JOURNAL AFRICAIN DES SCIENCES
VOL. 2, **2** (2025) 152-159

*https://jafrisci.org*

***ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930***

OPEN ACCESS

**ARTICLE DE RECHERCHE**

**Article Info.:**
Reçu : le 07/11/2025
Accepté : le 27/11/2025
Publié : le 01/12/2025

# DIAGNOSTIC ASSISTE DES SONS RESPIRATOIRES PAR FUSION MULTIMODALE MFCC-SPECTROGRAMMES AVEC RÉSEAUX DE NEURONES CONVOLUTIFS

**Tege Simboni Simboni** [1,2,4,*], **Fiston Oshasha Oshasha** [1,2,4], **Jordan Masakuna Felicien** [1], **Sylvestre Frey** [1], **Nestor Kibamba Anzola** [1], **André Musas-A-Musas** [1,2]

https://doi.org/10.70237/jafrisci.2025.v2.i2.12

**Resumé**

L'auscultation pulmonaire, bien qu'étant un pilier du diagnostic médical depuis des siècles, souffre d'une variabilité inter-observateurs atteignant 40% même entre experts. Cette étude propose un système innovant d'aide au diagnostic basé sur l'intelligence artificielle pour classifier automatiquement les sons respiratoires en quatre catégories cliniquement pertinentes. Nous avons développé un modèle de réseau de neurones convolutionnels hybride à double entrée qui exploite simultanément deux représentations complémentaires du signal audio : les coefficients MFCC (60×86) capturant l'enveloppe spectrale, et les mél-spectrogrammes (193×86) préservant les détails temps-fréquence. Notre architecture parallèle traite ces modalités indépendamment avant de fusionner leurs caractéristiques via concaténation pour la classification finale. Les expériences menées sur la base ICBHI 2017 (6 898 cycles respiratoires provenant de 126 patients) incluent un prétraitement rigoureux et un équilibrage par SMOTE (14 568 échantillons finaux). Les résultats démontrent une précision globale de 81% avec des performances différenciées : sons sains (précision 90%, rappel 95%, AUC 0.99), crépitements (précision 69%, rappel 77%, F1 0.73), Sifflements (précision 76%, rappel 62%, F1 0.68), et anomalies combinées (précision 86%, rappel 89%, F1 0.88). Ce système pourrait être déployé sur dispositifs mobiles pour améliorer l'accessibilité diagnostique dans les zones sous-médicalisées.

**Mots clés** : CNN, MFCC, mél-spectrogramme, sons respiratoires, diagnostic assisté par IA

**Abstract**

Pulmonary auscultation, despite being a cornerstone of medical diagnosis for centuries, suffers from inter-observer variability reaching 40% even among experts. This study proposes an innovative AI-based diagnostic support system to automatically classify respiratory sounds into four clinically relevant categories. We developed a dual-input hybrid convolutional neural network model that simultaneously exploits two complementary audio signal representations: MFCC coefficients (60×86) capturing spectral envelope, and mel-spectrograms (193×86) preserving time-frequency details. Our parallel architecture processes these modalities independently before fusing their features via concatenation for final classification. Experiments conducted on the ICBHI 2017 database (6,898 respiratory cycles from 126 patients) include rigorous preprocessing and SMOTE balancing (14,568 final samples). Results demonstrate an overall accuracy of 81% with differentiated performances: healthy sounds (precision 90%, recall 95%, AUC 0.99), crackles (precision 69%, recall 77%, F1 0.73), wheezes (precision 76%, recall 62%, F1 0.68), and combined anomalies (precision 86%, recall 89%, F1 0.88). This system could be deployed on mobile devices to improve diagnostic accessibility in underserved areas.

**Key words :** CNN, MFCC, mel-spectrogram, respiratory sounds, AI-assisted diagnosis

## 1. INTRODUCTION

Les maladies respiratoires représentent un fardeau sanitaire majeur à l'échelle mondiale, causant environ 4 millions de décès annuels selon l'Organisation Mondiale de la Santé [1]. La pneumonie, la bronchopneumopathie chronique obstructive (BPCO), l'asthme et les autres affections pulmonaires touchent disproportionnellement les populations vulnérables, particulièrement dans les pays en développement où l'accès aux soins spécialisés demeure limité. Dans ce contexte, l'auscultation pulmonaire constitue souvent le premier et parfois le seul outil diagnostique disponible pour les professionnels de santé.

***Correspondance :*** *tege.simboni1@gmail.com* ***(T. S. Simboni)***



[1] *Mention Mathématiques, Statistiques et Informatique, Faculté des sciences et Technologie, Université de Kinshasa, Kinshasa, R.D. Congo*
[2] *Département de Gestion Informatique, Institut Supérieur Pédagogique d'Isiro, Isiro, D.R. Congo*
[3] *Commissariat Général à l'Énergie Atomique, Centre Régional d'Études Nucléaires de Kinshasa, B.P. 868, Campus de l'Université de Kinshasa, R.D. Congo*
[4] *CRIA – Centre de Recherche en Informatique Appliquée, Kinshasa, R.D. Congo*

Pourtant, cette technique ancestrale présente des limitations bien documentées dans la littérature médicale. Les études cliniques ont révélé des taux de désaccord atteignant 40% entre pneumologues expérimentés lors de l'interprétation des mêmes enregistrements sonores [2]. Cette variabilité s'explique par plusieurs facteurs : la nature subjective de la perception auditive humaine, la variabilité de l'expérience clinique, les différences dans les conditions d'auscultation, et l'influence des bruits parasites environnementaux.

Heureusement, l'avènement des stéthoscopes électroniques, capables d'enregistrer et d'amplifier les sons respiratoires avec une fidélité supérieure aux stéthoscopes acoustiques traditionnels, ouvre de nouvelles perspectives. Couplés à des algorithmes d'intelligence artificielle, ils permettent d'envisager une auscultation objective, reproductible et accessible même en l'absence de spécialistes. C'est précisément dans cette direction que s'inscrit notre recherche.

Les sons respiratoires normaux présentent un spectre fréquentiel large et régulier. En revanche, les pathologies génèrent des signatures acoustiques distinctives [14]. Les crépitements (crackles), qui sont des sons très brefs durant typiquement 5 à 20 millisecondes, résultent de l'ouverture brusque de petites voies aériennes. Les sifflements (wheezes), qui sont des sons plus continus dépassant généralement 250 millisecondes, témoignent d'un rétrécissement bronchique. La présence simultanée de ces deux types d'anomalies signale généralement une pathologie complexe.

Pour l'analyse computationnelle de ces signaux audio médicaux, deux approches dominent actuellement la littérature. Les coefficients MFCC (Mel-Frequency Cepstral Coefficients) [3] fournissent une représentation compacte du spectre sonore en quelques dizaines de valeurs. Les mél-spectrogrammes [13] offrent une visualisation bidimensionnelle temps-fréquence conservant la dynamique temporelle fine du signal. Par ailleurs, les réseaux de neurones convolutionnels (CNN) ont révolutionné le traitement des signaux biomédicaux [4], mais force est de constater que la plupart des études publiées se limitent encore à l'utilisation d'une seule de ces modalités.

Notre hypothèse repose sur la complémentarité de ces deux représentations. Les MFCC excellent dans la caractérisation des sons quasi-stationnaires et continus comme les sifflements, tandis que les mél-spectrogrammes préservent les informations temporelles nécessaires à la détection des événements transitoires brefs comme les crépitements. Un modèle hybride capable d'exploiter simultanément et intelligemment ces deux sources d'information devrait donc logiquement offrir une robustesse et une performance supérieures à celles des approches uni-modales. C'est cette intuition que nous cherchons à valider dans cette étude.

Sur le plan théorique, ce travail contribue à clarifier le rôle et la complémentarité de deux représentations acoustiques majeures les MFCC et les mél-spectrogrammes dans un cadre de classification automatique des sons respiratoires. Alors que ces descripteurs ont souvent été utilisés séparément dans la littérature, nous montrons qu'une architecture bimodale, dans laquelle chaque modalité est traitée par une branche convolutionnelle dédiée avant une fusion tardive, permet de mieux exploiter leurs caractéristiques respectives pour distinguer les crépitements, les sifflements, les sons sains et les combinaisons pathologiques. Cette approche met en évidence l'intérêt d'une fusion explicite des flux de caractéristiques plutôt que d'un simple empilement de descripteurs, et propose un schéma générique de réseau multimodal pouvant être réutilisé pour d'autres tâches de diagnostic audio en médecine.

## 2. PROBLEMATIQUE

Plusieurs travaux récents ont exploré l'apprentissage profond pour l'analyse automatique des sons respiratoires, chacun apportant des contributions importantes mais révélant également des limites. Comprendre ces approches permet de situer la portée de notre contribution et d'identifier les défis encore ouverts.

Perna et Tagarelli (2019) [9] ont montré, avec leur approche « Deep Auscultation », que les CNN peuvent apprendre des représentations hiérarchiques pertinentes à partir de spectrogrammes. Toutefois, leur modèle repose sur une seule modalité acoustique. De leur côté, Aykanat et al. (2017) [10] ont utilisé un CNN spécialisé sur spectrogrammes, atteignant 76% de précision sur ICBHI 2017 et confirmant le potentiel des architectures convolutionnelles. Serbes et al. (2018) [16], quant à eux, ont proposé une classification hiérarchique combinant caractéristiques temporelles et spectrales, avec des résultats encourageants pour l'identification des anomalies respiratoires.

Les travaux de Palaniappan et al. (2013) [17] ont posé les bases méthodologiques du domaine en explorant diverses techniques de machine learning classiques. Chambers et al. (2018) [18] ont ensuite abordé la problématique du bruit ambiant avec une approche de dictionnaire couplée à un CNN, montrant une meilleure robustesse dans des conditions d'enregistrement réelles. Plus récemment, Ma et al. (2019) [22] ont proposé LungBRN, une architecture multimodale obtenant 79% de précision mais nécessitant des ressources computationnelles importantes. Enfin, Reichert et al. (2008) [15] ont fourni un état de l'art fondamental sur l'analyse spectrale des sons respiratoires, à l'origine de nombreuses approches modernes.

Face à ces avancées, une question demeure : comment optimiser la classification automatique des sons respiratoires en combinant plusieurs représentations acoustiques, tout en gardant un modèle exploitable en pratique clinique **?**

Notre travail apporte une réponse en s'appuyant sur trois éléments clés : (i) une fusion multimodale tardive des MFCC et mél-spectrogrammes pour exploiter leur forte complémentarité, (ii) une architecture CNN hybride à double branche permettant à chaque modalité d'être traitée de manière spécialisée avant la fusion, et (iii) l'utilisation rigoureuse de SMOTE pour contrer le déséquilibre naturel des classes. L'objectif est de tirer parti simultanément des caractéristiques spectrales globales et des détails temporels fins, tout en maintenant une complexité compatible avec un déploiement pratique.

## 3. MODELES

### 3.1. Vue d'ensemble du système

Notre système suit un pipeline complet allant de l'acquisition

brute des sons respiratoires jusqu'à la classification finale en quatre catégories cliniques. La Figure 1 présente schématiquement ce pipeline dans son ensemble. Le processus débute par la collecte des enregistrements audio au format numérique, suivie d'une phase de prétraitement essentielle pour normaliser et nettoyer les signaux. Vient ensuite l'extraction des caractéristiques acoustiques selon deux modalités parallèles (MFCC et mél-spectrogrammes), puis l'équilibrage du jeu de données par génération d'exemples synthétiques. Enfin, le modèle hybride CNN traite ces données préparées pour produire la prédiction diagnostique.

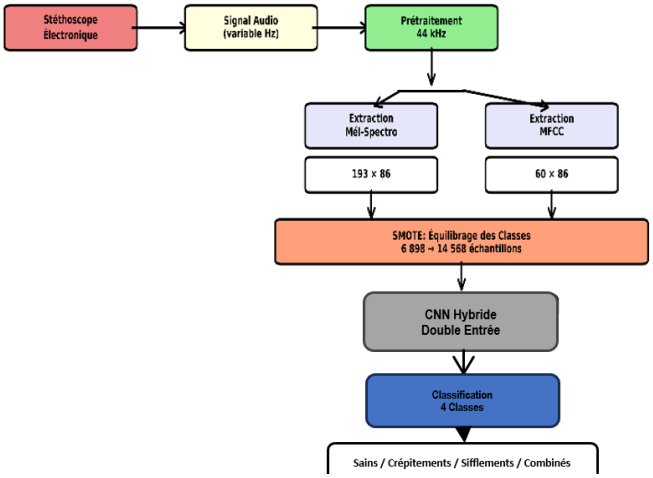


***Figure 1:*** *Pipeline système*

### 3.2. Base de données ICBHI 2017

Nos expériences reposent sur la base de données ICBHI 2017 [5], constituée spécifiquement pour l'évaluation standardisée des algorithmes de classification de sons respiratoires. Cette base de données résulte d'une collaboration internationale impliquant des équipes médicales pluridisciplinaires de quatre pays européens dont la Grèce, le Portugal, l'Italie et l'Allemagne. Les enregistrements ont été réalisés selon un protocole rigoureux dans des environnements cliniques contrôlés.

La base comprend 920 fichiers audio provenant de 126 patients, incluant des individus en bonne santé respiratoire et des patients présentant diverses pathologies respiratoires comme pneumonie, BPCO, asthme, etc. Tous les sons ont été capturés à l'aide d'un stéthoscope électronique [21,22] 3M Littmann 3200, dispositif médical certifié offrant une réponse en fréquence étendue (20 Hz à 2 kHz).

Après segmentation manuelle minutieuse par des experts médicaux, la base contient au final 6 898 cycles respiratoires individuels annotés. La distribution de ces cycles révèle un déséquilibre marqué mais attendu : 3 642 cycles sains (52,8% du total), 1864 cycles avec crépitements uniquement (27,0%), 886 cycles avec sifflements uniquement (12,8%), et enfin 506 cycles présentant une combinaison des deux anomalies (7,3%). Ce déséquilibre naturel reflète fidèlement la réalité de la distribution clinique des pathologies, mais pose inévitablement un défi algorithmique important que nous devrons gérer spécifiquement.

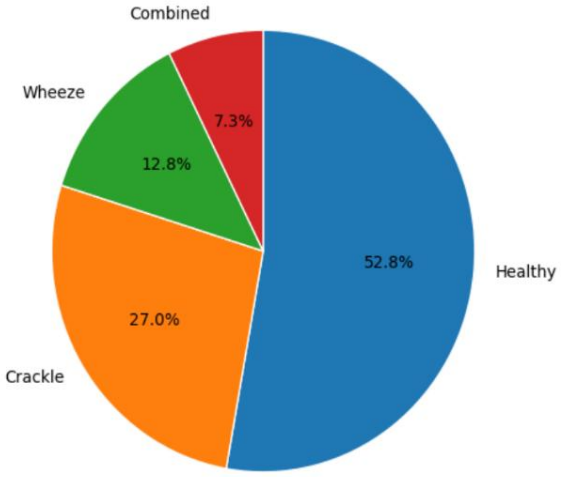


***Figure 2:*** *Distribution finale*

### 3.3. Prétraitement des signaux

Le prétraitement des signaux constitue une étape cruciale, car la qualité des données d'entrée conditionne directement les performances finales du modèle. La première opération consiste en une normalisation des taux d'échantillonnage à 44 kHz, offrant un bon compromis entre résolution fréquentielle et charge computationnelle. Cette normalisation utilise l'algorithme de rééchantillonnage polyphase implémenté dans Librosa [6], qui préserve bien la qualité du signal.

Ensuite, chaque cycle respiratoire est extrait individuellement et rééchantillonné à exactement 44 000 échantillons temporels, correspondant précisément à 1 seconde de signal. Pour atténuer les discontinuités, nous appliquons une fenêtre de Hann aux extrémités. Enfin, chaque segment est normalisé en amplitude pour une comparaison équitable.

### 3.4. Extraction des caractéristiques acoustiques

#### *3.4.1. Mél-spectrogrammes.*

Le mél-spectrogramme [13] représente une visualisation bidimensionnelle de l'énergie du signal dans un espace-temps-fréquence. Sa construction repose sur une transformation de Fourier à court terme (STFT) qui décompose le signal en petites fenêtres temporelles successives, suivie d'une conversion en échelle de Mel :

$$MelSpec_{dB}(t,m) = 10.\log(S_{mel}(t,m) + \epsilon) \quad (1)$$

Nous utilisons 193 bandes de Mel avec une fenêtre FFT de 2048 points et un hop length de 512 échantillons, générant 86 trames temporelles. Les valeurs sont converties en échelle logarithmique (décibels). La matrice résultante de dimensions 193 × 86 préserve les variations spectrales et temporelles du signal respiratoire.

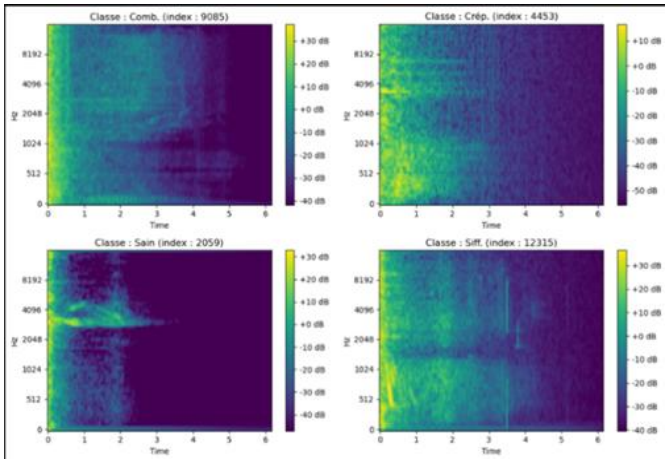


***Figure 3:*** *Représentation Mel-spectogramme des sons par classe : Comb, Sain, Crép et Siff*

### *3.4.2. Coefficients MFCC.*

Les coefficients MFCC offrent une alternative compacte au mél-spectrogramme. Leur calcul ajoute une transformée en cosinus discrète (DCT) qui décorrèle les valeurs d'énergie entre bandes de Mel voisines, permettant ainsi une compression efficace de l'information :

$$MFCC(t,k) = \sum_{m=1}^{M} \log\left(E_m(t)\right).cos\left[\frac{\pi k}{M}(m-0.5)\right] \quad (2)$$

Dans notre implémentation, nous extrayons 60 coefficients MFCC distincts pour chaque trame temporelle, produisant une matrice finale de dimensions 60 × 86. Les MFCC excellent dans la caractérisation des sons quasi-stationnaires comme les sifflements continus.

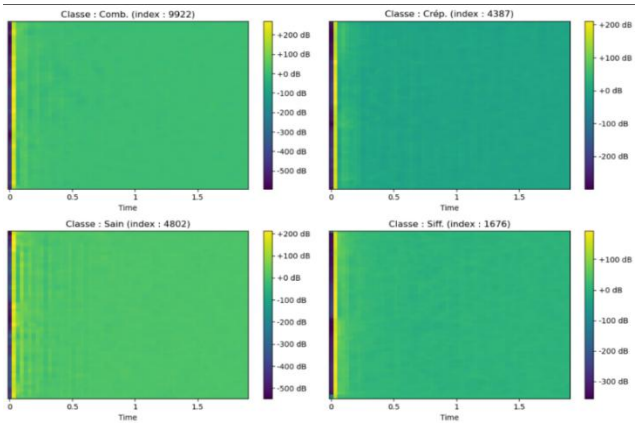


***Figure 4:*** *Représentation MFCC des sons par classe : Comb, Sain, Crép et Siff*

### 3.5. Équilibrage par SMOTE

Pour contrer efficacement le déséquilibre important des classes présent dans notre jeu de données initial, nous avons choisi d'implémenter la technique SMOTE (Synthetic Minority Over-sampling Technique) [7]. Cette approche intelligente est beaucoup plus sophistiquée qu'une simple duplication naïve des exemples minoritaires, car elle génère de nouveaux exemples véritablement synthétiques mais plausibles par interpolation géométrique dans l'espace des caractéristiques.

Le principe de fonctionnement est le suivant : pour chaque échantillon appartenant à une classe minoritaire, l'algorithme identifie intelligemment ses k plus proches voisins dans l'espace des caractéristiques (typiquement k=5). Il crée ensuite un nouvel exemple synthétique en interpolant linéairement entre l'échantillon original et un de ses voisins choisi aléatoirement :

$$x_{synth} = x_i + \lambda \times (x_{voisin} - x_i), \quad \lambda \sim U(0, 1) \quad (3)$$

où λ est un coefficient aléatoire tiré uniformément entre 0 et 1, garantissant que le nouvel exemple se situe quelque part sur le segment reliant les deux exemples originaux. Cette stratégie permet de générer des variations réalistes tout en restant dans des régions plausibles de l'espace des caractéristiques.

L'application systématique de SMOTE a porté notre jeu de données de 6 898 échantillons originaux à 14 568 échantillons finaux parfaitement équilibrés, avec exactement 3 642 exemples dans chacune des quatre classes. Cet équilibrage rigoureux évite que le modèle ne soit biaisé vers la prédiction systématique des classes majoritaires, garantissant ainsi un apprentissage équitable de toutes les catégories diagnostiques.

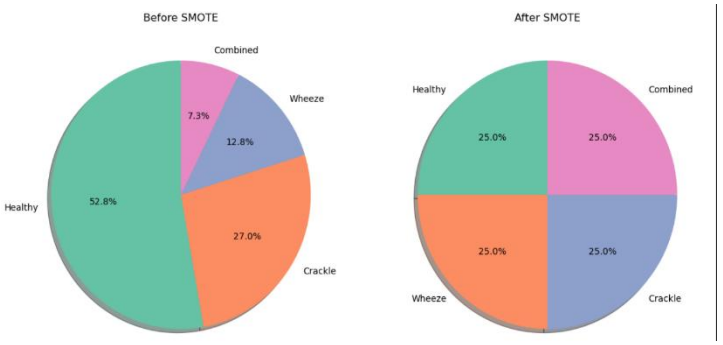


***Figure 5:*** *Distribution des classes Avant/Après SMOTE*

### 3.6. Architecture du réseau hybride

Notre architecture neuronale repose sur un principe élégant de fusion multimodale tardive : deux branches CNN parallèles et identiques dans leur structure traitent de manière totalement indépendante les deux modalités d'entrée (MFCC et mél-spectrogrammes) avant de combiner leurs sorties respectives au niveau de la couche de décision finale. Ce design permet à chaque branche de se spécialiser optimalement dans l'extraction des caractéristiques pertinentes de sa modalité spécifique, avant que l'information complémentaire ne soit fusionnée pour la classification.

Chaque branche CNN suit une architecture progressive classique mais éprouvée : elle débute par plusieurs couches convolutionnelles successives (avec des filtres de taille croissante 32, 64, 128) qui extraient hiérarchiquement des caractéristiques de plus en plus abstraites et complexes. Chaque convolution est suivie d'une fonction d'activation ReLU qui introduit la non-linéarité nécessaire, puis d'une opération de max-pooling qui réduit progressivement la dimensionnalité spatiale tout en conservant les caractéristiques les plus saillantes. Cette réduction dimensionnelle contrôlée permet de construire des représentations de plus en plus compactes et invariantes.

Après avoir extrait ces représentations riches mais compactes, chaque branche aplatit sa carte de caractéristiques en un vecteur unidimensionnel qui est ensuite passé à travers une ou deux couches denses (fully-connected) avec dropout pour prévenir le surapprentissage. C'est à ce niveau que les sorties des deux branches sont finalement concaténées en un unique vecteur qui combine l'information des deux modalités. Ce vecteur fusionné passe enfin à travers une dernière couche dense avec activation softmax qui produit les probabilités finales d'appartenance aux quatre classes diagnostiques. La Figure 6 illustre cette architecture détaillée.

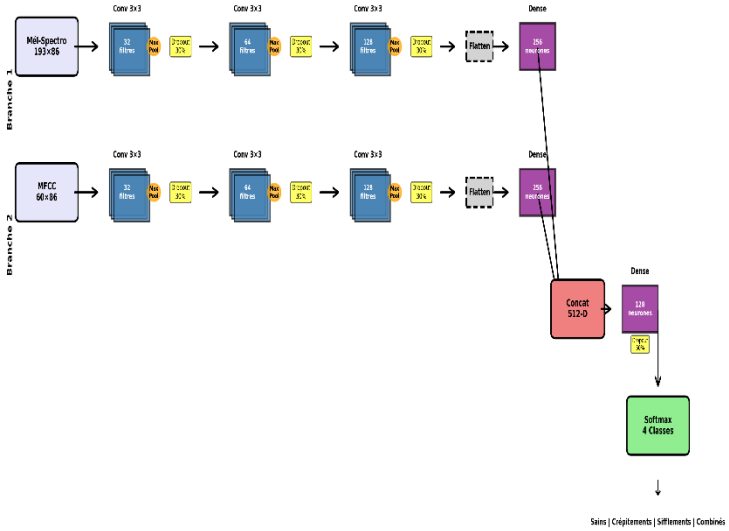


***Figure 6:*** *Architecture du CNN Hybride*

### 3.7. Protocole d'entraînement

L'entraînement de notre modèle utilise la fonction de perte d'entropie croisée catégorielle, qui est particulièrement bien adaptée aux problèmes de classification multi-classes. Pour l'optimisation des poids, nous employons l'algorithme Adam [8] avec un taux d'apprentissage initial de 0,001, qui s'est révélé être un bon compromis entre vitesse de convergence et stabilité de l'apprentissage.

Les données équilibrées sont divisées selon une stratégie classique : 70% sont réservés pour l'entraînement proprement dit (soit environ 10 200 échantillons), tandis que les 30% restants (environ 4 370 échantillons) constituent l'ensemble de test indépendant qui servira à évaluer objectivement les performances finales. Pour éviter le surapprentissage, nous avons implémenté un mécanisme d'arrêt précoce (early stopping) qui surveille continuellement la perte de validation et interrompt automatiquement l'entraînement s'il n'observe aucunes améliorations durant 20 époques consécutives.

Dans notre cas particulier, l'entraînement a naturellement convergé après environ 50 époques, moment où le mécanisme d'arrêt précoce s'est déclenché, indiquant que le modèle avait atteint ses meilleures performances possibles sans surapprentissage significatif. Cette convergence relativement rapide témoigne de l'efficacité de notre architecture et de la qualité du prétraitement des données.

## 4. RESULTATS

### 4.1. Évolution de l'entraînement

La Figure 7 illustre de manière visuelle comment notre modèle a progressivement appris au fil des époques successives d'entraînement. On peut observer une convergence assez rapide pendant les 20 premières époques, période durant laquelle la précision grimpe rapidement jusqu'à atteindre environ 95% sur les données d'entraînement. Parallèlement, la précision sur l'ensemble de validation se stabilise autour de 81%, ce qui représente un bon équilibre indiquant que le modèle généralise correctement sans surapprentissage excessif. L'écart modéré entre les performances sur l'entraînement et la validation suggère que notre modèle a bien appris les patterns généraux plutôt que de simplement mémoriser les exemples d'entraînement.

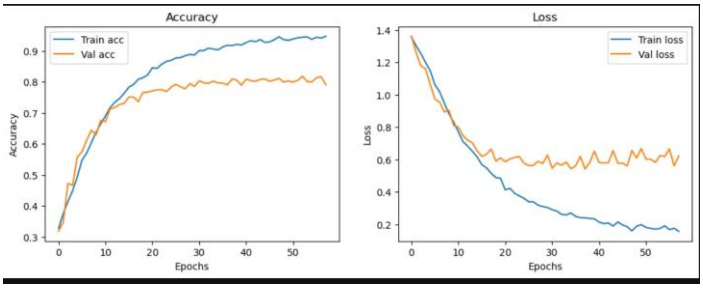


***Figure 7:Evolution de la précision et de la perte***

### 4.2. Performances globales

Sur l'ensemble de test indépendant, notre modèle hybride atteint une précision globale très satisfaisante de 81%. Le Tableau 1 présente en détail les résultats pour chacune des quatre classes, avec les métriques standards de précision, rappel et F1-score.

***Tableau 1:*** *Rapport détaillé de classification par classe*

| Classe | Précision | Rappel | F1 | Support |
|---|---|---|---|---|
| **Healthy** | 0.90 | 0.95 | 0.92 | 1115 |
| **Crackle** | 0.69 | 0.77 | 0.73 | 1084 |
| **Wheeze** | 0.76 | 0.62 | 0.68 | 1091 |
| **Combined** | 0.86 | 0.89 | 0.88 | 1081 |
| | | | | |
| **Accuracy** | | | 0.81 | 4371 |
| **Macro avg** | 0.81 | 0.81 | 0.80 | 4371 |

Les sons sains affichent d'excellentes performances avec une précision de 90% et un rappel impressionnant de 95%, ce qui signifie que le modèle est très bon pour identifier correctement les patients en bonne santé respiratoire. Les sons combinés (présentant à la fois crépitements et sifflements) obtiennent aussi de très bons résultats avec un F1-score de 0.88, indiquant que ces cas pathologiques complexes sont bien détectés.

En revanche, les sifflements isolés présentent le rappel le plus faible de toutes les classes (62%), ce qui indique une certaine difficulté du modèle à détecter systématiquement ce type d'anomalie. Les crépitements isolés se situent dans une position intermédiaire avec un F1-score honorable de 0.73. Ces différences de performance entre classes reflètent probablement la plus grande variabilité acoustique de certains types de sons pathologiques.

### 4.3. Analyse de la matrice de confusion

La Figure 8 présente la matrice de confusion complète qui révèle de manière détaillée les patterns d'erreurs commises par notre modèle. Cette visualisation nous permet de comprendre précisément quels types de confusions le système tend à faire.

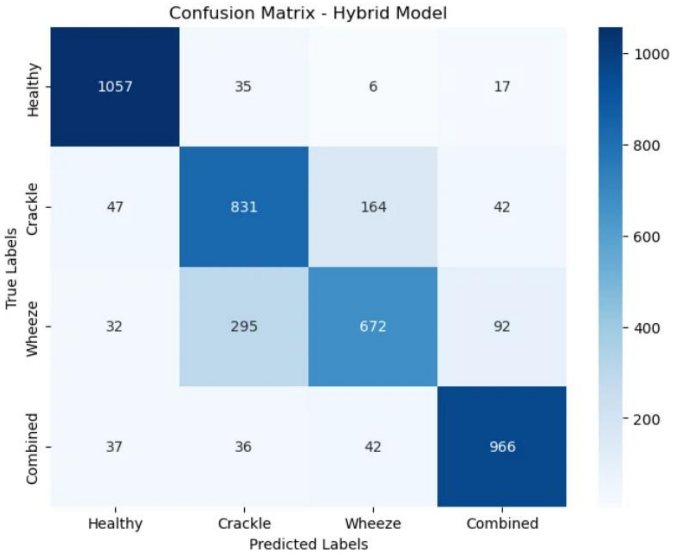


***Figure 8:*** *Matrice de confusion sur l'ensemble de test*

On constate que les sons sains (1057 correctement classifiés sur 1115 au total) et les sons combinés (966 correctement classifiés sur 1081) sont remarquablement bien identifiés par le modèle. La confusion principale et la plus problématique concerne les crépitements et les sifflements entre eux : environ 295 exemples de sifflements sont incorrectement classifiés comme crépitements (représentant 27% des sifflements), ce qui reflète probablement une certaine similarité acoustique naturelle entre ces deux types d'anomalies. Cette confusion bidirectionnelle suggère que les frontières acoustiques entre crépitements et sifflements ne sont pas toujours parfaitement nettes dans les signaux réels, et que même pour un modèle d'apprentissage profond, la distinction peut s'avérer délicate dans certains cas ambigus.

### 4.4. Courbes ROC

Les courbes ROC (Receiver Operating Characteristic) présentées en Figure 9 permettent d'évaluer de manière globale la capacité discriminative de notre modèle pour chacune des quatre classes, indépendamment du seuil de décision choisi.

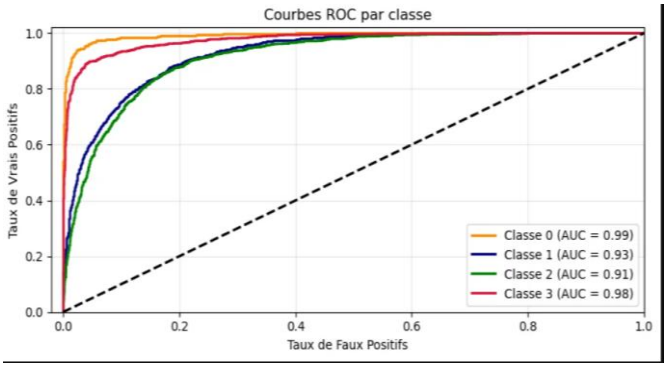


***Figure 9: Courbes ROC par classe***

Les aires sous la courbe (AUC) obtenues sont toutes très élevées et cliniquement prometteuses : sons sains (0.99), sons combinés (0.98), crépitements (0.93), et sifflements (0.91). Ces valeurs remarquablement élevées, toutes supérieures à 0.90, confirment objectivement la bonne capacité générale de discrimination du modèle. Même la classe la plus difficile (sifflements) maintient une AUC excellente de 0.91, ce qui est bien au-dessus du seuil cliniquement acceptable. Ces résultats suggèrent que malgré les confusions ponctuelles observées dans la matrice de confusion, le modèle possède intrinsèquement une bonne capacité à séparer les différentes classes lorsqu'on considère l'ensemble des seuils de décision possibles.

## 5. DISCUSSION

Nos résultats montrent que la combinaison de MFCC et de mél-spectrogrammes fonctionne réellement bien, avec une précision globale de 81%, ce qui représente une performance très honorable pour une tâche médicale aussi délicate. Ce n'est pas surprenant : les deux représentations sont naturellement complémentaires. Les mél-spectrogrammes capturent mieux les événements courts comme les crépitements, tandis que les MFCC décrivent plus efficacement les sons continus tels que les sifflements. En laissant chaque branche du réseau se spécialiser avant fusion, on tire le meilleur de chaque modalité.

Il est cependant important de reconnaître certaines limites. L'utilisation de SMOTE [7] pour équilibrer les classes, bien qu'efficace, introduit des exemples synthétiques qui ne reflètent pas toujours la complexité réelle des signaux cliniques. De plus, la base ICBHI 2017 [5], même si elle fait référence dans le domaine, reste relativement modeste et essentiellement européenne. Cela soulève naturellement la question de la généralisabilité du modèle à d'autres populations et contextes médicaux, en particulier dans des environnements africains, asiatiques ou américains où les profils pathologiques, les morphologies thoraciques, et les dispositifs d'enregistrement peuvent varier.

Un autre point à mentionner concerne notre protocole expérimental : les performances ont été évaluées à partir d'une seule division des données en entraînement et test. Par manque de temps et de ressources computationnelles, une validation croisée complète n'a pas pu être mise en place. Elle aurait pourtant permis d'obtenir une estimation encore plus robuste des performances. Cette limitation méthodologique sera traitée dans des travaux futurs, avec l'implémentation d'une validation croisée stratifiée et, idéalement, une validation externe sur d'autres bases de sons respiratoires.

Malgré ces limites, le système présente un vrai potentiel clinique, notamment pour du triage en première ligne dans les zones où les pneumologues sont rares. Avec 90% de précision sur les sons sains, il peut rassurer efficacement, et avec un F1-score de 0.88 sur les cas combinés, il identifie correctement les situations nécessitant une prise en charge spécialisée. Le modèle est léger et pourrait fonctionner en temps réel sur un simple smartphone [20], ouvrant la voie à des outils d'aide au diagnostic accessibles dans les milieux à faibles ressources.

Enfin, plusieurs pistes d'amélioration sont envisageables :

ajouter des mécanismes d'attention pour mieux localiser les parties pertinentes du signal, enrichir l'entraînement par des augmentations de données réalistes, ou intégrer des informations cliniques simples (âge, sexe, historique) pour contextualiser l'analyse. L'apprentissage semi-supervisé représente également une perspective prometteuse pour exploiter les nombreux enregistrements non annotés et réduire la dépendance à l'expertise humaine.

Lors de la comparaison avec les principales approches publiées dans la littérature, notre modèle atteint une précision globale de 81%, ce qui le positionne dans la moyenne haute des méthodes CNN appliquées à la base ICBHI. Les travaux d'Aykanat et al. [10], utilisant un CNN simple basé sur des spectrogrammes, rapportent 76%, tandis que Rocha et al. [5], avec un pipeline CNN-SVM, atteignent 65,5%. L'approche de Perna et Tagarelli [9], fondée sur des spectrogrammes et des architectures profondes, se situe à 80,6%, soit un niveau comparable au nôtre. Le modèle multimodal LungBRN proposé par Ma et al. [22] obtient 79%, confirmant l'intérêt des architectures hybrides.

Bien que certaines études récentes non incluses dans nos références rapportent des performances plus élevées (par exemple Bansal 2021 ou Shovo 2024), leurs résultats sont souvent obtenus dans des conditions différentes (protocole d'entraînement propriétaire, bases enrichies, prétraitement personnalisé), rendant la comparaison directe difficile. Dans ce contexte, notre modèle se distingue par un compromis pertinent entre simplicité architecturale, fusion multimodale explicite et performance robuste, ce qui confirme la pertinence clinique de l'approche proposée.

***Tableau 2 :*** *Comparaison avec des méthodes existantes*

| Référence | Méthode | Accuracy |
|---|---|---|
| **Aykanat et al. (2017) [10]** | CNN simple | 76.0% |
| **Rocha et al. (2018) [5]** | CNN+SVM | 65.5% |
| **Perna et Tagarelli (2019) [9]** | Spectrogram CNN | 80.6% |
| **Ma et al. (2020) [22]** | LungBRN | 79.0% |
| **Cette étude (2025)** | CNN bimodal (MFCC + Mél) | 81.0% |

## 6. CONCLUSION

Cette étude démontre la viabilité d'une approche hybride CNN pour la classification des sons respiratoires. Avec 81% de précision globale et des AUC entre 0.91-0.99, le système montre un potentiel clinique pour le triage en première ligne. La contribution principale réside dans la validation empirique de la complémentarité MFCC/mél-spectrogrammes, chacune capturant des aspects différents du signal respiratoire.

Le modèle proposé pourrait être déployé sur dispositifs mobiles [20] pour assister les cliniciens en zones sous-médicalisées. Toutefois, des validations prospectives en contextes cliniques réels restent indispensables avant un déploiement à grande échelle. Cette étude pose une brique vers des systèmes d'aide au diagnostic respiratoire plus accessibles, particulièrement dans les pays en développement où les besoins médicaux sont criants.

## RÉFÉRENCES


[1] Organisation Mondiale de la Santé. (2023). Maladies respiratoires chroniques : Principaux faits. WHO Global Health Estimates. https://www.who.int/news-room/fact-sheets/detail/chronic-respiratory-diseases

[2] Bohadana, A., Izbicki, G., & Kraman, S. S. (2014). Fundamentals of lung auscultation. New England Journal of Medicine, 370(8), 744-751. doi:10.1056/NEJMra1302901

[3] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366. doi:10.1109/TASSP.1980.1163420

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539

[5] Rocha, B. M., Filos, D., Mendes, L., Vogiatzis, I., Perantoni, E., Kaimakamis, E., ... & Maglaveras, N. (2018). A respiratory sound database for the development of automated classification. In Precision Medicine Powered by pHealth and Connected Health (pp. 33-37). Springer, Singapore. doi:10.1007/978-981-10-7419-6_6

[6] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference (Vol. 8, pp. 18-25).

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. doi:10.1613/jair.953

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). doi:10.1109/CVPR.2016.90

[9] Perna, D., & Tagarelli, A. (2019). Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 50-55). IEEE. doi:10.1109/CBMS.2019.00020

[10] Aykanat, M., Kılıç, Ö., Kurt, B., & Saryal, S. (2017). Classification of lung sounds using convolutional neural networks. EURASIP Journal on Image and Video Processing, 2017(1), 65. doi:10.1186/s13640-017-0213-2

[11] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations (ICLR). arXiv preprint arXiv:1412.6980.

[12] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

[13] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America, 8(3), 185-190. doi:10.1121/1.1915893

[14] Pasterkamp, H., Kraman, S. S., & Wodicka, G. R. (1997). Respiratory sounds: advances beyond the stethoscope.

American Journal of Respiratory and Critical Care Medicine, 156(3), 974-987. Fundamentals of lung auscultation. New England Journal of Medicine, 370(8), 744-751. doi:10.1164/ajrccm.156.3.9701115

[15] Reichert, S., Gass, R., Brandt, C., & Andrès, E. (2008). Analysis of respiratory sounds: state of the art. Clinical Medicine: Circulatory, Respiratory and Pulmonary Medicine, 2, CCRPM-S530. doi:10.4137/CCRPM.S530

[16] Serbes, G., Ulukaya, S., & Kahya, Y. P. (2018). An automated lung sound preprocessing and classification system based on spectral analysis methods. In Precision Medicine Powered by pHealth and Connected Health (pp. 45-49). Springer, Singapore.

[17] Palaniappan, R., Sundaraj, K., & Ahamed, N. U. (2013). Machine learning in lung sound analysis: a systematic review. Biocybernetics and Biomedical Engineering, 33(3), 129-135. doi:10.1016/j.bbe.2013.07.001

[18] Chambres, G., Hanna, P., & Desainte-Catherine, M. (2018). Automatic detection of patient with respiratory diseases using lung sound analysis. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-6). IEEE. doi:10.1109/CBMI.2018.8516489

[19] Ntalampiras, S., & Potamitis, I. (2020). Transfer learning for improved audio-based human activity recognition. Biosensors, 10(5), 48. doi:10.3390/bios10050048

[20] Acharya, J., & Basu, A. (2020). Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. IEEE Transactions on Biomedical Circuits and Systems, 14(3), 535-544. doi:10.1109/TBCAS.2020.2981172

[21] Fraiwan, M., Fraiwan, L., Khassawneh, B., & Ibnian, A. (2021). A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. Data in Brief, 35, 106913. doi:10.1016/j.dib.2021.106913

[22] Ma, Y., Xu, X., Yu, Q., Zhang, Y., Li, Y., Zhao, J., & Wang, G. (2019). LungBRN: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm. In 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS) (pp. 1-4). IEEE. doi:10.1109/BIOCAS.2019.8919021