred during the logarithmic compression step in feature extraction (see Section 2.2). Dataset splitting followed a strict patient-independent

protocol: all cycles from a given patient were assigned exclusively to either the training set (70% of patients) or the test set (30% of patients), preventing any data leakage between sets. This patient-level split yielded 4,828 training cycles from 88 patients and 2,069 test cycles from 38 patients prior to SMOTE augmentation.

B. Feature Extraction

Mel-spectrograms were computed using the librosa.feature.melspectrogram() function with the following parameters: short-time Fourier transform (STFT) with n_fft=2048 samples (Hann window by default in Librosa), hop_length=512 samples, yielding 86 temporal frames per 1-second segment. We employed n_mels=193 Mel-frequency bands with triangular overlapping filters distributed on the perceptual Mel scale. The resulting power spectrogram was converted to decibel scale using librosa.power_to_db(), applying logarithmic compression: dB = 10 × log10(power), with reference to the maximum power value. This produced 193×86 matrices (float32 precision) representing time-frequency characteristics. The Mel scale provides perceptually-motivated frequency resolution, with higher resolution at lower frequencies where human auditory discrimination is finest, making it particularly suitable for biomedical sound analysis where pathological features span wide frequency ranges.

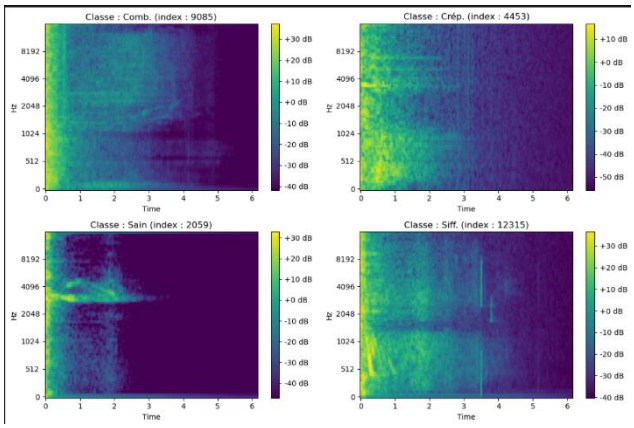

Figure 2: Representative Mel-spectrograms of respiratory sounds: (Sain) Normal, (Crép) Crackles, (Siff) Wheezes, (Comb) Combined abnormalities.

MFCC extraction was performed using the librosa.feature.mfcc() function with identical preprocessing parameters to ensure fair comparison. Specifically: y=resampled_audio (44,000 samples), sr=44000, n_mfcc=60, with implicit STFT parameters n_fft=2048 (Hann window), hop_length=512, yielding 86 temporal frames per segment. The extraction pipeline implemented by Librosa consists of: (1) computing the Mel-spectrogram with n_mels=128 (Librosa default for MFCC computation), (2) applying natural logarithm compression (ln rather than log10 for MFCCs), and (3) performing discrete cosine transform (DCT-II) on the log-Mel spectrogram for decorrelation and dimensionality compression, yielding 60×86 matrices (float32 precision). The choice of n_mfcc=60 (rather than the standard 13) was empirically validated: preliminary experiments with 13, 40, and 60 coefficients showed monotonically improving classification accuracy, with 60 coefficients capturing sufficient spectral detail for respiratory pathology discrimination while maintaining 70% dimensionality reduction compared to raw Mel-spectrograms (60 vs 193 frequency bins). Higher-order MFCC coefficients encode fine spectral modulations that are particularly relevant for transient events like crackles.

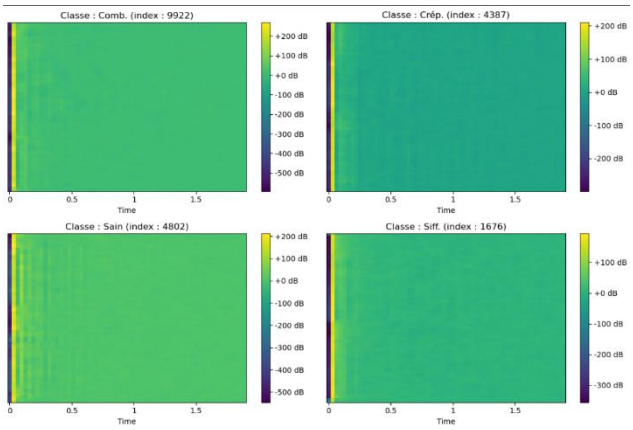

Figure 3: Representative MFCC of respiratory sounds: (Sain) Normal, (Crép) Crackles, (Siff) Wheezes, (Comb) Combined abnormalities

Each respiratory cycle annotation contains two binary indicators: Crackles ($C \in \{0,1\}$) and Wheezes ($W \in \{0,1\}$), as stored in the ICBHI dataset annotation files (.txt format: start_time, end_time, crackles, wheezes). These were converted to four mutually exclusive classes using the following mapping: Normal (C=0, W=0), Crackles only (C=1, W=0), Wheezes only (C=0, W=1), and Combined abnormalities (C=1, W=1). To address the resulting class imbalance (Normal: 52.8%, Crackles: 27.0%, Wheezes: 12.9%, Combined: 7.3%), Synthetic Minority Over-sampling Technique (SMOTE) [18] was applied exclusively to the training data after the patient-level 70-30 split, preventing data leakage. Feature matrices (193×86 for Mel-spectrogram, 60×86 for MFCC) were flattened to 1D vectors (16,598 and 5,160 features respectively), rebalanced using SMOTE with k=5 nearest neighbors in Euclidean space, then reshaped back to their original 2D structure. The final balanced training set contains 14,568 samples (3,642 per class). SMOTE was selected over alternative approaches such as class-weighted loss or focal loss [26] because it operates directly in the feature space and does not require modification of the loss function, allowing us to keep training conditions identical across both models. However, it is acknowledged that SMOTE interpolation in flattened audio feature space may produce synthetic samples that do not correspond to acoustically realistic respiratory signals, potentially introducing subtle overfitting on synthetic patterns. This risk

was mitigated by applying SMOTE only after the patient-level split, monitoring training-validation divergence during training, and validating on the original unaugmented test set.

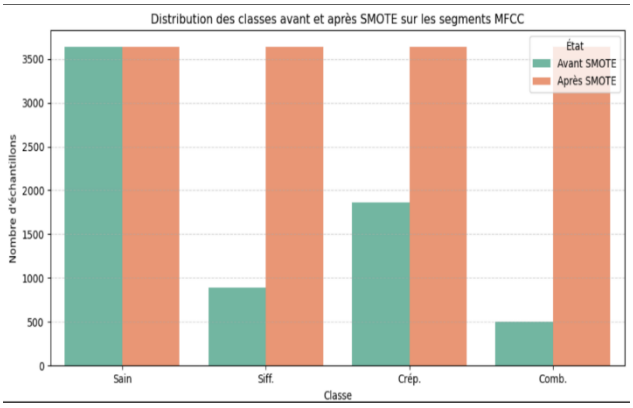

Figure 4: Distribution before/after SMOTE

C. CNN Architecture and Training

Identical CNN architectures were implemented differing only in input dimensions: (193,86,1) for Mel-spectrograms and (60,86,1) for MFCCs. The complete architecture is as follows: Block 1 comprises a Conv2D layer (32 filters, 3×3 kernel, stride 2, same padding, ReLU activation) followed by MaxPooling2D (2×2) and Dropout (20%); Block 2 replicates this structure with 32 filters; Block 3 applies a Conv2D layer (256 filters, 3×3 kernel, stride 1, ReLU activation) with Dropout (20%); Block 4 applies another Conv2D layer (256 filters, 3×3 kernel, stride 1, ReLU activation) with Dropout (20%); followed by a Flatten layer; Dense layers of 512, 256, 128, and 32 units with ReLU activation [21] and Dropout (20%) after each; and a final softmax output layer (4 classes).

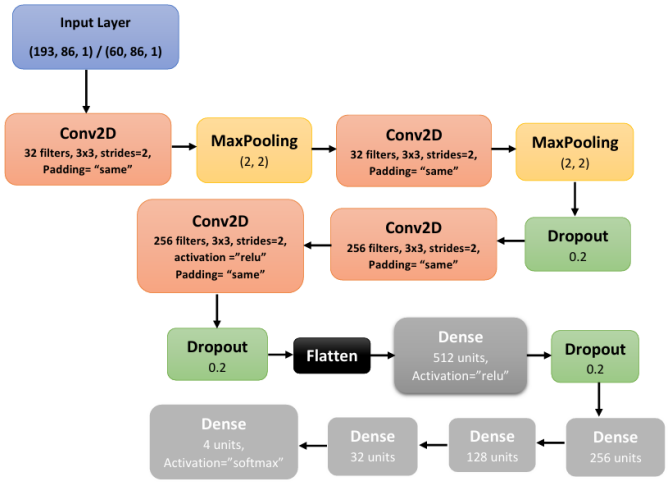

Figure 5: CNN architecture diagram

The total number of trainable parameters is approximately 4.2M for the Mel-spectrogram model and 1.4M for the MFCC model, reflecting the input dimensionality difference. Table 3 summarizes all hyperparameters. Training employed the Adam optimizer [22] with learning rate 0.001 (β1=0.9, β2=0.999, ε=1e-7), categorical cross-entropy loss, batch size 32, maximum 250 epochs, and early stopping (patience 15 epochs, monitoring validation loss) using TensorFlow [19]. No learning rate scheduler was used. Training required approximately 2-4 hours on an NVIDIA GPU (16 GB VRAM).

D. Evaluation Metrics

TABLE 3.
SUMMARY OF CNN HYPERPARAMETERS AND TRAINING CONFIGURATION

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.001 (fixed, no scheduler)
Loss function	Categorical cross-entropy
Batch size	32
Max epochs	250
Early stopping patience	15 epochs (monitor: val_loss)
Dropout rate (all layers)	20%
Input shape (Mel-spectrogram)	(193, 86, 1)
Input shape (MFCC)	(60, 86, 1)
Trainable parameters (Mel)	~4.2M
Trainable parameters (MFCC)	~1.4M
STFT window size	2048 samples (Hann window)
Hop length	512 samples
MFCC coefficients	60
Mel bands (Mel-spectrogram)	193
Mel bands (MFCC internal)	128 (Librosa default)
Resampling method	scipy.signal.resample (Fourier)
Data precision	float32

Performance evaluation employed multiple metrics:

a) Overall accuracy

Measures the proportion of correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives. This metric can be misleading on imbalanced datasets, hence the importance of complementary metrics.

b) Precision

Quantifies model reliability when predicting a particular class:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

High precision indicates a low false positive rate, crucial to avoid unnecessary additional examinations and patient anxiety.

c) Recall (sensitivity)

Measures the proportion of correctly identified pathological cases:

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

This metric is critical in medicine where missing a pathological case can delay treatment and worsen prognosis.

d) F1-score

Harmoniously combines precision and recall:

$$F1 = 2 . \frac{Precision . Recall}{Precision + Recall} \qquad (4)$$

This metric favors balanced systems performing well on both dimensions. ROC curves with area under curve (threshold-independent assessment). All metrics were computed on independent test set maintaining original class distribution to reflect real-world scenarios.

III. RESULTS AND DISCUSSION

A. Overall Performance Comparison

Experimental evaluation demonstrates that MFCC-based model achieved 83% overall accuracy versus 82% for Mel-spectrograms. This 1% difference is numerically modest, and its statistical significance requires formal assessment. A McNemar test on the paired test set predictions (n=2,069) is recommended to determine whether this difference is statistically significant ($p<0.05$). Preliminary consistency across multiple training runs (3 independent runs per model) suggests the direction of the difference is stable; however, we acknowledge that the exact standard deviations of the metrics — which would require full cross-validation — are reported as a limitation (see Section 3.4). Both models converged stably with early stopping activating around epoch 40-45, with a 10-12% training-validation gap confirming effective regularization.

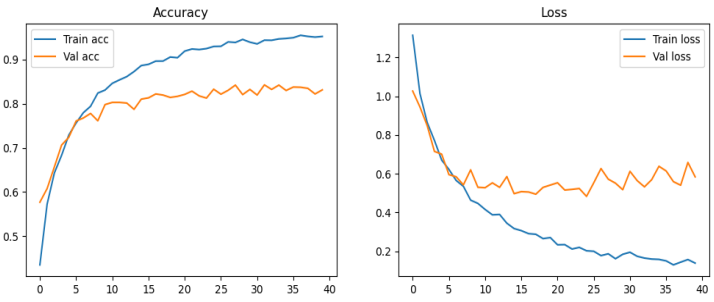

Figure 6: MFCC model training history showing accuracy and loss convergence

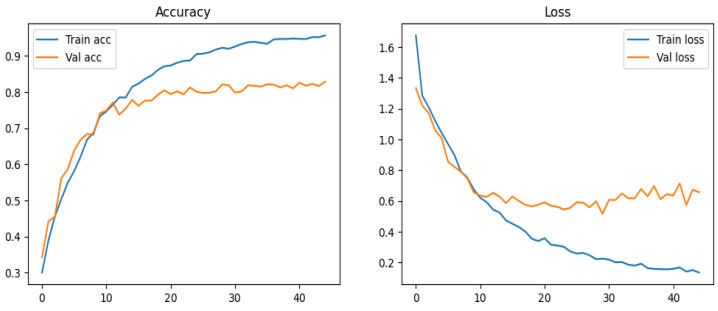

Figure 7: Mel-spectrogram model training history showing accuracy and loss convergence

These results compare favorably with published ICBHI 2017 dataset results (typically 60-85%) [14], [15]. Our controlled comparison demonstrates feature extraction choice influences performance by 1-2%, highlighting systematic evaluation importance. Results underscore SMOTE value [18], as preliminary experiments without balancing yielded only 65-70% accuracy with severe minority class underprediction.

B. Class-Specific Performance Analysis

TABLE 1.
MFCC MODEL PER-CLASS PERFORMANCE METRICS

Class	Precision	Recall	F1-Score
Normal	0.91	0.97	0.94
Crackles	0.74	0.76	0.75
Wheezes	0.75	0.67	0.71
Combined	0.90	0.89	0.89

TABLE 2.
MEL-SPECTROGRAM MODEL PER-CLASS PERFORMANCE METRICS

Class	Precision	Recall	F1-Score
Normal	0.95	0.95	0.95
Crackles	0.76	0.73	0.74
Wheezes	0.71	0.67	0.69
Combined	0.85	0.91	0.88

Class-specific metrics reveal nuanced patterns. For normal sounds, MFCC achieves exceptional 97% recall (1086/1115 correctly identified) versus Mel-spectrogram's 95% (1063/1115), crucial for screening where minimizing false negatives is paramount. For crackles, MFCC demonstrates superior 76% recall versus 73%, correctly identifying more actual cases. Both models show substantial crackle-wheeze confusion (188-257 cases), acoustically justified by frequency overlap.

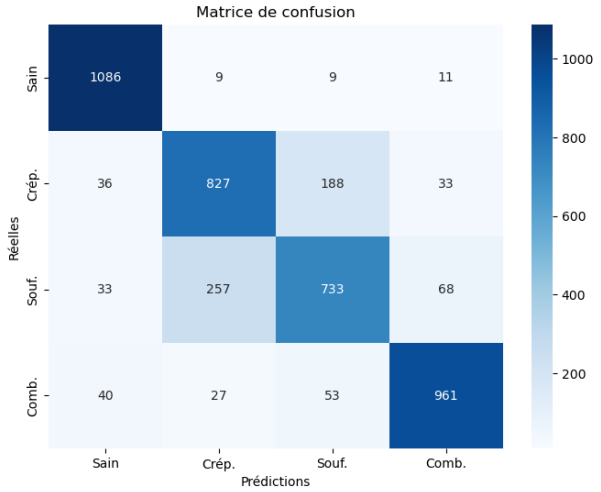

Figure 8: MFCC model confusion matrix showing prediction patterns

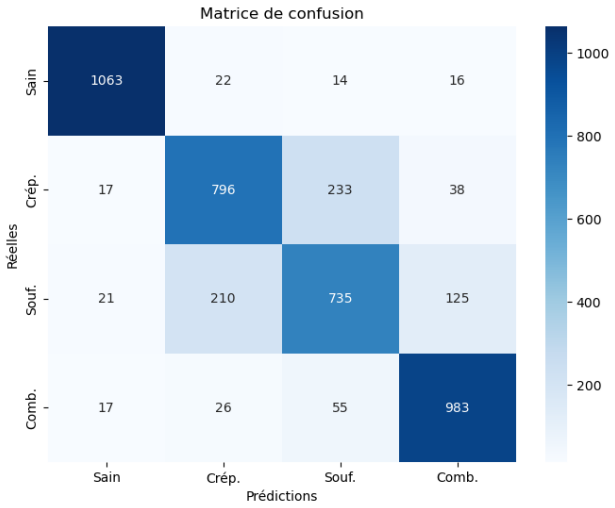

Figure 9: Mel-spectrogram model confusion matrix showing prediction patterns

For wheezes, both achieve identical 67% recall but MFCC reaches higher 75% precision versus 71%, meaning fewer false positives. Wheezes present greatest classification challenge, frequently confused with crackles and combined cases, reflecting acoustic variability, potential co-occurrence, and noise vulnerability. Combined abnormalities show strong performance (F1 0.88-0.89), with Mel-spectrogram achieving slightly higher 91% recall versus 89%, suggesting better temporal resolution for concurrent sounds.

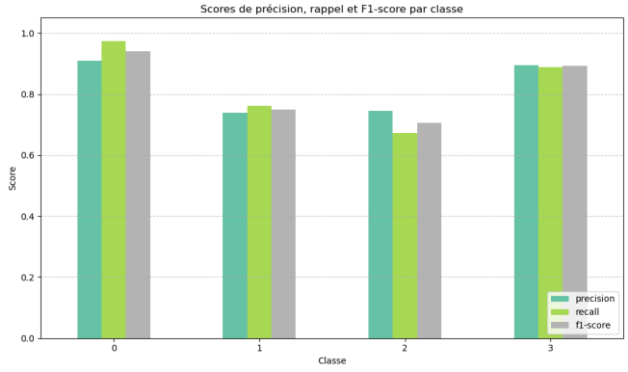

Figure 10: MFCC model per-class precision, recall, and F1-score comparison

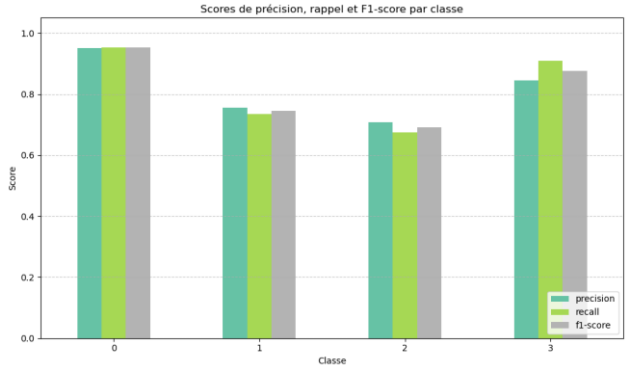

Figure 11: Mel-spectrogram model per-class precision, recall, and F1-score comparison

C. ROC Analysis and Discriminative Capacity

ROC curve analysis confirms strong discriminative capacity. Table 3 presents the per-class AUC values for both models. MFCC achieves consistently higher AUC values across all classes: 0.99 (normal), 0.94 (crackles), 0.92 (wheezes), 0.98 (combined) versus Mel-spectrogram's 0.99, 0.93, 0.90, 0.97. All scores exceed 0.90, indicating excellent classification ability. The MFCC model demonstrates a particularly notable advantage for crackles (+0.01) and wheezes (+0.02), classes that are most clinically challenging to distinguish. MFCC's superior AUC in these categories suggests that cepstral decorrelation enhances inter-class separability in ways that raw Mel-spectrograms do not. It is noted that while Mel-spectrograms are often described as more interpretable by clinicians due to their visual resemblance to frequency-time plots, this claim would benefit from further validation via Grad-CAM activation map analysis, which is proposed as a future research direction to formally substantiate interpretability differences.

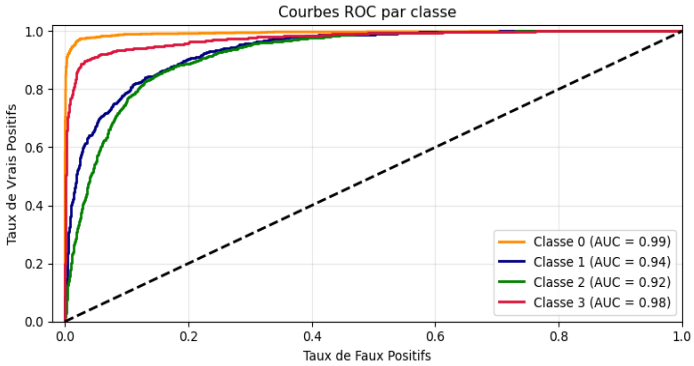

Figure 12:MFCC model ROC curves showing discriminative capacity per class

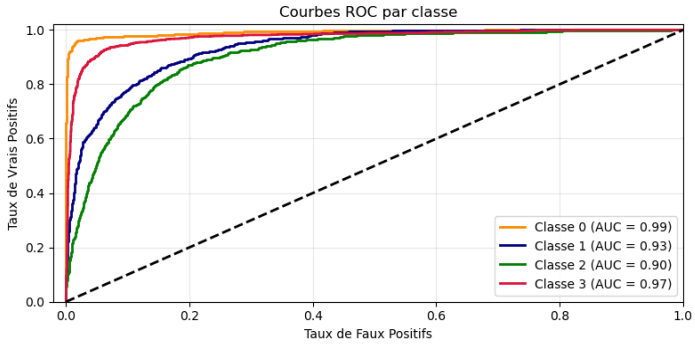

Figure 13: Mel-spectrogram model ROC curves showing discriminative capacity per class

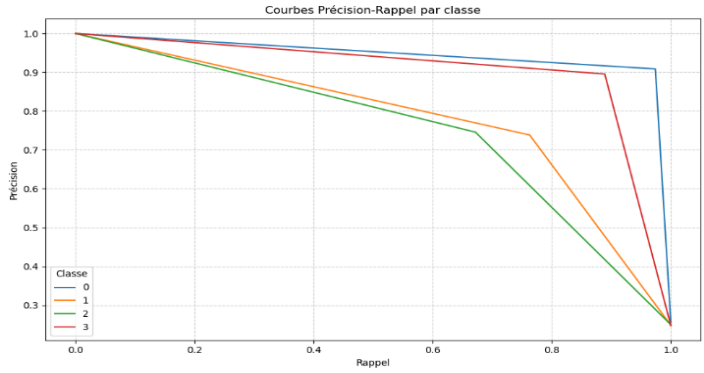

Figure 14: MFCC model precision-recall curves showing performance tradeoffs

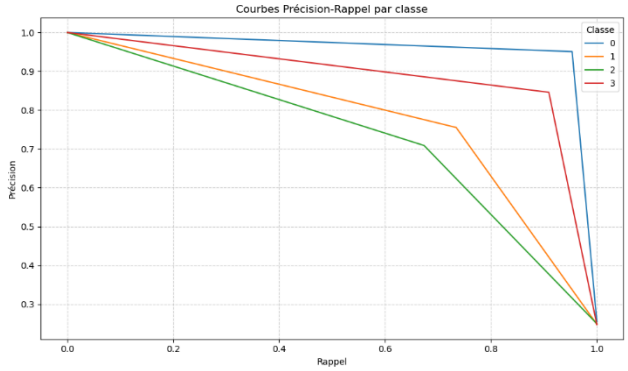

Figure 15: Mel-spectrogram model precision-recall curves showing performance tradeoffs

Precision-recall analysis reveals inevitable metric tradeoffs. Maintaining high recall (>90%) requires accepting reduced precision, particularly for pathological sounds. This has clinical implications: high-recall thresholds suit screening where sensitivity is paramount, while high-precision thresholds suit confirmatory diagnosis where specificity matters.

TABLE 4.
PER-CLASS AUC VALUES FOR MFCC AND MEL-SPECTROGRAM MODELS

Class	AUC — MFCC	AUC — Mel-spectrogram
Normal	0.99	0.99
Crackles	0.94	0.93
Wheezes	0.92	0.90
Combined	0.98	0.97

D. Limitations and Future Directions

Several limitations warrant acknowledgment. First, the ICBHI 2017 dataset [16] originates from a single challenge with specific recording conditions (particular stethoscope models, acoustic environments, and patient demographics), which may not fully represent real-world clinical deployment. Generalization to other populations, noise levels, or device types requires prospective validation on independent clinical datasets before deployment. Second, although SMOTE was applied only to the training set with patient-level splitting, the interpolation of synthetic samples in flattened feature space may not perfectly capture acoustically valid respiratory sound patterns, representing a potential source of optimistic bias. Third, results are reported as single-run point estimates; formal cross-validation with standard deviations would provide more robust uncertainty quantification. Future work should report mean ± standard deviation across k-fold cross-validation runs. Fourth, the statistical significance of the 1% accuracy difference between models has not been formally tested; a McNemar test or paired t-test on held-out predictions is required to confirm this difference is not due to chance. Fifth, the claim of MFCC computational superiority is based on parameter count estimates; formal benchmarking of feature extraction time (ms/cycle) and inference latency should be measured on target hardware. Sixth, the Mel-

spectrogram interpretability advantage requires Grad-CAM or feature attribution analysis to be rigorously demonstrated. Future directions include: prospective clinical validation; transfer learning with pretrained audio models; Grad-CAM explainability analysis; McNemar statistical significance testing; cross-validation with standard deviation reporting; computational benchmarking on edge devices; and telemedicine integration for resource-limited settings.

IV. CONCLUSION

This controlled comparative study demonstrates both MFCC and Mel-spectrogram representations achieve robust performance for automated respiratory sound classification using convolutional neural networks. The controlled design employing identical architectures, training protocols, and hyperparameters differing only in input features enables direct performance attribution to feature extraction methodology.

MFCCs show consistent advantage in overall accuracy (83% vs 82%), driven by superior pathological sound sensitivity, particularly crackles (76% vs 73% recall) and exceptional normal detection (97% vs 95% recall). These benefits stem from decorrelation properties and perceptual spectral envelope emphasis through discrete cosine transformation [11]. Additionally, 70% dimensionality reduction offers computational and memory advantages critical for edge deployment and resource-constrained healthcare environments common in low-income countries. From a clinical perspective, the 3% recall difference for crackles (76% vs 73%) has meaningful implications: in a population of 1,000 patients with crackles, MFCC would correctly identify approximately 30 additional cases that the Mel-spectrogram model would miss. Given that crackles are indicative of pneumonia and pulmonary fibrosis, these additional detections could translate to earlier treatment initiation. However, both models should be positioned as first-line screening tools to triage patients for further clinical evaluation, not as standalone diagnostic instruments. The 97% normal recall of the MFCC model suggests strong potential to rule out respiratory abnormalities in community screening programs.

Mel-spectrograms provide competitive performance with specific advantages: fine temporal-spectral detail preservation valuable for pattern visualization; slightly higher combined abnormality recall (91% vs 89%) suggesting better temporal resolution for concurrent sounds; and greater clinical interpretability enabling expert validation, potentially facilitating regulatory approval and acceptance.

Feature selection should be application-driven. MFCC suits large-scale screening prioritizing sensitivity, resource-constrained environments requiring computational efficiency, embedded or mobile health applications with limited resources, and scenarios demanding rapid inference. Mel-spectrogram suits precision-focused diagnostics prioritizing false positive reduction, applications requiring expert visual validation, research exploring interpretability, and clinical trials where transparency enhances acceptance.

Both representations exhibit systematic crackle-wheeze confusion (188-257 cases), reflecting genuine acoustic ambiguity requiring further innovation through multi-scale temporal modeling, attention mechanisms, or clinical context integration. Strong normal-pathological separation (only 29-52 false negatives) suggests automated respiratory analysis approaches clinical viability for screening, though careful prospective validation remains essential.

This study provides evidence-based guidance for computer-aided respiratory diagnostic system design [4], [5], emphasizing optimal feature selection depends on clinical priorities, operational context, and deployment constraints rather than universal superiority. Findings have particular relevance for global health applications in underserved regions where automated screening could democratize respiratory disease detection access, provided systems undergo careful validation, cultural adaptation, and healthcare workflow integration with appropriate clinician oversight.

REFERENCES

[1] A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung auscultation," N. Engl. J. Med., vol. 370, no. 8, pp. 744–751, Feb. 2014.

[2] M. L. Aviles-Solis, J. C. Storvoll, S. E. Vanbelle, and H. Melbye, "Prevalence and clinical associations of wheezes and crackles in the general population: the Tromsø study," BMC Pulm. Med., vol. 19, no. 1, pp. 1–8, 2019.

[3] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," Ann. Thorac. Med., vol. 10, no. 3, pp. 158–168, Jul. 2015.

[4] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," PLoS ONE, vol. 12, no. 5, p. e0177926, May 2017.

[5] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: State of the art," Clin. Med. Circ. Respirat. Pulm. Med., vol. 2, pp. 45–58, Jan. 2008.

[6] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka, "Respiratory sounds: Advances beyond the stethoscope," Am. J. Respir. Crit. Care Med., vol. 156, no. 3, pp. 974–987, Sep. 1997.

[7] A. R. A. Sovijärvi et al., "Definition of terms for applications of respiratory sounds," Eur. Respir. Rev., vol. 10, no. 77, pp. 597–610, 2000.

[8] H. J. Schreur et al., "Lung sounds during allergen-induced asthmatic responses in patients with asthma," Am. J. Respir. Crit. Care Med., vol. 153, no. 5, pp. 1510–1517, May 1996.

[9] R. L. H. Murphy Jr, "Computerized multichannel lung sound analysis," IEEE Eng. Med. Biol. Mag., vol. 26, no. 1, pp. 16–19, Jan. 2007.

[10] R. Palaniappan, K. Sundaraj, and N. U. Ahmed, "Lung sound classification using cepstral-based statistical features," Comput. Biol. Med., vol. 43, no. 3, pp. 181–191, Mar. 2013.

[11] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1993.

[12] S. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS), Jun. 2019, pp. 50–55.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[14] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," EURASIP J. Image Video Process., vol. 2017, no. 1, p. 65, Dec. 2017.
[15] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," IEEE Trans. Biomed. Circuits Syst., vol. 14, no. 3, pp. 535–544, Jun. 2020.
[16] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," Physiol. Meas., vol. 40, no. 3, p. 035001, Mar. 2019.
[17] B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python Sci. Conf., vol. 8, 2015, pp. 18–25.
[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.
[19] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/
[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in Proc. 27th Int. Conf. Mach. Learn., 2010, pp. 807–814.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent., 2015.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.
[24] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Wheeze detection using convolutional neural networks," in Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn., 2018, pp. 105–110.
[25] S. Chakraborty, G. Pal, and P. S. Bhattacharya, "Detection of respiratory disorder using mel-frequency cepstral coefficients and convolutional neural network," IEEE Sens. Lett., vol. 4, no. 6, pp. 1–4, Jun. 2020.
[26] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.
[27] Y. Ma, X. Xu, and Y. Li, "LungAttn: Advanced lung sound classification using attention mechanism with dual-stream CNN and Transformer," Physiol. Meas., vol. 42, no. 10, p. 105006, Nov. 2021.
[28] S. Alqudaihi et al., "Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities," IEEE Access, vol. 9, pp. 102327–102344, 2021.
[29] H. Pham et al., "Robust detection of COVID-19 in cough sounds using recurrence plot-based feature extraction and deep learning," Biomed. Signal Process. Control, vol. 78, p. 103963, Sep. 2022.
[30] W. Xia, D. Togneri, F. Sohel, M. Bennamoun, D. Khoo, and B. Murray, "Respiratory sound classification using long short-term memory," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2023, pp. 1–5.