Détection automatique de la tuberculose à partir de sons de toux

par

Assaad CHIBOUB

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE EN TECHNOLOGIE DE L'INFORMATION
M. Sc. A.

MONTRÉAL, LE 02 MARS 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. David Labbé, directeur de mémoire
Département de génie logiciel et TI à l'École de technologie supérieure

Mme. Neila Mezghani, codirectrice
Université TÉLUQ

M. Rachid Aissaoui, président du jury
Département de génie logiciel à l'École de technologie supérieure

M. Carlos Vázquez, membre du jury
Département de génie logiciel et TI à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 27 JANVIER 2026

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

Les travaux présentés ont été effectués entre 2024 et 2025 au laboratoire de recherche en imagerie et orthopédie (LIO), sous la direction du Prof. David Labbé et de la Prof. Neila Mezghani, en collaboration avec des chercheurs et des professionnels de la recherche.

L'objectif de ce mémoire est de développer une contribution scientifique et méthodologique visant à améliorer le dépistage automatisé de la tuberculose à partir de sons de toux. Plus précisément, il s'agit d'explorer, d'implémenter et de valider des approches d'intelligence artificielle appliquées à des données audio cliniques, en mobilisant des techniques avancées de traitement du signal et d'apprentissage profond, afin de concevoir des outils de dépistage non invasifs, accessibles et fiables.

Ce travail s'inscrit dans un contexte de santé mondiale, où le besoin d'outils de triage rapides et à faible coût demeure crucial, notamment dans les régions à ressources limitées. Il contribue à la littérature scientifique en mettant en évidence le potentiel des sons de toux comme biomarqueur objectif et en évaluant la pertinence de leur intégration dans des systèmes intelligents de dépistage.

REMERCIEMENTS

La réalisation de ce mémoire a été rendue possible grâce au soutien et à l'accompagnement de nombreuses personnes, que je souhaite remercier sincèrement.

Je tiens avant tout à exprimer ma profonde gratitude à Dre Neila Mezghani et Dr David Labbé, qui m'ont accueilli au laboratoire de recherche en imagerie et orthopédie (LIO) et ont guidé mes travaux avec rigueur et bienveillance. Leurs orientations scientifiques, leurs conseils pertinents et leur disponibilité ont joué un rôle déterminant dans l'aboutissement de ce projet.

Je suis également reconnaissant envers M. Rachid Raïssaoui et M. Carlos Vázquez, qui ont accepté d'évaluer ce mémoire et dont les remarques constructives ont contribué à son amélioration.

Mes remerciements vont aussi à M. Youssef Ouakrim, M. Johannes Ayena et M. Karim Loulou, pour leur disponibilité, leurs encouragements et leurs conseils avisés. Je remercie également l'ensemble de l'équipe du LIO pour la convivialité, la collaboration et l'esprit d'entraide qui ont enrichi mon expérience scientifique et humaine.

Je souhaite aussi exprimer ma plus profonde gratitude à ma famille, en particulier à mes parents et à mon grand frère. Leur amour, leurs encouragements et leurs sacrifices m'ont donné la force nécessaire pour mener ce parcours à terme. Je leur rends un hommage particulier pour leurs prières, leur confiance et leur soutien indéfectible, qui ont constitué le socle moral et spirituel de ce projet.

À toutes celles et ceux qui ont contribué, de près ou de loin, je dis simplement : Merci.

Enfin, je me remercie moi-même pour la persévérance et les efforts qui m'ont permis d'achever ce travail.

Détection automatique de la tuberculose à partir de sons de toux

Assaad CHIBOUB

RÉSUMÉ

La tuberculose demeure l'une des principales menaces sanitaires mondiales, avec plus de dix millions de nouveaux cas chaque année selon l'OMS, et constitue un enjeu majeur de santé publique, particulièrement dans les pays à faibles ressources. Les méthodes actuelles de dépistage, bien qu'efficaces (tests moléculaires, cultures, imagerie), présentent des contraintes importantes en termes de coût, de délai et d'accessibilité, limitant leur déploiement à grande échelle.

Ce mémoire propose une approche innovante de détection automatisée de la tuberculose à partir de sons de toux, reposant sur des techniques avancées de traitement du signal et d'apprentissage profond. À partir du corpus fourni par le CODA TB DREAM Challenge 2022, un pipeline complet a été conçu et implémenté : prétraitement audio, extraction de caractéristiques (MFCC, spectrogrammes de Mel et statistiques globales), augmentation des données, architectures neuronales hybrides CNN–BiGRU–Attention, et optimisation systématique des hyperparamètres par Optuna. Une validation croisée stricte par sujet a été mise en place pour garantir la robustesse et la généralisabilité des résultats.

Les résultats obtenus confirment le potentiel de l'analyse acoustique de la toux comme outil complémentaire, non invasif et abordable, pour améliorer le dépistage précoce de la tuberculose et contribuer à la réduction de sa transmission communautaire.

Mots-clés: tuberculose, dépistage précoce, sons de toux, traitement du signal, apprentissage profond

Automatic Tuberculosis Detection from Cough Sounds

Assaad CHIBOUB

ABSTRACT

Tuberculosis remains one of the leading global health threats, with over ten million new cases reported annually according to WHO. Despite their effectiveness, current diagnostic methods (molecular assays, cultures, imaging) are still limited by cost, infrastructure requirements, and turnaround time, which restricts their widespread use in low-resource settings.

This thesis introduces an innovative automated tuberculosis detection system based on cough sounds, leveraging advanced signal processing and deep learning. Using the dataset provided by the CODA TB DREAM Challenge 2022, we designed a complete pipeline comprising audio preprocessing, feature extraction (MFCC, Mel spectrograms, and global statistics), data augmentation, hybrid deep neural architectures CNN–BiGRU–Attention, and hyperparameter optimization with Optuna. To ensure robust and generalizable results, strict subject-level cross-validation was applied throughout the experiments.

The findings demonstrate the potential of cough sound analysis as a complementary, non-invasive, and cost-effective approach for early tuberculosis screening, with significant implications for reducing community transmission and improving global health outcomes.

Keywords: tuberculosis, early screening, cough sounds, signal processing, deep learning

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AUC	Aire sous la courbe (Area Under the Curve)
AUC-ROC	Aire sous la courbe ROC
BCG	Bacille de Calmette et Guérin (vaccin)
BiGRU	GRU bidirectionnelle (Bidirectional GRU)
CNN	Réseau de neurones convolutionnel (Convolutional Neural Network)
CODA	CODA TB DREAM Challenge (jeu de données / challenge sur la toux)
CRNN	Réseau convolutionnel-récurrent (Convolutional Recurrent Neural Network)
DREAM	Dialogue for Reverse Engineering Assessments and Methods
FN	Faux négatif (False Negative)
FP	Faux positif (False Positive)
FPR	Taux de faux positifs (False Positive Rate)
GRU	Unité récurrente à portes (Gated Recurrent Unit)
GeneXpert	Xpert MTB/RIF (test moléculaire GeneXpert)
IGRA	Interferon-Gamma Release Assay (test de libération d'interféron-gamma)
LIO	Laboratoire de recherche en imagerie et orthopédie
LSTM	Mémoire à long court terme (Long Short-Term Memory)
MFCC	Coefficients cepstraux en fréquences de Mel (Mel-Frequency Cepstral Coefficients)
MTB	*Mycobacterium tuberculosis*
OMS	Organisation mondiale de la Santé (WHO)
OOF	Out-of-fold (validation croisée hors-plis)
PCR	Réaction en chaîne par polymérase (Polymerase Chain Reaction)
RF	Forêt aléatoire (Random Forest)
RIF	Rifampicine
RMS	Valeur quadratique moyenne (Root Mean Square)
RNN	Réseau neuronal récurrent (Recurrent Neural Network)

ROC	Receiver Operating Characteristic (courbe ROC)
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Machine à vecteurs de support (Support Vector Machine)
TB	Tuberculose
TN	Vrai négatif (True Negative)
TP	Vrai positif (True Positive)
TPR	Taux de vrais positifs (sensibilité, True Positive Rate)
TS	Time stretching (étirement temporel)
TST	Test cutané à la tuberculine (Tuberculin Skin Test)
VIH	Virus de l'immunodéficience humaine (HIV)
XGBoost	eXtreme Gradient Boosting
ZCR	Taux de passages par zéro (Zero-Crossing Rate)
dB	Décibel
pAUC	Aire sous la courbe partielle (partial AUC)
PH	Philippines
VN	Vietnam
SA	Arabie Saoudite
UG	Ouganda
IN	Inde
MG	Madagascar
TZ	Tanzanie

INTRODUCTION

La tuberculose est une maladie infectieuse d'une grande complexité biomédicale et épidémiologique, causée par Mycobacterium tuberculosis. Elle représente encore aujourd'hui l'une des principales menaces sanitaires mondiales. Selon l'Organisation mondiale de la santé (OMS, 2023), environ 10,6 millions de personnes ont développé la maladie et 1,3 million en sont décédées en 2022, faisant de la tuberculose la deuxième cause de mortalité infectieuse après la COVID-19. Cette charge demeure particulièrement lourde dans les pays à faibles ressources, où les infrastructures médicales sont limitées et l'accès au dépistage reste restreint.

Le diagnostic de la tuberculose repose actuellement sur deux grandes familles d'approches. La première regroupe les examens de laboratoire tels que la microscopie des frottis, la culture bactérienne ou les tests moléculaires comme le GeneXpert (World Health Organization, 2024). Ces méthodes offrent une sensibilité élevée, mais exigent des équipements spécialisés, un personnel formé et des délais d'analyse parfois trop longs pour un dépistage rapide. La seconde famille d'approches repose sur l'évaluation clinique en cabinet médical, examen physique, antécédents du patient et observation des signes tels que la toux chronique, la fièvre ou la perte de poids (Pai et al., 2016). Bien que plus accessibles, ces méthodes manquent de spécificité et ne permettent pas de confirmer le diagnostic de manière fiable.

Ces limites justifient le besoin de développer des alternatives de dépistage plus rapides, accessibles et non invasives, capables de compléter les outils diagnostiques existants, notamment dans les contextes à faibles ressources. Dans ce cadre, la toux se présente comme un biomarqueur prometteur : symptôme cardinal de la tuberculose pulmonaire, elle reflète les altérations physiopathologiques des voies respiratoires. Plusieurs travaux récents ont montré que les signatures acoustiques contenues dans les sons de toux peuvent être exploitées par des techniques avancées de traitement du signal et d'intelligence artificielle pour identifier la maladie (Pahar et al., 2021).

L'objectif général de ce projet de recherche est de concevoir une méthode de détection automatique de la tuberculose à partir de sons de toux. Ce travail s'appuie sur les données du CODA TB DREAM Challenge 2022, qui regroupe un corpus de milliers d'enregistrements de toux annotés, et propose une méthodologie rigoureuse combinant extraction de caractéristiques acoustiques, augmentation des données, entraînement de modèles de réseaux neuronaux profonds et optimisation d'hyperparamètres.

CHAPITRE 1

PROBLÉMATIQUE ET OBJECTIFS

1.1 Problématique scientifique

La tuberculose (TB) est une maladie infectieuse causée par la bactérie *Mycobacterium tuberculosis*, touchant principalement les poumons mais pouvant atteindre d'autres organes. Elle demeure l'une des principales causes de mortalité infectieuse dans le monde, avec plus de 10 millions de nouveaux cas par an et 1,3 million de décès estimés en 2022 (World Health Organization, 2023a). Dans les régions à ressources limitées, le dépistage reste un défi majeur : les méthodes de référence, telles que la culture bactérienne, qui consiste à faire croître le bacille sur un milieu spécifique, ou la PCR en temps réel (par exemple le test *GeneXpert MTB/RIF*, permettant la détection rapide de l'ADN du bacille et de la résistance à la rifampicine), sont précises mais coûteuses, lentes et difficilement accessibles. Dans ce contexte, l'analyse acoustique de la toux suscite un intérêt croissant comme outil de triage rapide, non invasif et peu coûteux (le triage clinique désigne l'étape préliminaire consistant à identifier les individus à risque avant des tests confirmatoires).

Cependant, plusieurs verrous scientifiques et techniques freinent encore le transfert clinique à grande échelle.

Premièrement, l'ancrage physiopathologique des marqueurs audio doit être clarifié : les descripteurs temps–fréquence extraits, tels que les coefficients cepstraux en fréquences de Mel (MFCC), qui représentent la structure spectrale perçue du son selon une échelle inspirée de la sensibilité auditive humaine, et les spectrogrammes, qui illustrent la distribution de l'énergie du signal dans le temps et la fréquence, captent-ils réellement des signatures spécifiques de la tuberculose ou sont-ils influencés par des artefacts liés aux conditions d'enregistrement (bruit, dispositif, protocole) (Kapetanidis et al., 2024) ?

Deuxièmement, la capacité de généralisation des modèles demeure limitée : des performances encourageantes observées sur un corpus peuvent se dégrader fortement lorsqu'on change de

population, d'environnement ou d'appareil d'acquisition (Pahar et al., 2021). La généralisabilité correspond ici à la capacité du modèle à maintenir ses performances sur de nouvelles données non vues.

Enfin, la transférabilité clinique exige des protocoles d'évaluation rigoureux, incluant une validation indépendante par sujet, une estimation des performances au niveau participant et un calibrage du seuil de décision, c'est-à-dire l'ajustement du point de coupure pour maximiser la sensibilité ou la spécificité selon le contexte. Conformément aux recommandations de l'OMS, la sensibilité (proportion de vrais positifs correctement identifiés) doit être priorisée (≥ 80 %), tout en maintenant une spécificité (proportion de vrais négatifs correctement identifiés) suffisante afin d'éviter un excès de tests confirmatoires inutiles (World Health Organization, 2024).

De récents travaux illustrent ces défis. Par exemple, DMRNet (Xu et al., 2022), un réseau à double module convolutionnel et récurrent, CNNX, un modèle convolutionnel profond, et HGBoost (Suda, 2023a), une méthode d'ensemble basée sur le gradient boosting, atteignent de hautes performances dans certains contextes, mais leur reproductibilité et leur validité externe restent discutables. Le véritable enjeu est donc de concevoir un pipeline robuste, reproductible et transparent, capable de concilier exigences méthodologiques et contraintes cliniques, afin de transformer l'analyse acoustique de la toux en un outil fiable de santé publique.

Ces constats motivent la formulation d'une hypothèse centrale et de questions de recherche spécifiques, présentées dans la section suivante.

1.2 Hypothèse et questions de recherche

Hypothèse centrale

Les toux sollicitées contiennent des signatures acoustiques discriminantes permettant, en combinaison avec des métadonnées cliniques simples, d'effectuer un triage fiable de la tuberculose. Un tel système doit être généralisable à de nouveaux sujets, robuste aux conditions d'acquisition et compatible avec un déploiement en contexte de terrain.

Questions de recherche

À partir de cette hypothèse centrale, cinq questions guident ce travail :

Représentations : quelles représentations (spectrogrammes Mel, MFCC, descripteurs spectraux tels que le centroïde spectral, le *roll-off* et l'entropie spectrale, qui caractérisent respectivement le centre de gravité du spectre, la fréquence seuil d'énergie et la complexité du signal) offrent le meilleur compromis sensibilité/spécificité aux niveaux enregistrement et participant ?

Architectures : quelle configuration architecturale, par exemple un réseau convolutionnel (CNN, *Convolutional Neural Network*), capable d'extraire automatiquement des motifs locaux dans des données structurées, ou une combinaison CNN–GRU (unité récurrente à portes, *Gated Recurrent Unit*) modélisant les dépendances temporelles, permet la meilleure généralisation sujet-indépendante dans la détection de la tuberculose à partir de sons de toux ?

Métadonnées : l'intégration de métadonnées cliniques minimales (âge, sexe, symptômes rapportés) améliore-t-elle significativement la décision par rapport à l'audio seul ?

Validation et équité : quels protocoles de validation et quelles stratégies d'équilibrage des classes permettent d'obtenir une évaluation fiable, non biaisée et équitable entre les différents sous-groupes démographiques (sexe, âge, site), tout en assurant une robustesse face au bruit et à la variabilité des données ? L'équité désigne ici la constance des performances entre sous-groupes.

Décision clinique : quelles règles de décision (optimisation de seuil, vote majoritaire multi-toux, agrégation par patient) maximisent l'utilité clinique en triage ?

Ces questions de recherche orientent la définition des objectifs spécifiques et la conception de la méthodologie présentée ci-après.

1.3 Objectifs du mémoire

L'objectif général de ce travail est de concevoir et valider un pipeline automatisé de triage de la tuberculose à partir de toux sollicitées, visant une sensibilité élevée, une spécificité compatible avec un usage clinique et une robustesse face à la variabilité inter-sujets et inter-appareils.

Pour atteindre cet objectif général, trois objectifs spécifiques ont été fixés :

1. **Étudier la qualité du corpus audio.** Mettre en place un protocole rigoureux de prétraitement et de contrôle qualité des enregistrements de toux et des métadonnées associées, afin d'assurer l'intégrité, la normalisation et la représentativité de la base de données.
2. **Identifier les représentations acoustiques et les modèles d'apprentissage les plus discriminants.** Analyser et comparer différentes représentations du signal de toux (coefficients cepstraux en fréquences de Mel, spectrogrammes Mel, descripteurs spectraux) et concevoir des architectures d'apprentissage profond adaptées (réseaux convolutionnels et récurrents) afin de capturer les signatures temporelles et spectrales caractéristiques de la tuberculose.
3. **Optimiser et évaluer la performance, la robustesse et l'équité du système.** Optimiser les paramètres d'entraînement, mettre en œuvre des protocoles de validation croisée sujet-indépendante (méthode consistant à évaluer la performance sur des sous-ensembles exclusifs de participants), évaluer la généralisabilité du modèle face aux sources de variabilité (âge, sexe, site d'enregistrement) et analyser les performances selon des sous-groupes afin de garantir équité, stabilité et reproductibilité.

Ces objectifs orientent la démarche méthodologique décrite dans la section suivante.

1.4 Démarche méthodologique générale

La méthodologie adoptée repose sur un pipeline structuré en plusieurs étapes successives et complémentaires, allant du prétraitement des signaux audio à l'évaluation avancée du modèle.

Chaque composante du processus vise à renforcer la robustesse, la reproductibilité et la validité scientifique des résultats obtenus.

- **Prétraitement et contrôle de qualité** : cette étape comprend la conversion et le rééchantillonnage des signaux à 16 kHz, la normalisation d'amplitude et la pré-accentuation (filtrage amplifiant les hautes fréquences avant l'analyse), suivies d'un contrôle d'intégrité portant sur la durée, le *clipping* (saturation du signal au-delà de la plage de mesure) et le décalage DC (composante continue déplaçant le signal autour d'une valeur moyenne non nulle). L'objectif est de réduire les biais techniques et d'assurer une homogénéité optimale des entrées.
- **Représentations et caractéristiques** : les signaux prétraités sont convertis en représentations temps–fréquence telles que les spectrogrammes Mel et log-Mel, où les amplitudes spectrales sont exprimées sur une échelle logarithmique. Des descripteurs acoustiques (MFCC, centroïde spectral, *roll-off*, entropie spectrale) sont extraits pour capturer différentes dimensions du signal. Des combinaisons audio–tabulaires sont également explorées afin d'intégrer les métadonnées associées aux participants.
- **Modélisation** : un ensemble d'architectures neuronales est conçu et comparé, incluant un modèle CNN–GRU intégrant un mécanisme d'attention (procédé qui pondère dynamiquement les parties les plus informatives du signal dans le temps) pour l'agrégation temporelle, et une variante CNN pure servant de référence. Une fusion tardive est mise en œuvre pour combiner les informations issues des données acoustiques et tabulaires.
- **Optimisation et validation** : l'ajustement automatique des hyperparamètres est réalisé à l'aide de la bibliothèque **Optuna** (*Optimization Utilities for Neural Architectures*), une méthode d'optimisation bayésienne open-source. La performance de chaque configuration est évaluée par une validation croisée stricte à cinq plis, groupée par sujet, garantissant qu'aucun enregistrement d'un même participant ne soit présent simultanément dans l'entraînement et la validation. Enfin, un calibrage du seuil de décision est appliqué afin d'adapter le modèle à un usage de dépistage.
- **Évaluation avancée** : des analyses d'ablation (retrait de certains composants du modèle pour mesurer leur contribution) sont effectuées pour évaluer l'apport de chaque élément, complétées

par des tests de robustesse (ajout de bruit, variabilité des appareils) et une évaluation d'équité entre sous-groupes démographiques (sexe, âge, site). Ces étapes garantissent la validité interne et externe du modèle et assurent la transparence scientifique requise pour une future validation clinique.

1.5 Conclusion

Ce chapitre a présenté la problématique scientifique du dépistage automatisé de la tuberculose par la toux, formulé l'hypothèse centrale et les questions de recherche, et défini les objectifs ainsi que la démarche méthodologique générale. L'ambition de ce mémoire est de démontrer qu'un pipeline rigoureux, multimodal et reproductible peut constituer une base solide pour un futur outil de triage clinique de la tuberculose, en particulier dans les contextes à ressources limitées. Le chapitre suivant aborde le cadre médical et biomédical de la tuberculose, en présentant ses caractéristiques épidémiologiques, ses mécanismes de transmission et ses méthodes de diagnostic actuelles.

CHAPITRE 2

LA TUBERCULOSE : CONTEXTE MÉDICAL ET BIOMÉDICAL

2.1 Introduction

Dans ce chapitre, nous posons le cadre médical et épidémiologique de la tuberculose, examinons l'état du dépistage et ses limites, et présentons les biomarqueurs qui sous-tendent le recours à l'intelligence artificielle (IA) pour l'analyse de la toux. L'objectif est d'offrir un socle conceptuel clair et complet pour les développements méthodologiques ultérieurs du mémoire.

2.2 Aspects épidémiologiques et cliniques de la tuberculose

2.2.1 Origine et agents pathogènes

La tuberculose est causée par une bactérie du complexe *Mycobacterium tuberculosis* (MTBC), c'est-à-dire un groupe d'espèces mycobactériennes pathogènes partageant une très forte identité génomique et des caractéristiques phénotypiques communes. La plus courante chez l'humain est *M. tuberculosis* sensu stricto ; on parle historiquement du *bacille de Koch* pour désigner cette espèce découverte par Robert Koch en 1882. Cette bactérie a coévolué avec l'espèce humaine depuis des millénaires, comme en témoignent des traces génétiques retrouvées dans des restes momifiés datant de plus de 9 000 ans (Hershkovitz et al., 2008). Son origine zoonotique probable serait liée à une transmission croisée avec des espèces mycobactériennes bovines (*M. bovis*), bien que l'origine précise fasse toujours l'objet de débats scientifiques (Brosch et al., 2002).

M. tuberculosis est un bacille strictement aérobie, non sporulé, à croissance lente, caractérisé par une paroi cellulaire riche en lipides qui lui confère une résistance élevée aux stress environnementaux et à certaines réponses immunitaires. Cette structure explique sa capacité à persister dans l'organisme sous forme latente pendant des années ainsi qu'une résistance partielle à certains antibiotiques. En microscopie, après la *coloration de Ziehl–Neelsen* (technique ciblant

les bactéries acido-alcoolo-résistantes via la fuchsine basique), le bacille apparaît sous forme de bâtonnets rouges.

Parmi les autres espèces pathogènes du complexe figurent *M. africanum*, plus fréquent en Afrique de l'Ouest, *M. bovis*, transmissible par le lait non pasteurisé, ou encore *M. canettii*, une forme rare identifiée principalement à Djibouti (Supply et al., 2013). Ces espèces partagent une grande homologie génétique tout en différant légèrement en termes de transmission, de pathogénicité et de distribution géographique. La compréhension de la biologie et de la diversité génomique du MTBC constitue un fondement essentiel pour le développement de stratégies de dépistage et de traitement efficaces.

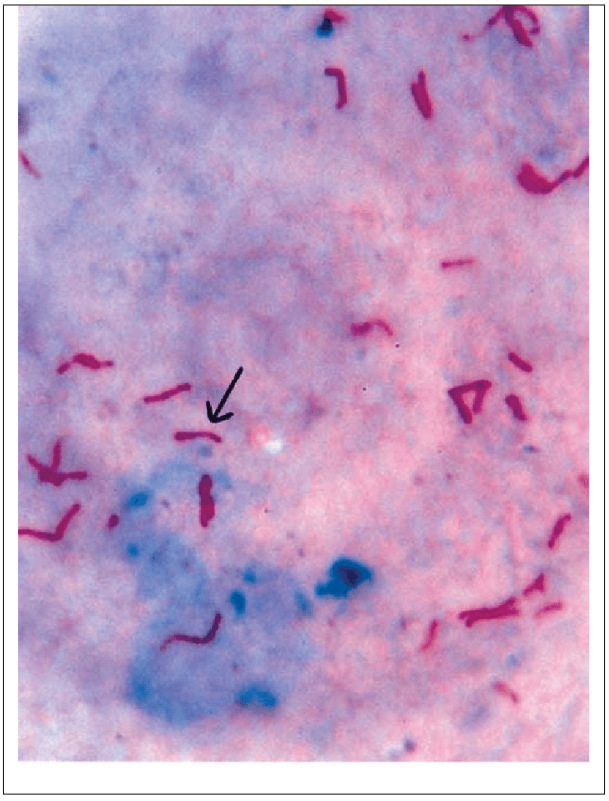


Figure 2.1 Observation microscopique de *Mycobacterium tuberculosis* après coloration de Ziehl–Neelsen.
Tirée de Dubey et al. (2012)

2.2.2 Statistiques mondiales et fardeau sanitaire

Selon le rapport mondial de l'Organisation mondiale de la santé (OMS) de 2023, la tuberculose a touché environ 10,6 millions de personnes dans le monde en 2022, entraînant 1,3 million de décès parmi les personnes non porteuses du VIH, et 167 000 décès chez les personnes vivant avec le VIH. La maladie demeure la treizième cause de mortalité dans le monde et la principale cause de décès d'origine bactérienne (World Health Organization, 2023a). La majorité des cas se concentre dans des régions à faibles et moyens revenus, notamment en Asie du Sud-Est, en Afrique subsaharienne et dans certaines zones de l'Europe de l'Est. Huit pays concentrent à eux seuls plus des deux tiers des cas mondiaux, dont l'Inde, l'Indonésie, la Chine, les Philippines et le Nigéria.

Le fardeau sanitaire est aggravé par les formes multirésistantes de la tuberculose (TB-MDR), définies par une résistance au moins à l'isoniazide et à la rifampicine, deux antituberculeux majeurs. En 2023, près de 410,000 cas de TB-MDR ont été détectés, dont seulement 50 % ont reçu un traitement adéquat (World Health Organization, 2023a). Comme l'illustre la Figure 2.2, l'incidence mondiale de la tuberculose demeure particulièrement élevée dans certaines régions d'Asie et d'Afrique, reflétant une forte inégalité géographique dans la charge de la maladie.

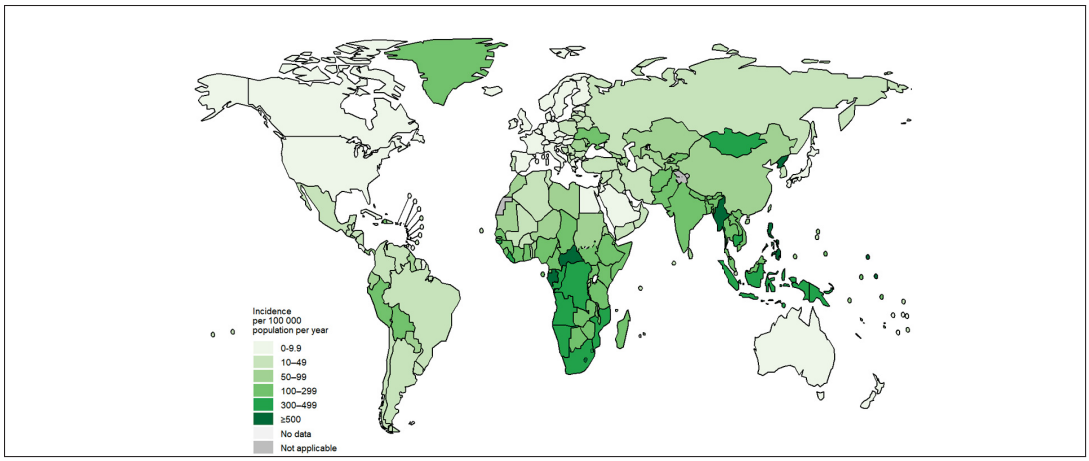


Figure 2.2 Épidémiologie mondiale de la tuberculose (incidence 2023).
Tirée de World Health Organization (2023c)

2.2.3 Groupes à risque et facteurs socio-économiques associés

La tuberculose affecte de manière disproportionnée certains groupes vulnérables en raison de facteurs biologiques, sociaux et économiques. Parmi les populations à risque figurent les personnes vivant avec le VIH, les enfants, les personnes âgées, les migrants, les détenus, les personnes sans domicile fixe, ainsi que les travailleurs de la santé exposés régulièrement à des cas infectieux (Lönnroth et al., 2010). Les *déterminants sociaux de la santé* désignent les conditions économiques et sociales qui influencent la santé et l'accès aux soins ; ils incluent notamment la pauvreté, la malnutrition, la surpopulation et la difficulté d'accès au système de santé. La stigmatisation associée à la maladie contribue également à des retards de diagnostic et à une observance thérapeutique insuffisante, renforçant la transmission communautaire.

2.2.4 Présentations cliniques

La forme pulmonaire constitue la présentation la plus fréquente de la tuberculose et représente environ 75 % des cas. Elle se manifeste par une toux persistante, parfois accompagnée d'expectorations, de la fièvre, des sueurs nocturnes, de la perte de poids et de la fatigue chronique. Lorsque la maladie progresse, elle peut provoquer des lésions *cavitaires* visibles à l'imagerie thoracique ; une cavitation pulmonaire correspond à une zone nécrotique creusée dans le parenchyme (World Health Organization, 2021a).

La tuberculose extrapulmonaire concerne environ 15 % des cas, tandis qu'environ 10 % associent simultanément une atteinte pulmonaire et extrapulmonaire. Parmi les localisations extrapulmonaires fréquentes figurent l'épanchement pleural (20 %), l'atteinte ganglionnaire (15 %), les atteintes ostéo-articulaires (10 %), la méningite tuberculeuse (8 %), la tuberculose miliaire (8 %), ainsi que des formes plus rares telles que les atteintes génito-urinaires (5 %) ou abdominales. Ces formes sont souvent plus difficiles à diagnostiquer, en particulier chez les enfants ou les personnes immunodéprimées, et nécessitent des approches spécialisées comme les biopsies ou l'imagerie ciblée (Golden and Vikram, 2005).

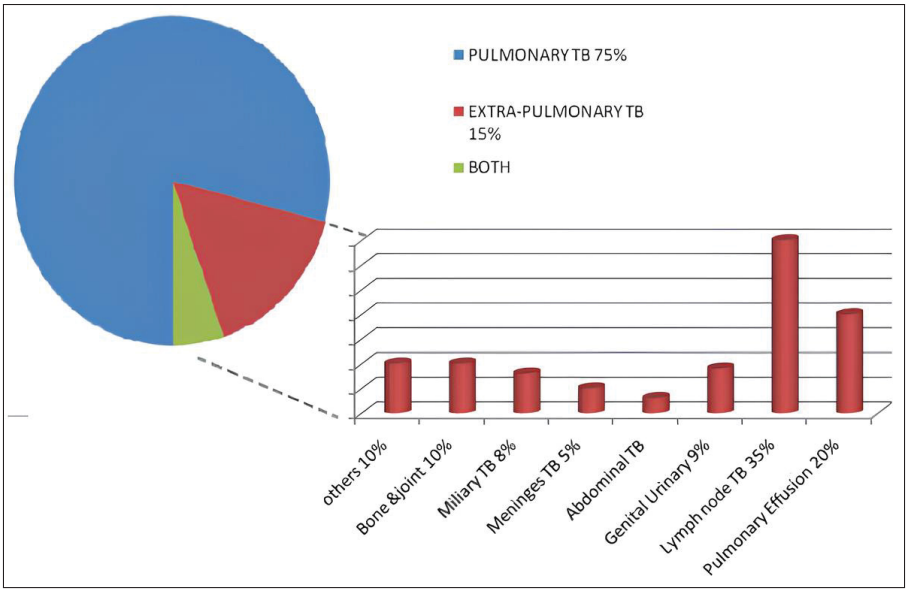


Figure 2.3 Radiographie thoracique montrant une cavitation apicale droite.
Tirée de Centers for Disease Control and Prevention (2000)

La Figure 2.3 illustre une présentation typique de tuberculose pulmonaire avancée, avec cavitation apicale visible à la radiographie.

2.3 Mécanismes de transmission et évolution de la maladie

2.3.1 Modes de transmission

La tuberculose se transmet principalement par voie aérienne lorsque des personnes infectées expulsent des gouttelettes contenant les bacilles de Koch en toussant, parlant, éternuant ou chantant. On distingue classiquement les *gouttelettes* de taille supérieure à 5 µm, qui retombent rapidement, et les *aérosols* de taille inférieure à 5 µm, qui peuvent rester en suspension dans l'air pendant de longues périodes et favoriser la contagion, notamment dans des environnements clos et mal ventilés (Tellier et al., 2019). L'inhalation d'un petit nombre de bactéries peut suffire à provoquer une infection chez une personne sensible. La Figure 2.4 schématise la transmission par aérosols.

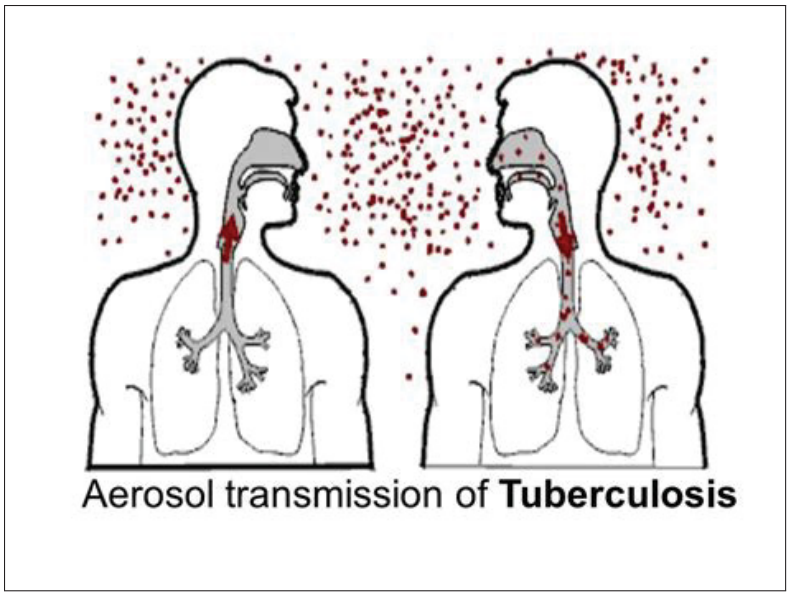


Figure 2.4 Schéma simplifié de la transmission de la tuberculose par aérosols.
Tirée de Chatham County Public Health Department (2025)

2.3.2 Incubation et progression de l'infection

Après l'exposition, la période d'incubation de la tuberculose varie de quelques semaines à plusieurs mois. Durant cette phase, les bacilles inhalés sont phagocytés par les *macrophages alvéolaires*, cellules immunitaires résidentes chargées d'englober et d'éliminer les agents pathogènes, au sein desquelles *M. tuberculosis* peut survivre et se multiplier. Le système immunitaire peut ensuite contenir l'infection sans l'éliminer, menant à une forme latente (Flynn and Chan, 2001). La progression vers une forme active dépend de nombreux facteurs, dont l'état immunitaire, l'âge et l'état nutritionnel de l'hôte.

2.3.3 Formes latente et active

On distingue deux états principaux de l'infection tuberculeuse : la *tuberculose latente*, où le patient est porteur du bacille sans symptômes cliniques ni contagiosité, et la *tuberculose active*, caractérisée par des symptômes respiratoires, un état général altéré et une capacité de transmission.

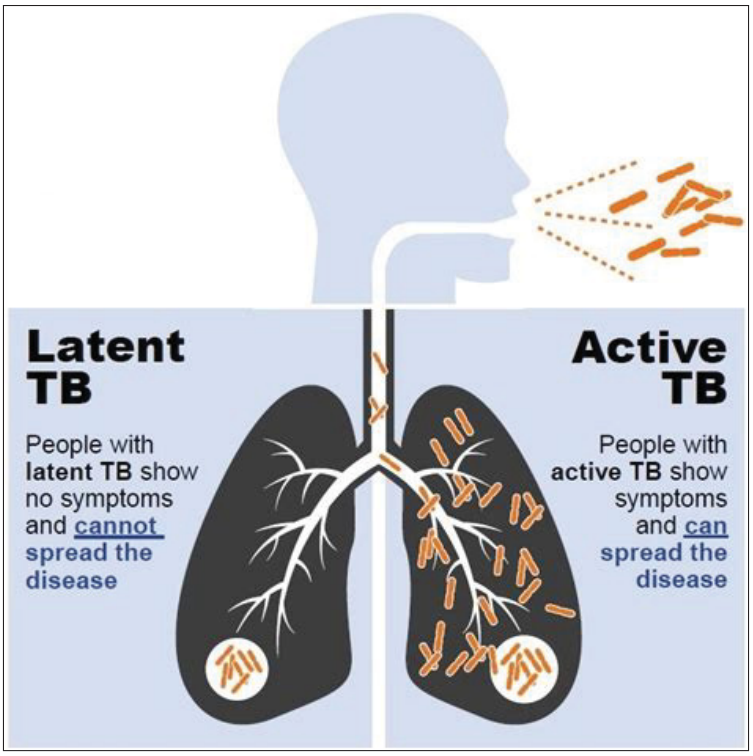


Figure 2.5 Schéma des états latent et actif de la tuberculose.
Tirée de World Health Organization (2021b)

On estime qu'environ un quart de la population mondiale est porteur d'une infection latente (Houben and Dodd, 2016). Le risque de réactivation vers une forme active est plus élevé chez les personnes immunodéprimées, notamment les personnes vivant avec le VIH. La Figure 2.5 illustre la dynamique entre latence et maladie active.

2.4 Méthodes de diagnostic conventionnelles

2.4.1 Test de dépistage cutané (Mantoux), IGRA

Le test cutané à la tuberculine (TST, *Tuberculin Skin Test*) consiste en l'injection intradermique de *PPD* (*Purified Protein Derivative*, extrait protéique purifié issu de *M. tuberculosis*) au niveau de l'avant-bras, suivie d'une lecture 48 à 72 heures plus tard. La réaction est mesurée en

millimètres d'induration et interprétée selon des seuils définis par le risque et l'état immunitaire du patient (Menzies et al., 2007). Bien qu'il soit largement utilisé, ce test présente des limites de spécificité, notamment chez les individus ayant reçu la vaccination par le *BCG* (*Bacillus Calmette–Guérin*, vaccin vivant atténué dérivé de *M. bovis*), en raison de réactions croisées. Les tests *IGRA* (*Interferon-Gamma Release Assays*), tels que le QuantiFERON-TB Gold et le T-SPOT.TB, mesurent la production d'interféron gamma par les lymphocytes T après exposition à des antigènes spécifiques de *M. tuberculosis*. Ces tests ne sont pas influencés par la vaccination BCG et offrent une meilleure spécificité, mais ils restent coûteux et nécessitent un laboratoire équipé (Menzies et al., 2007; Mazurek et al., 2010).

2.4.2 Radiographie thoracique et tomodensitométrie

La radiographie thoracique est un examen d'imagerie rapide permettant de détecter des anomalies pulmonaires compatibles avec une tuberculose active, telles que des infiltrats, cavernes ou fibroses (Kendall and Furin, 2021). Toutefois, ces anomalies ne sont pas spécifiques et doivent être confirmées par des examens microbiologiques. La tomodensitométrie (*CT-scan*) offre une résolution supérieure, permettant la détection de lésions plus subtiles, notamment chez les patients immunodéprimés ou présentant des formes atypiques. Cependant, son coût élevé, la nécessité d'équipements spécialisés et l'exposition accrue aux rayonnements limitent son usage, en particulier dans les pays à ressources limitées (MacPherson et al., 2014).

2.4.3 Analyse de l'expectoration (culture, PCR, GeneXpert)

L'analyse microbiologique de l'expectoration reste la référence diagnostique (*gold standard*) pour la tuberculose pulmonaire (World Health Organization, 2023a). La culture bactérienne sur milieu de Löwenstein–Jensen ou en milieu liquide *MGIT* (*Mycobacteria Growth Indicator Tube*, système détectant la croissance via un indicateur fluorescent d'oxygène) présente la plus grande sensibilité, mais nécessite plusieurs semaines avant l'obtention des résultats. La microscopie directe (bacilloscopie) permet une détection rapide, mais sa sensibilité est limitée, notamment chez les patients paucibacillaires. Les tests moléculaires rapides, tels que la PCR en temps réel

et la plateforme GeneXpert MTB/RIF, permettent la détection de l'ADN de *M. tuberculosis* et l'identification simultanée d'une résistance à la rifampicine en moins de deux heures (Boehme et al., 2010). Ces tests ont amélioré le diagnostic, mais leur coût et la nécessité d'une alimentation électrique fiable limitent leur déploiement universel.

2.4.4 Défis persistants du diagnostic de la tuberculose

Les méthodes conventionnelles de diagnostic de la tuberculose présentent plusieurs limites majeures. Dans les contextes à faibles ressources, l'accès aux tests de culture ou aux techniques moléculaires est souvent restreint en raison de leur coût et des infrastructures nécessaires. Par ailleurs, la sensibilité de certains examens, notamment la microscopie, demeure insuffisante chez les enfants et les patients immunodéprimés, entraînant un risque de sous-diagnostic. De plus, les délais liés à la culture, parfois de plusieurs semaines, retardent la mise en place d'un traitement approprié et favorisent la transmission continue de la maladie (Pai et al., 2016). Ces limites opérationnelles motivent l'exploration de stratégies de dépistage complémentaires.

2.5 Besoin d'approches alternatives pour le dépistage

2.5.1 Contraintes dans les zones rurales ou à faible infrastructure

Dans de nombreux pays à revenus faibles ou intermédiaires, l'accès à un diagnostic de qualité reste un défi majeur. Les infrastructures de santé sont souvent limitées, les laboratoires peu équipés et les professionnels de santé sous-formés. Les populations rurales sont particulièrement vulnérables, car les distances vers les centres de diagnostic sont longues, et les coûts de déplacement ou d'analyses peuvent être prohibitifs (World Health Organization, 2023a). Ces contraintes entraînent des retards de diagnostic et favorisent la propagation silencieuse de la maladie.

2.5.2 Nécessité d'outils de dépistage rapides, non invasifs et abordables

Compte tenu des limites des méthodes conventionnelles, l'OMS et plusieurs initiatives internationales soulignent l'importance de développer des outils de dépistage accessibles, rapides et acceptables pour les patients. Idéalement, ces dispositifs devraient être utilisables en *point-of-care* (réalisables directement au lieu de soin, sans laboratoire spécialisé), tout en permettant un triage efficace des cas suspects. La non-invasivité, la portabilité et la simplicité d'utilisation sont désormais considérées comme des critères essentiels pour l'évaluation des nouvelles technologies (Pai et al., 2022).

2.5.3 Vers des approches fondées sur les données et l'intelligence artificielle

L'essor de l'IA et de l'analyse de données biomédicales ouvre de nouvelles perspectives pour le dépistage de la tuberculose. Des approches innovantes telles que l'analyse automatisée de la toux, l'interprétation d'images radiologiques ou encore l'étude de biomarqueurs vocaux permettent d'envisager des solutions de dépistage rapides, peu coûteuses et potentiellement intégrables dans des dispositifs mobiles. Ces méthodes représentent une opportunité pour lever les barrières actuelles d'accès au diagnostic et améliorer le triage à grande échelle (Sounderajah and et al., 2021). Elles servent d'appui aux stratégies de santé publique, en complément des standards actuels.

2.6 La toux comme indicateur biomédical de la tuberculose

2.6.1 Caractéristiques physiopathologiques de la toux tuberculeuse

La toux est un mécanisme réflexe de défense des voies respiratoires, déclenché par l'activation de récepteurs sensoriels et visant à expulser particules, mucosités ou agents pathogènes. Dans le contexte de la tuberculose pulmonaire, elle est souvent chronique et persistante en raison d'une inflammation profonde et d'une altération structurale des tissus pulmonaires causée par *M. tuberculosis*. Physiologiquement, la toux se déroule en trois phases successives : (1) inspiration

profonde, (2) fermeture de la glotte avec contraction des muscles expiratoires, (3) ouverture brusque de la glotte entraînant l'expulsion d'air à haute vitesse.

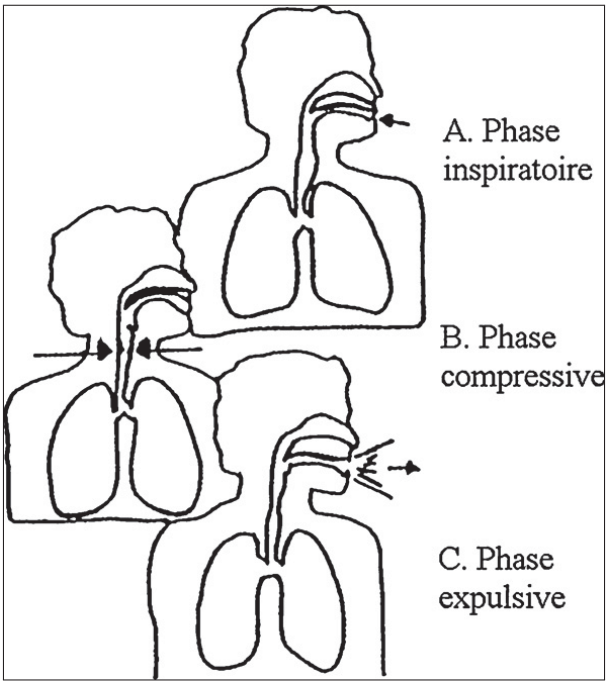


Figure 2.6 Mécanisme de la toux et altérations pulmonaires dans la tuberculose.
Tirée de Barton et al. (2012)

Chez les patients tuberculeux, cette séquence peut être perturbée. L'infection provoque des lésions cavitaires, une sécrétion excessive de mucus et une altération de la compliance pulmonaire. Ces changements affectent directement les propriétés acoustiques de la toux, rendant celle-ci plus rauque, prolongée, irrégulière ou multiphase.

2.6.2 Différences acoustiques entre toux saine et toux pathologique

L'analyse acoustique de la toux met en évidence des signatures distinctes selon l'état de santé de l'individu. La toux tuberculeuse se différencie de la toux saine ou de celle associée à d'autres affections respiratoires par plusieurs paramètres mesurables tels que la durée totale du signal, la fréquence fondamentale, la densité spectrale, la forme temporelle et la présence de phases

multiples. Ces caractéristiques peuvent être enregistrées à l'aide de microphones standards puis analysées par des méthodes de traitement du signal.

Dans ce contexte, les MFCCs constituent des descripteurs acoustiques qui représentent la structure spectrale du son selon une échelle perceptuelle inspirée de la sensibilité auditive humaine. Le *spectrogramme* est quant à lui une représentation temps–fréquence qui visualise la distribution de l'énergie du signal au cours du temps ; sa variante *Mel* applique une banque de filtres perceptuels, et la version *log-Mel* exprime les amplitudes sur une échelle logarithmique. Les informations extraites peuvent ensuite être interprétées par des algorithmes d'apprentissage automatique afin de discriminer les différents types de toux.

Plusieurs études ont montré que ces paramètres acoustiques permettent de distinguer les cas de tuberculose avec une précision dépassant 80 % (Pahar et al., 2021). Ces résultats confirment le potentiel de la toux en tant que biomarqueur utile, ouvrant la voie au développement d'outils de dépistage non invasifs, rapides et économiquement accessibles.

2.7 Conclusion

La tuberculose demeure l'une des principales menaces sanitaires mondiales, avec un impact disproportionné sur les populations vulnérables vivant dans des contextes à faibles ressources. Ce chapitre a présenté un panorama des aspects épidémiologiques, cliniques et diagnostiques de la maladie, en soulignant l'importance de la compréhension de l'agent pathogène, des mécanismes de transmission et des défis liés au dépistage précoce. Les limites des méthodes conventionnelles, combinées aux contraintes socio-économiques, justifient le développement d'approches innovantes, accessibles et non invasives. L'analyse des sons de toux, soutenue par l'IA et le traitement du signal, représente une piste prometteuse pour améliorer le diagnostic précoce et réduire la transmission communautaire. Cette approche, en complément des outils actuels, pourrait jouer un rôle dans les stratégies futures de lutte contre la tuberculose.

CHAPITRE 3

REVUE DE LA LITTÉRATURE

3.1 Introduction

Ces dernières années, les avancées conjointes en traitement du signal, apprentissage automatique et puissance de calcul ont permis l'émergence de nouvelles approches pour l'analyse de signaux biomédicaux, notamment les sons de toux. Dans le contexte de la tuberculose (TB), la toux constitue un symptôme central, potentiellement riche en informations diagnostiques. L'analyse acoustique de ce symptôme ouvre la voie à des outils de dépistage non invasifs, accessibles et peu coûteux, une perspective particulièrement prometteuse pour les régions à ressources limitées (World Health Organization, 2023a; Pahar et al., 2021).

Plusieurs travaux ont ainsi exploré l'exploitation de l'intelligence artificielle (IA) pour extraire des biomarqueurs de la toux et améliorer le diagnostic de la TB (Zimmer and Pai, 2022). Les sections suivantes présentent un état de l'art des approches de représentation du signal de toux et des modèles d'apprentissage automatique utilisés dans ce domaine.

3.2 Représentation des sons de toux

3.2.1 Descripteurs acoustiques

Pour permettre l'analyse automatique des sons de toux, ceux-ci doivent être convertis en représentations numériques pertinentes, appelées descripteurs acoustiques. Parmi ces représentations, les coefficients cepstraux en fréquences de Mel (MFCC, *Mel Frequency Cepstral Coefficients*) figurent parmi les plus utilisés dans le traitement de la parole et des sons. Ils modélisent la perception auditive humaine en projetant le spectre du signal sur une échelle perceptuelle, dite échelle de Mel. En pratique, les MFCC capturent la distribution énergétique globale du spectre de la toux, de manière condensée et robuste au bruit, ce qui les rend très efficaces pour

caractériser son timbre. Ils sont d'ailleurs largement employés dans diverses applications de reconnaissance audio et de détection de pathologies respiratoires.

Parallèlement, les spectrogrammes (souvent en échelle de Mel également) offrent une représentation temps–fréquence du signal, où l'énergie acoustique est visualisée en fonction du temps (axe horizontal) et de la fréquence (axe vertical). Cette représentation bidimensionnelle permet d'observer les variations énergétiques dans le temps et d'identifier les motifs temporels (quintes, silences) et fréquentiels (harmoniques, bruit, dispersion) propres à chaque toux.

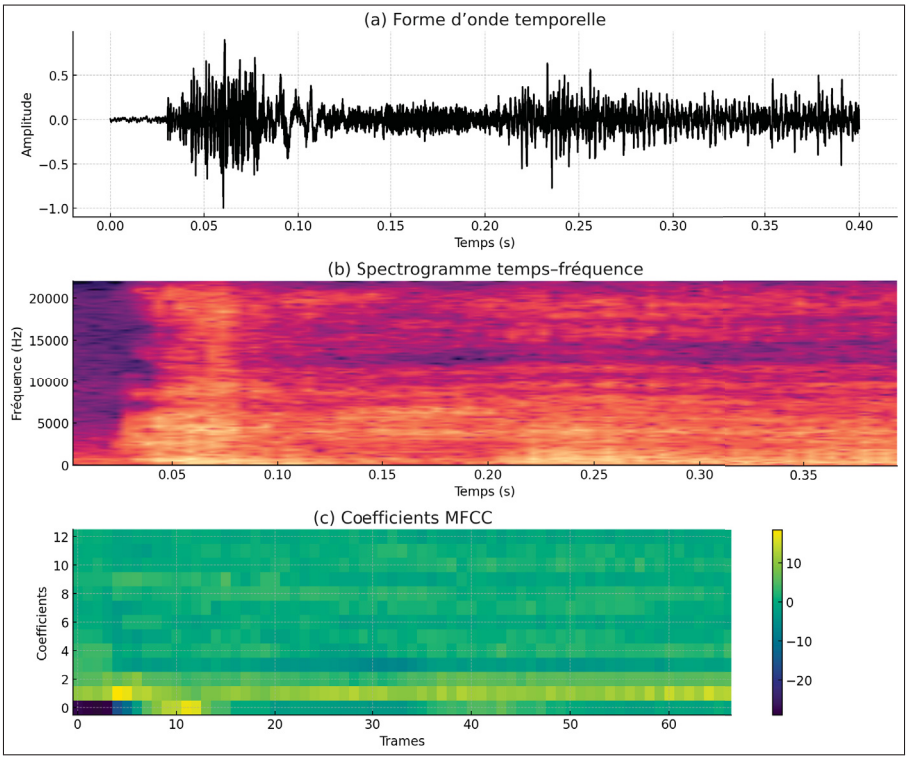


Figure 3.1 Exemples de représentations d'un son de toux. (a) Forme d'onde temporelle illustrant l'amplitude en fonction du temps ; (b) Spectrogramme temps–fréquence mettant en évidence la distribution énergétique du signal ; (c) Coefficients MFCC utilisés comme caractéristiques acoustiques pour l'analyse par apprentissage automatique

Dans cet exemple, on distingue nettement deux expirations rapprochées correspondant à deux quintes de toux successives. Le spectrogramme associé représente la distribution de l'énergie

acoustique en fonction du temps et de la fréquence, avec une échelle de couleurs allant des faibles intensités (bleu) aux fortes (jaune). On y observe les composantes fréquentielles propres à chaque quinte de toux ainsi que leur atténuation progressive. Enfin, la représentation MFCC (environ 13 coefficients par trame) illustre, sous forme de carte de chaleur, les coefficients cepstraux extraits du spectre filtré selon l'échelle de Mel. Cette représentation constitue une description compacte et discriminante de la signature fréquentielle de la toux.

3.2.2 Avantages et limites des caractéristiques acoustiques classiques

Les caractéristiques dites « classiques » présentent plusieurs atouts : elles sont simples à extraire, bien maîtrisées et souvent interprétables physiquement. Par exemple, un taux de passage par zéro (ZCR, *Zero Crossing Rate*) élevé indique un signal contenant davantage de hautes fréquences, tandis que l'énergie quadratique moyenne (RMS, *Root Mean Square*) reflète l'intensité moyenne du son. D'autres descripteurs spectraux, tels que le flux spectral (variation du spectre entre trames successives) et le *roll-off* spectral (fréquence sous laquelle se concentre une proportion donnée de l'énergie spectrale), fournissent des indications sur la texture et la rugosité du signal.

Ces descripteurs se révèlent efficaces pour des tâches de classification basiques sur des ensembles de données de taille modeste. Cependant, ils présentent également plusieurs limites. D'une part, ils sont sensibles au bruit et aux conditions d'enregistrement : un changement de dispositif ou de contexte sonore peut faire varier les valeurs extraites de manière non pertinente (par exemple, un bruit de fond peut artificiellement accroître l'énergie RMS ou perturber la forme du spectrogramme). D'autre part, la variabilité interindividuelle des toux (fréquence, intensité, timbre) complique l'établissement de seuils universels, car une toux « typique TB » peut différer considérablement d'un patient à l'autre.

Surtout, ces descripteurs manuels peinent à capturer des motifs complexes ou subtils propres à certaines pathologies. Les distinctions entre une toux tuberculeuse et d'autres toux peuvent résider dans des combinaisons d'attributs ou des dynamiques temporelles difficilement exprimables par quelques mesures linéaires. Ces limites ont motivé le recours à des modèles d'apprentissage

profond capables d'apprendre automatiquement des représentations hiérarchiques riches à partir des données brutes, sans passer par des *features* prédéfinies. Ces réseaux neuronaux profonds extraient des caractéristiques complexes et robustes, optimisant directement la séparation entre classes d'intérêt (par exemple TB vs non-TB), au lieu de s'appuyer sur des indices heuristiques.

3.3 Apports de l'intelligence artificielle dans le diagnostic par analyse de toux

3.3.1 Approches par traitement du signal

Historiquement, la détection assistée par ordinateur de maladies respiratoires à partir de sons (toux, respiration, phonocardiogrammes, etc.) reposait sur des approches classiques de traitement du signal. L'objectif consistait à extraire manuellement des caractéristiques jugées pertinentes fréquences dominantes, durées, enveloppes temporelles, puis à appliquer des règles déterministes ou des méthodes statistiques pour distinguer les signaux pathologiques. Par exemple, dans le cas de la toux, on pouvait filtrer le son pour isoler certaines bandes fréquentielles ou détecter des motifs temporels caractéristiques d'une toux productive versus sèche.

Ces méthodes, fondées sur l'expertise, présentaient l'avantage d'être interprétables et peu coûteuses en calcul. Toutefois, leur efficacité restait limitée par la variabilité des conditions réelles : bruit de fond, hétérogénéité des dispositifs d'enregistrement et diversité interindividuelle compromettaient la fiabilité de règles fixes. En pratique, le traitement du signal constituait un préalable à l'application de l'IA, transformant le signal brut en représentations numériques exploitables par des algorithmes. Ces représentations pouvaient être temporelles (forme d'onde), fréquentielles (spectre, cepstre) ou temps–fréquence (spectrogramme). L'essor des méthodes d'apprentissage automatique n'a pas supprimé ces techniques, mais les a intégrées dans des pipelines hybrides combinant prétraitement, extraction et classification (Kapetanidis et al., 2024).

3.3.2 Représentations audio pour la classification médicale

Pour l'analyse automatique des sons de toux, plusieurs représentations numériques sont couramment utilisées. Les coefficients cepstraux en fréquence Mel (MFCC) modélisent la perception auditive humaine en tenant compte de l'échelle de Mel (Rabiner and Schafer, 2011). Les spectrogrammes constituent une autre représentation clé : ils indiquent la répartition de l'énergie du signal au cours du temps et selon les fréquences, souvent exprimée en décibels. Cette visualisation permet d'identifier les motifs locaux (pics d'énergie, zones de silence, structure harmonique) caractéristiques de certaines pathologies. D'autres descripteurs complètent cette analyse : le ZCR, l'énergie RMS, le flux spectral et le *roll-off* spectral décrivent la régularité, la puissance et la texture du son.

Ces descripteurs présentent plusieurs avantages : ils sont faciles à calculer, interprétables et bien documentés. Cependant, ils sont sensibles au bruit, aux variations d'enregistrement et à la variabilité interindividuelle (intensité, timbre, morphologie respiratoire). En conséquence, leurs performances diminuent fortement dans des contextes réels ou bruités (Nidadavolu et al., 2020; Botha, 2021). Cette observation a conduit la recherche à privilégier des modèles d'apprentissage automatique et profond capables d'extraire directement des représentations hiérarchiques à partir des données brutes, plus robustes et mieux adaptées à la classification de la toux tuberculeuse.

3.4 Méthodes d'apprentissage automatique appliquées à l'audio médical

3.4.1 Approches classiques supervisées

Les premières tentatives d'analyse automatique des sons de toux ont souvent reposé sur des algorithmes d'apprentissage supervisé classiques. Parmi les plus utilisés, on retrouve les machines à vecteurs de support (SVM, *Support Vector Machine*) et les forêts d'arbres décisionnels aléatoires (*Random Forest*). Ces méthodes reçoivent en entrée des vecteurs de caractéristiques extraites manuellement (par exemple MFCC, ZCR, énergie, etc.) et apprennent à prédire la classe cible (toux pathologique ou saine) en fonction de ces caractéristiques.

Une SVM cherche à séparer les données dans l'espace des caractéristiques à l'aide d'un hyperplan optimal, maximisant la marge entre les deux classes. La forêt aléatoire, quant à elle, combine de multiples arbres de décision construits sur des sous-échantillons aléatoires des données afin de réduire la variance et d'améliorer la robustesse. Sur de petits ensembles de données et dans des conditions contrôlées, ces modèles peuvent atteindre des performances satisfaisantes pour détecter la tuberculose (TB) ou d'autres pathologies respiratoires à partir de la toux.

Certaines études pilotes, comme celle de Pahar et al. (2021), ont rapporté des précisions élevées en entraînant des classificateurs SVM ou Random Forest sur des jeux de données de taille limitée, bien annotés et enregistrés dans des conditions optimales. Cependant, dès que l'on cherche à généraliser à des contextes plus variés ou à des bases de données de grande taille, les modèles classiques montrent rapidement leurs limites. Leur capacité à capturer la complexité du signal dépend entièrement de la pertinence et de la complétude du jeu de caractéristiques choisi. Si une information diagnostique n'est pas explicitement codée dans les *features*, le modèle ne pourra pas l'inférer.

La distinction entre la toux d'un patient TB et celle d'une autre maladie respiratoire repose souvent sur des motifs acoustiques subtils, tels que des micro-pauses, des harmoniques spécifiques ou des dynamiques temporelles fines que les descripteurs simples ne capturent pas toujours. De plus, les modèles linéaires ou à base d'arbres peinent à modéliser les relations non linéaires complexes entre les caractéristiques et l'étiologie de la toux. Des travaux récents, tels que ceux de Botha (2021), ont montré que dans des enregistrements réalisés en conditions réelles (présence de bruit, microphones variés, port du masque, etc.), les performances des SVM ou des réseaux de neurones peu profonds chutent significativement, se situant en deçà des exigences cliniques, même après optimisation du jeu de *features*. Ces constats ont conduit la recherche à se tourner vers des approches plus puissantes et flexibles, en particulier les réseaux de neurones profonds, qui dominent désormais l'état de l'art.

3.4.2 Approches profondes

Les approches d'apprentissage profond (*deep learning*) ont profondément transformé le traitement audio médical. Les réseaux de neurones convolutionnels (CNN, *Convolutional Neural Networks*) appliqués aux spectrogrammes de toux se sont révélés particulièrement efficaces pour la détection de la TB. Pahar et al. (2022) ont comparé plusieurs architectures, dont CNN, LSTM et ResNet-50, sur des milliers d'enregistrements de toux couvrant des cas de TB, de COVID-19 et des individus sains. Ils ont obtenu un F1-score [1] de 0,93 pour la distinction TB/COVID-19 et de 0,86 pour une classification à trois classes.

Dans le domaine des réseaux séquentiels, Frost et al. (2022) ont montré qu'un modèle BiLSTM (*Bidirectional Long Short-Term Memory*) enrichi d'un mécanisme d'attention améliore la généralisation et offre une interprétabilité accrue. L'attention permet au modèle d'identifier les régions temporelles les plus discriminantes dans un signal de toux, facilitant la visualisation des zones contribuant le plus à la décision.

À plus grande échelle, Suda (2023b) ont proposé un modèle hybride combinant un CNN 2D et XGBoost, entraîné sur plus de 720 000 échantillons de toux collectés dans plusieurs pays. Le système atteint un score AUROC [2] de 0,88, dépassant les recommandations de l'OMS pour les outils de triage de la TB.

De même, Rajasekar et al. (2024) ont comparé différentes architectures, incluant les *Capsule Networks* et les réseaux entièrement connectés (FCNN, *Fully Connected Neural Network*), sur près de 10 000 enregistrements. Leur modèle hybride a obtenu des résultats remarquables : précision de 0,97, sensibilité de 0,98, spécificité de 0,96 et F1-score de 0,97. Enfin, une méta-analyse récente de Sahoo et al. (2025) a compilé les performances de divers modèles d'IA audio pour la détection de la TB, montrant des valeurs moyennes d'AUC d'environ 0,95 et un rapport de cotes diagnostique supérieur à 80, confirmant le potentiel de ces approches pour le

[1] Le F1-score est la moyenne harmonique entre la précision et le rappel. Il mesure la qualité globale d'un classificateur binaire.

[2] L'AUROC (*Area Under the Receiver Operating Characteristic*) évalue la capacité du modèle à distinguer deux classes. Une valeur proche de 1 indique une excellente performance.

dépistage clinique. Les modèles profonds, qu'ils soient convolutionnels, récurrents ou hybrides, atteignent aujourd'hui des performances diagnostiques élevées et démontrent leur supériorité sur les approches classiques.

3.4.3 Mécanismes d'attention et architectures hybrides

L'introduction des mécanismes d'attention a encore renforcé les performances des modèles audio récents. Le principe de l'attention consiste à pondérer l'importance relative des différentes parties d'une séquence lors de la prédiction. Appliqué aux sons de toux, un modèle attentionnel apprend à se concentrer sur les segments les plus informatifs, comme les pics sonores, et à ignorer les parties non pertinentes telles que les silences ou le bruit de fond.

Le modèle de transformeur proposé par Vaswani et al. (2017) a popularisé l'attention multi-têtes (*multi-head attention*), capable de capter simultanément des dépendances locales et globales dans une séquence. En traitement audio, des modules de *self-attention* ou de *cross-attention* ont été intégrés à des architectures CNN–RNN pour mieux exploiter l'information temporelle. Imran et al. (2020) ont montré que l'ajout d'une couche d'attention par-dessus un modèle CNN–BiGRU (*Bidirectional Gated Recurrent Unit*) améliore la détection des cas positifs difficiles en valorisant des indices acoustiques faibles mais pertinents.

Des architectures hybrides combinant convolution, récurrence et attention ont ensuite été proposées afin de tirer parti des atouts de chacune. Imran et al. (2022) décrivent un pipeline typique comprenant plusieurs couches CNN pour extraire les caractéristiques locales d'un spectrogramme, suivies d'une couche Bi-GRU condensant ces caractéristiques dans le temps, et enfin d'un bloc d'attention mettant en évidence les *frames* temporelles les plus pertinentes avant la décision finale. Ce type d'architecture a démontré des performances accrues dans les tâches de dépistage automatique de la TB.

Un modèle représentatif est celui de *MetforNet*, proposé par l'équipe *Metformin* dans le cadre du *CODA TB DREAM Challenge* (2023). Ce modèle intègre cinq blocs convolutionnels, un Bi-GRU et une couche d'attention multi-têtes, atteignant un score AUC d'environ 0,74 sur des

données de toux enregistrées en conditions réelles (Jaganath et al., 2024). L'attention permet ici non seulement d'améliorer les performances, mais aussi d'accroître l'interprétabilité du modèle en identifiant les parties du signal ayant contribué au diagnostic.

3.5 Architectures de modèles appliquées à l'audio

3.5.1 Réseaux CNN et GRU pour l'audio

Comme évoqué précédemment, les réseaux convolutifs (CNN) et récurrents (GRU, *Gated Recurrent Unit*) constituent le cœur des meilleures architectures pour l'analyse de sons médicaux. De nombreuses variantes ont été explorées pour optimiser la modélisation des toux.

Sur la partie CNN, certaines études ont proposé des convolutions unidimensionnelles (1D) directement sur le signal audio brut ou sur les coefficients MFCC, tandis que d'autres privilégient des convolutions bidimensionnelles (2D) appliquées aux spectrogrammes. L'avantage des CNN 1D est de modéliser la forme d'onde sans étape de transformation temps–fréquence, comme l'ont expérimenté Orlandic et al. (2021) pour la détection de la COVID-19 via la toux. Cependant, la majorité des études sur la TB utilisent des représentations temps–fréquence (spectrogrammes ou scalogrammes), qui rendent les motifs acoustiques plus explicites.

Des architectures profondes préentraînées sur des bases audio génériques, telles que VGGish ou ResNet-50 (préentraînées sur AudioSet), ont été testées en transfert d'apprentissage pour la TB, avec des succès variables selon la similarité acoustique entre les sons généraux et les toux pathologiques (Yadav et al., 2024). Côté RNN, l'utilisation de variantes bidirectionnelles (Bi-GRU, Bi-LSTM) est fréquente, car elles tiennent compte du contexte passé et futur d'une séquence sonore, ce qui permet de mieux modéliser la dynamique complète d'une toux. Par exemple, un Bi-GRU peut apprendre qu'une forte explosion sonore suivie d'une traînée de bruit correspond à une toux complète, tandis qu'un pic isolé pourrait être un bruit externe.

Les modèles récents intègrent également des mécanismes de *gating* sophistiqués pour filtrer l'information pertinente à chaque instant, ce qui est essentiel pour des signaux longs comportant

des segments non informatifs. En somme, les architectures CNN+GRU représentent un compromis efficace entre capacité de modélisation et coût computationnel. Elles capturent à la fois les caractéristiques locales du signal et son organisation temporelle, tout en demeurant légères comparativement aux transformeurs entièrement attentifs.

3.5.2 Mécanismes d'attention intégrés

Les mécanismes d'attention sont désormais intégrés de manière quasi systématique aux architectures CNN/RNN appliquées à l'audio. Deux grandes approches sont distinguées : l'attention temporelle et l'attention multimodale.

Dans l'attention temporelle, une couche d'attention est appliquée aux sorties d'un RNN pour pondérer chaque pas de temps selon son importance. Par exemple, dans un enregistrement comportant plusieurs quintes de toux, la couche d'attention peut attribuer des poids élevés aux segments où la toux est la plus marquée et de faibles poids aux zones silencieuses. Ce principe a été utilisé avec succès dans le cadre du *CODA TB DREAM Challenge*, améliorant le rappel des modèles sur des signaux longs contenant peu de toux effectives (Huddart et al., 2024). Techniquement, un vecteur contexte est calculé comme combinaison pondérée des états cachés du RNN, les poids d'attention étant produits par un petit réseau *feed-forward* entraîné conjointement (Vaswani et al., 2017). Cette approche a permis au modèle *MetforNet* d'atteindre les meilleures performances du concours en filtrant les informations non pertinentes du signal (Jaganath et al., 2024).

L'attention multimodale, quant à elle, vise à fusionner des sources d'information hétérogènes, par exemple l'audio et les métadonnées cliniques. Yadav et al. (2024) ont introduit un module de *cross-attention* bidirectionnelle où chaque modalité interagit avec l'autre pour affiner ses représentations internes. Cette technique illustre la capacité de l'attention à intégrer de façon cohérente des données issues de sources diverses.

Enfin, l'attention offre un avantage majeur en matière d'explicabilité. Les cartes de poids d'attention peuvent être visualisées comme un spectrogramme du « focus » du modèle, indiquant

quelles parties du signal ont le plus influencé la prédiction (TB ou non-TB). Des études récentes, comme celle de Zhang et al. (2022), exploitent cette propriété pour vérifier que le modèle se base sur des éléments acoustiques cliniquement pertinents, renforçant ainsi la confiance dans les outils de diagnostic fondés sur l'IA.

Ces avancées ouvrent la voie à des modèles multimodaux combinant les signaux de toux et les informations patient, une thématique abordée dans la section suivante.

3.6 Présentation du challenge CODA TB DREAM

3.6.1 Objectifs et données fournies

Le challenge CODA TB (Cough Diagnostic Algorithm for TB) du programme DREAM a été lancé en 2022 pour accélérer le développement de modèles d'IA de dépistage de la tuberculose à partir de sons de toux. Le concours, international, mettait à disposition un large jeu de données de toux accompagnées de métadonnées cliniques afin que les équipes conçoivent, en quatre mois, des algorithmes performants et comparables. Les données proviennent d'une étude prospective incluant 2 143 adultes présentant une toux depuis au moins deux semaines, recrutés dans des dispensaires de sept pays (Inde, Madagascar, Philippines, Afrique du Sud, Tanzanie, Ouganda et Vietnam) (Jaganath et al., 2024).

Pour chaque participant, au moins trois toux sollicitées ont été enregistrées via l'application mobile Hyfe Research, fournissant le signal audio d'entrée. Parallèlement, des variables cliniques et démographiques étaient collectées (âge, sexe, exposition, présence d'autres symptômes tels que fièvre ou perte de poids), susceptibles d'améliorer le triage. Le statut TB patient-niveau était établi par les tests de référence Xpert MTB/RIF Ultra et la culture, garantissant des étiquettes de vérité terrain de bonne qualité. L'objectif principal était de classer, pour chaque patient, TB versus non-TB à partir des enregistrements de toux, avec des performances compatibles avec un test de triage communautaire : maximiser la sensibilité tout en conservant une spécificité suffisante pour limiter les faux positifs. Deux sous-défis reflétaient ces usages : un volet « audio

seul » et un volet « audio + données cliniques ». Les organisateurs ont fourni un ensemble d'entraînement librement exploitable et conservé un ensemble de test masqué pour l'évaluation indépendante, assurant l'équité des comparaisons (Jaganath et al., 2024).

3.6.2 Règles et critères d'évaluation

Les équipes devaient soumettre un algorithme final empaqueté dans un conteneur exécutable (Docker) applicable au jeu de test sans intervention manuelle. Le conteneur devait inclure le modèle entraîné et l'intégralité de la chaîne de traitement (prétraitement du signal, extraction de caractéristiques, inférence) afin de garantir la reproductibilité et une évaluation en aveugle sur l'infrastructure des organisateurs (Jaganath et al., 2024). Les sorties attendues comprenaient idéalement un score continu par patient (risque de TB), permettant le calcul de courbes ROC, en plus d'une décision binaire.

Le critère de classement principal était l'aire sous la courbe ROC (AUC) mesurée sur l'ensemble de test. Pour coller aux objectifs de dépistage, une AUC partielle a également été considérée dans la zone de sensibilité prioritaire. Conformément aux cibles de l'OMS pour un test de triage TB (sensibilité minimale 90 %, spécificité minimale 70 % (World Health Organization, 2023b)), Le challenge a retenu une pAUC, définie comme l'aire sous la courbe ROC restreinte à la plage de sensibilités comprise entre 80 % et 100 %. Cette métrique permet de privilégier les modèles efficaces dans une zone critique pour le dépistage, où la réduction des faux négatifs est prioritaire. Un seuil minimal de 60 % de spécificité à 80 % de sensibilité était également requis. Les modèles étaient d'abord filtrés par pAUC, puis classés selon leur AUC globale. Cette métrique favorise les approches capturant un maximum de cas TB (peu de faux négatifs) tout en maintenant un niveau de faux positifs acceptable. En pratique, 11 systèmes « audio seul » et 6 systèmes « audio + clinique » ont été soumis par les finalistes (Jaganath et al., 2024). L'évaluation des performances (AUC, pAUC, sensibilité/spécificité à des points cibles) a été réalisée de manière indépendante sur des données jamais vues, avec restitution d'un classement et de scores détaillés à chaque équipe. Des analyses complémentaires ont examiné la robustesse

par sous-groupes (pays, sexe, statut VIH), afin d'identifier d'éventuels biais et d'apprécier la généralisabilité des modèles (Jaganath et al., 2024).

3.6.3 Méthodes testées (CRNN, DMRNet, approches hybrides)

Les meilleures équipes du challenge CODA TB DREAM ont convergé vers des architectures profondes de type CRNN : des réseaux convolutifs (pour extraire des motifs locaux sur spectrogrammes) suivis de couches récurrentes bidirectionnelles (Bi-GRU/Bi-LSTM) pour modéliser la dynamique des quintes, puis d'un mécanisme d'attention temporelle qui met en avant les trames informatives et atténue les segments silencieux ou bruités. Dans l'évaluation indépendante du challenge, les modèles « audio seul » présentent des AUROC compris entre 0,69 et 0,74 ; au point opérationnel de 80 % de sensibilité, la meilleure soumission atteint 55,5 % de spécificité (IC 95 % : 47,7–64,2) (Jaganath et al., 2024). La disponibilité d'un corpus multi-pays patient-niveau (733 756 toux, 2 143 participants) a permis d'entraîner ces modules d'attention dans des conditions proches du terrain (Huddart et al., 2024).

En parallèle des CRNN, plusieurs équipes ont exploré des pipelines « à caractéristiques » : extraction d'un grand jeu de descripteurs acoustiques (par exemple familles MFCC/énergie/flux/roll-off), agrégation statistique par fenêtre ou par patient, réduction de dimension (ACP) puis apprentissage avec des classifieurs non linéaires (forêts aléatoires, gradient boosting). Sur des données réelles collectées par application mobile, ce type d'approche atteint typiquement un AUC d'environ 0,70 (±0,05) avec l'audio seul, porté à ≈0,81 (±0,05) lorsque l'on ajoute des métadonnées cliniques simples, illustrant l'intérêt des schémas multimodaux (Kafentzis et al., 2023).

Des architectures plus avancées ont également été proposées dans la littérature et ont inspiré certaines soumissions. DMRNet introduit des convolutions dynamiques (filtres modulés par l'entrée) pour affiner l'extraction de caractéristiques, ainsi qu'un double mécanisme d'attention, attention polarisée locale puis attention multi-têtes globale, afin d'améliorer l'agrégation

temporelle. Le modèle rapporte des performances élevées sur un jeu contrôlé (AUC > 0,90), bien qu'il n'ait pas été directement évalué sur le corpus du challenge (Xu et al., 2023).

Un certain nombre de leviers transversaux se retrouvent parmi les meilleures pratiques :

(i) une augmentation de données ciblée (bruit additif réaliste, variations de gain, décalage temporel, convolution par réponses impulsionnelles) afin d'améliorer la robustesse au contexte d'enregistrement ;

(ii) une validation croisée « par sujet » (par exemple, *Stratified Group K-Fold*) pour éviter toute fuite entre l'entraînement et le test ;

(iii) une calibration du seuil de décision afin d'atteindre la sensibilité cible (80 % pour le classement primaire du challenge), en complément de l'AUC et de la pAUC définie dans la zone de haute sensibilité (Jaganath et al., 2024).

Au total, les approches les plus performantes combinent un encodeur convolutionnel, une couche récurrente bidirectionnelle, un bloc d'attention, et, lorsque disponible, la fusion avec des métadonnées patient pour rapprocher les performances des cibles de triage fixées par les lignes directrices.

Tableau 3.1 Comparaison synthétique d'approches pour la détection de la tuberculose à partir de la toux

Approche	**Modèle**	**AUC**	**Sens. @80% spéc.**	**Spéc. @80% sens.**
Features + SVM	MFCC + SVM	0.70	0.80	0.62
Features + RF	MFCC + RF	0.72	0.81	0.64
CNN spectrogramme	CNN 2D simple	0.78	0.85	0.70
CNN + BiGRU	CNN 2D + BiGRU	0.82	0.87	0.74
CNN + BiGRU + Attention	MetforNet	0.84	0.89	0.78

3.6.4 Limitations identifiées

Malgré des progrès tangibles, plusieurs limites persistent. Le bruit de fond et la qualité hétérogène des enregistrements constituent un frein majeur en conditions réelles, les toux étant captées via téléphones dans des environnements non contrôlés (Jaganath et al., 2024). La variabilité

des protocoles de collecte complique la généralisabilité, certains sites ayant procédé en milieu clinique supervisé et d'autres en auto-enregistrement (Jaganath et al., 2024).

Le déséquilibre des classes est un autre point critique. Dans le challenge, la cohorte de suspects présentait une proportion de cas TB confirmés plus élevée qu'en dépistage de population. Les équipes ont donc utilisé pondérations de classes et échantillonnages stratégiques. En pratique, une bonne AUC peut coexister avec trop de faux positifs lorsque la prévalence réelle est faible, d'où l'importance d'un réglage de seuil et d'un test de confirmation. S'ajoute la question de la vérité terrain : Xpert MTB/RIF Ultra et la culture offrent une référence robuste, mais des cas paucibacillaires peuvent rester non confirmés, introduisant du bruit d'annotation. Par ailleurs, d'autres étiologies de toux non-TB ne sont pas distinguées de manière systématique, ce qui peut dégrader la spécificité si la diversité diagnostique est insuffisante (Jaganath et al., 2024).

Enfin, l'estimation honnête des performances reste délicate. Des écarts entre validation interne et scores sur le test indépendant suggèrent du sur-apprentissage ou une spécialisation à certains contextes. Les schémas de validation sujets-indépendants recommandés réduisent ce risque, sans le supprimer totalement. À terme, des études prospectives en déploiement réel sur des populations et contextes nouveaux seront nécessaires (Jaganath et al., 2024).

3.7 Positionnement de notre approche

Au regard de l'état de l'art et des contraintes de terrain, notre approche vise à concilier performance et robustesse pour le triage de la tuberculose à partir de la toux. Elle s'appuie sur une chaîne cohérente qui commence par une extraction audio optimisée combinant des représentations temps–fréquence (notamment des spectrogrammes Mel en décibels) avec une segmentation stricte des épisodes de toux et une agrégation au niveau patient lorsque plusieurs enregistrements sont disponibles. Les métadonnées cliniques normalisées (âge, sexe, symptômes, facteurs de risque) sont intégrées de façon explicite après encodage homogène, traitement des valeurs manquantes et standardisation, de manière à exploiter un signal multimodal pertinent pour le triage. Le cœur du modèle repose sur une architecture profonde hybride de type

CNN–BiGRU dotée d'un mécanisme d'attention temporelle, qui capte à la fois les motifs locaux et la dynamique globale des quintes, avec une tête de classification calibrée pour des décisions stables à l'usage. L'évaluation utilise une validation croisée par groupes stratifiée (*StratifiedGroupKFold*), qui préserve la proportion des classes tout en séparant les données par sujet afin d'assurer l'indépendance entre l'entraînement et la validation. Cette méthode limite les fuites de données, complétée par un arrêt anticipé pour réduire le sur-apprentissage. Les hyperparamètres sont ajustés par optimisation bayésienne (par exemple avec Optuna) en ciblant des objectifs alignés sur le triage, tels que la pAUC entre 80 % et 100 % de sensibilité et la spécificité au point 80 % de sensibilité, tandis que le seuil de décision est réglé pour respecter ces cibles. Le déséquilibre de classes est pris en compte par des pondérations adaptées, des pertes tolérantes au déséquilibre et, le cas échéant, un échantillonnage au niveau sujet afin de préserver la sensibilité en faible prévalence. Enfin, les sorties segmentaires sont agrégées et calibrées au niveau patient (vote, moyenne de scores, calibration isotone ou de Platt) et l'influence des enregistrements aberrants est réduite par la détection des segments non pertinents (contrôles d'énergie et de durée, détection d'activité vocale) et l'atténuation des outliers acoustiques. L'ensemble de ces choix vise à rapprocher les performances des cibles de triage, tout en restant compatible avec des enregistrements mobiles et des contextes d'acquisition hétérogènes.

3.8 Conclusion

La revue a montré que les représentations acoustiques classiques demeurent utiles mais atteignent rapidement leurs limites face à la variabilité des conditions réelles d'enregistrement. Les approches profondes dominent désormais l'analyse de la toux : les réseaux convolutionnels appliqués aux spectrogrammes extraient des motifs locaux discriminants, les réseaux récurrents modélisent la dynamique des quintes et les mécanismes d'attention renforcent la sélection des segments informatifs. La fusion de l'audio avec des métadonnées cliniques améliore la spécificité et la robustesse, en particulier lorsque la prévalence est faible et que les enregistrements sont bruyants ou hétérogènes. Les enseignements du challenge CODA TB ont confirmé l'intérêt de l'augmentation de données, de la validation croisée par sujet, de la calibration des sorties

et d'une évaluation indépendante, tout en mettant en évidence des marges de progrès sur la sensibilité au bruit, l'hétérogénéité d'appareils et de contextes, le déséquilibre de classes et la généralisabilité inter-populations. Dans cette continuité, l'approche proposée combine encodeurs convolutionnels, récurrence, attention et fusion multimodale, avec une évaluation rigoureuse orientée triage et des mécanismes de calibration et d'agrégation au niveau patient, afin d'améliorer la fiabilité du dépistage automatique de la tuberculose à partir de la toux.

CHAPITRE 4

MÉTHODOLOGIE

Ce chapitre détaille la méthodologie adoptée pour concevoir et mettre en œuvre un système automatisé de détection de la tuberculose à partir de sons de toux. Le développement de cette approche s'appuie sur le corpus fourni dans le cadre du challenge international CODA TB DREAM 2022, ainsi que sur les travaux récents de la détection de tuberculose. Cette méthodologie expérimentale est rigoureusement structurée : elle intègre plusieurs étapes successives allant de la constitution du jeu de données à l'évaluation du modèle. L'objectif est de garantir la robustesse, la fiabilité et la reproductibilité des résultats obtenus, tout en maintenant une rigueur scientifique exemplaire.

4.1 Données utilisées

La première étape de notre méthodologie concerne la description et la préparation du jeu de données. Celui-ci comprend à la fois des enregistrements audios de toux et des métadonnées associées, comme détaillé ci-dessous.

4.1.1 Protocole de collecte des données

Les données utilisées proviennent du *CODA TB DREAM Challenge 2022*, ce projet vise à encourager la recherche ouverte en santé mondiale à travers la mise à disposition publique de jeux de données de haute qualité sur la tuberculose.

L'étude inclut des adultes ($\geq$ 18 ans) se présentant en consultation externe avec une toux persistante ($\geq$ 2 semaines). Pour chaque participant, les informations cliniques et démographiques ont été recueillies à l'inclusion, suivies d'un bilan standard de dépistage de la tuberculose (TB). Le statut TB est établi à partir des résultats microbiologiques de référence, tests PCR *Xpert MTB/RIF* (ou *Ultra*) et culture bactérienne. Les enregistrements de sons de toux ont été réalisés avant toute initiation de traitement, selon un protocole harmonisé entre les différents sites d'étude.

Un sous-échantillon a également enregistré des toux longitudinales pendant deux semaines via smartphone, destinées exclusivement à l'entraînement des modèles. La collecte des toux sollicitées a été effectuée à une distance fixe à l'aide d'un smartphone Android utilisant l'application *Hyfe Research*. Cette application capture des clips de 0,5 s d'événements acoustiques explosifs et les étiquette comme « toux » lorsque le score de prédiction est supérieur ou égal à 0,85. Les participants étaient instruits de tousser au moins trois fois ; toute toux spontanée déclenchée à la suite d'une toux sollicitée était également enregistrée. L'évaluation finale repose uniquement sur les toux sollicitées, les enregistrements longitudinaux servant exclusivement à l'entraînement.

Les fichiers sources sont hébergés sur la plateforme *Synapse*, qui assure la conservation, l'anonymisation et la traçabilité des données. L'auteur tient à exprimer sa reconnaissance envers les équipes de recherche et les participants ayant contribué à la collecte, à la validation et au partage de ces données, essentielles à l'avancement du dépistage automatisé de la tuberculose.

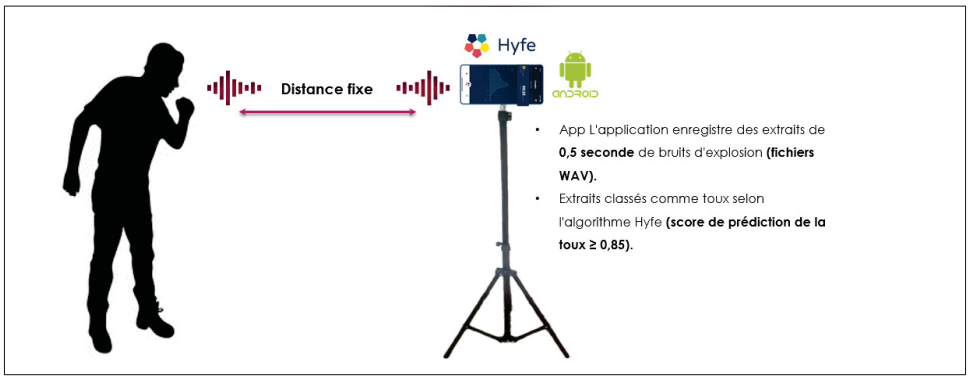


Figure 4.1 Protocole de collecte des toux via l'application Hyfe Research

4.1.2 Données audio (fichiers de toux)

Les enregistrements de toux constituent la base essentielle de notre système de détection. Ils proviennent du corpus officiel du challenge CODA TB DREAM 2022, qui regroupe plusieurs milliers d'échantillons de toux volontairement collectés dans des conditions contrôlées. Chaque enregistrement est encodé au format WAV avec un taux d'échantillonnage de 16 kHz et présente

une durée moyenne comprise entre 0,5 et 1 seconde, correspondant uniquement à la phase explosive de la toux. Ce choix de format non compressé garantit une qualité sonore élevée, indispensable à une analyse acoustique précise des signaux.

Une analyse préliminaire par visualisation des formes d'onde illustrée à la Figure 4.2, qui compare la forme d'onde d'une toux saine à celle d'une toux tuberculeuse met en évidence des différences notables entre ces deux types de signaux. Dans la majorité des enregistrements, on distingue des écarts perceptibles même à l'oreille humaine en termes d'intensité, de régularité du signal et de structure temporelle des impulsions sonores.

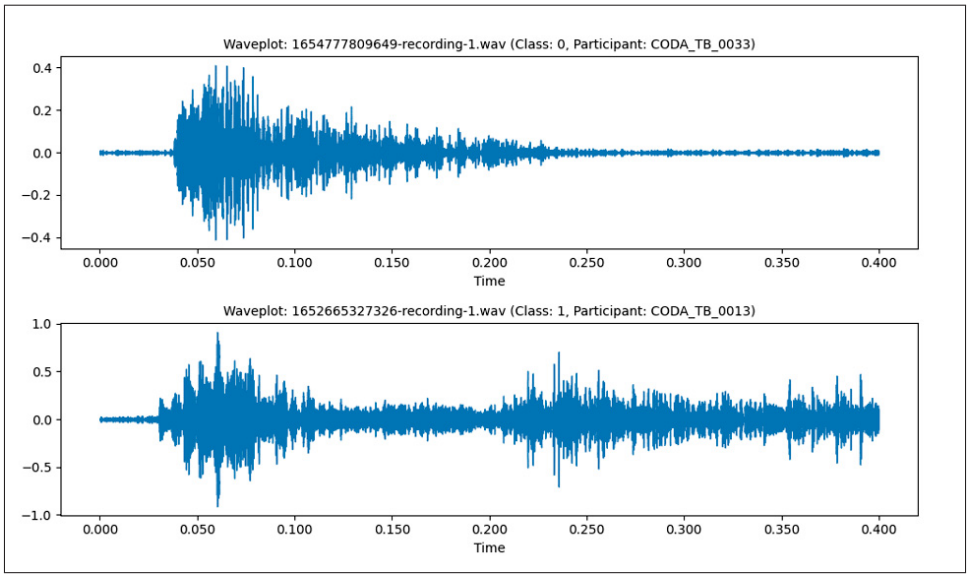


Figure 4.2 Comparaison visuelle des formes d'onde d'une toux saine (en haut) et d'une toux tuberculeuse (en bas). Les différences d'intensité, de régularité et de structure temporelle illustrent la complexité de la classification des toux

4.1.3 Métadonnées cliniques et démographiques

En parallèle des signaux audio, nous disposons pour chaque participant d'un ensemble de métadonnées cliniques et démographiques sous forme tabulaire. Celles-ci comprennent des variables sociodémographiques (âge, sexe, pays), des informations cliniques (poids, taille, symptômes déclarés) ainsi que diverses données contextuelles pertinentes.

Ces informations complémentaires sont exploitées dans une approche multimodale afin d'enrichir le modèle. En effet, la fusion des métadonnées avec les caractéristiques audio permet de capturer d'éventuelles interactions entre les paramètres acoustiques d'une toux et les facteurs physiologiques ou contextuels propres au patient. Par exemple, l'efficacité de détection peut varier selon l'âge ou le profil médical du sujet, d'où l'intérêt d'inclure ces données dans le modèle.

Tableau 4.1 Dictionnaire des métadonnées cliniques et démographiques

Nom technique	**Définition (FR)**
participant_id	Identifiant unique de chaque participant à l'étude
sex_birth	Sexe à la naissance déclaré par le participant
age_years	Âge au moment de la collecte (calculé à partir de la date de naissance si connue, sinon âge rapporté)
height_cm	Taille mesurée en centimètres
weight_kg	Poids mesuré en kilogrammes
cough_duration_days	Durée auto-rapportée de la toux actuelle, en jours
tb_prior	Antécédent déclaré de tuberculose (a déjà eu ou a été informé avoir eu la TB)
tb_prior_type	Type d'antécédent de TB (pulmonaire, extrapulmonaire, inconnu ; choix multiples possibles)
hemoptysis_30d	Hémoptysie au cours des 30 derniers jours (a craché du sang)
heart_rate_bpm	Fréquence cardiaque mesurée à l'inclusion (battements par minute)
temperature_c	Température corporelle mesurée à l'inclusion (°C)
weight_loss_30d	Perte de poids auto-rapportée au cours des 30 derniers jours

Suite à la page suivante.

Suite du tableau 4.1.

Nom technique	**Définition (FR)**
tobacco_or_vape_7d	Usage de tabac combustible et/ou vapotage au cours des 7 derniers jours
fever_30d	Fièvre auto-rapportée au cours des 30 derniers jours
night_sweats_30d	Sueurs nocturnes auto-rapportées au cours des 30 derniers jours
country	Pays d'inclusion (PH, VN, SA, UG, IN, MG, TZ)
cough_collected_baseline	Indique si des enregistrements de toux ont été collectés à l'inclusion (oui/non)
n_cough_sounds_baseline	Nombre d'enregistrements de toux collectés à l'inclusion
hiv_status	Statut VIH (positif déclaré ou test positif, négatif au test, ou inconnu/refus)
tb_ref_standard	Statut de référence TB basé sur les résultats Xpert/Xpert Ultra sur expectoration (positif, négatif, indéterminé)
xpert_ultra_semiquant	Plus haut grade semi-quantitatif Xpert Ultra au baseline (trace, très faible, faible, moyen, élevé ; disponible seulement si test positif)
tb_status	Statut TB analytique dérivé des résultats Xpert/Xpert Ultra (positif, négatif, indéterminé)

4.1.4 Alignement et validation de la cohérence

Avant de passer à l'étape d'apprentissage automatique, les données ont été soigneusement alignées afin d'assurer la correspondance exacte entre chaque enregistrement audio et ses métadonnées. Chaque fichier de toux est identifié par un identifiant unique, facilitant la jointure et la vérification de cohérence entre les deux sources.

Un contrôle de qualité a ensuite permis d'écarter les enregistrements jugés incomplets ou incohérents, notamment ceux dont la durée était insuffisante, les métadonnées manquantes ou les identifiants non appariés. Ces vérifications, réalisées automatiquement à l'aide de scripts dédiés, ont permis de constituer un jeu de données final cohérent, homogène et prêt pour les étapes de prétraitement décrites en section 4.2.

4.2 Prétraitement des données

Le prétraitement vise à homogénéiser les entrées, atténuer les artefacts d'acquisition et préparer des descripteurs audio et tabulaires fiables pour l'apprentissage. Pour prévenir toute fuite d'information, tous les modules qui "apprennent" des données (seuils, normalisations, détecteurs d'anomalies, imputations) sont ajustés exclusivement sur les partitions d'entraînement au sein de chaque pli de validation croisée (Wong, 2015), puis réappliqués à l'identique aux partitions de validation et au jeu de test. Cette discipline méthodologique garantit que les performances estimées reflètent la capacité de généralisation du modèle, et non un effet d'optimisme dû à un calibrage sur l'ensemble complet.

4.2.1 Nettoyage et normalisation des signaux audio

La phase explosive de la toux contient l'information diagnostique majeure. Le prétraitement vise à améliorer la qualité acoustique tout en préservant l'intégrité temporelle du signal.

Un filtrage de base, adapté au contexte, est appliqué afin de supprimer les basses fréquences indésirables et le bourdonnement secteur, tout en conservant la bande informative principale du signal. Un passe-bande de 50 à 6000 Hz a été retenu, plage généralement utilisée dans l'analyse des sons respiratoires, car elle couvre la majorité du contenu énergétique utile tout en éliminant les fréquences très basses (souvent liées aux bruits de manipulation ou au flux d'air) et les hautes fréquences dominées par le bruit électronique (Serrurier et al., 2022).

Lorsque nécessaire, les enregistrements sont ensuite normalisés en amplitude afin d'harmoniser les niveaux sonores issus de conditions d'enregistrement différentes (Husssain et al., 2024). Un

léger recadrage temporel (trim) est enfin appliqué pour supprimer les silences en début et fin d'enregistrement.

Ces étapes filtrage léger, normalisation harmonisée et recadrage suffisent à stabiliser les données tout en préservant la dynamique pertinente pour la détection acoustique.

4.2.2 Détection et exclusion des enregistrements aberrants

Même après nettoyage et normalisation, une hétérogénéité notable des signaux subsiste et aucun patron de toux clair ne se dégage voir (Figure 4.3). Afin de limiter l'impact de ces enregistrements atypiques, nous appliquons une détection non supervisée par Isolation Forest (Liu et al., 2008). Le taux de contamination est fixé à 7 %, et l'ajustement est réalisé uniquement sur la partition d'entraînement de chaque pli. Le seuil appris est ensuite figé et appliqué tel quel aux partitions de validation et de test, ce qui prévient toute fuite d'information.

En pratique, cette procédure conduit à l'exclusion d'environ 140 fichiers par pli, un ordre de grandeur relativement stable. Ce nombre est légèrement inférieur au 7 % théorique, en raison du caractère plus conservateur du seuil transféré hors apprentissage et des différences de distribution entre partitions. Toutes les décisions sont consignées (identifiant, motif synthétique, valeurs de métriques), et la stabilité des répartitions par classe et par participant est systématiquement contrôlée.

À l'issue de cette étape, l'ensemble d'entraînement est recomposé uniquement à partir des fichiers retenus. La suite du pipeline (segmentation, extraction de caractéristiques, normalisation, augmentation et apprentissage) est alors appliquée exclusivement sur ce corpus filtré.

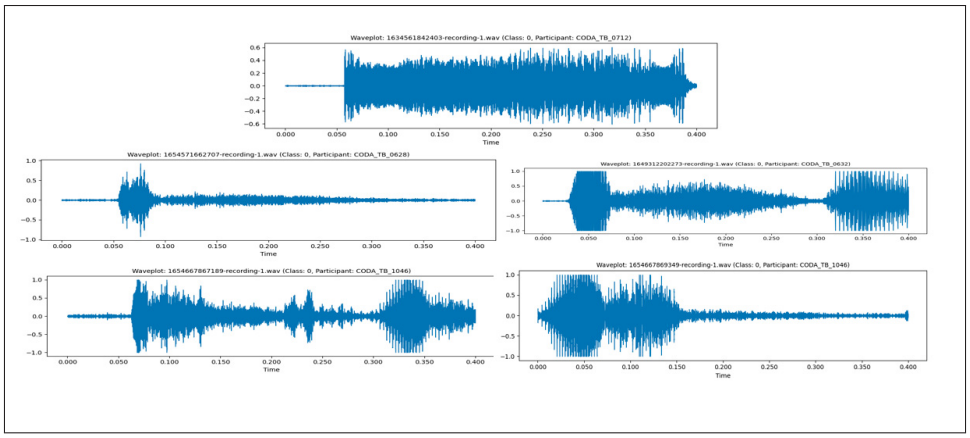


Figure 4.3 Signaux de toux extraits après nettoyage et normalisation

4.2.3 Prétraitement des métadonnées

Le traitement des variables cliniques et démographiques a pour finalité de fiabiliser la fusion avec les descripteurs audio. Un contrôle d'harmonisation est d'abord effectué pour vérifier les unités (p. ex. taille en cm vs m), détecter les valeurs hors domaine (extrêmes physiologiquement improbables) et regrouper les modalités rares (fréquences < 1–2 %) afin de limiter la sparsité en encodage. La gestion des valeurs manquantes s'appuie sur une imputation apprise sur l'entraînement : médiane pour les variables continues et modalité majoritaire pour les catégorielles, avec possibilité d'un indicateur « valeur manquante » pour les champs critiques. Les estimateurs d'imputation sont conservés et réappliqués sans ré-apprentissage à la validation et au test.

Les variables continues sont ensuite standardisées (Z-score, moyenne 0, écart-type 1) et les catégorielles encodées en one-hot, le dictionnaire de colonnes étant déduit du train puis gelé pour garantir la compatibilité structurelle entre splits. Un contrôle de colinéarité (corrélations fortes, VIF) peut être mené afin de réduire l'instabilité des estimateurs au moment de la fusion audio–tabulaire. Toutes les transformations sont systématiquement tracées (versions, graines pseudo-aléatoires, correspondance des catégories), ce qui facilite la reproductibilité et

l'audit méthodologique. Le Tableau 4.2 présente le dictionnaire des variables retenues pour l'entraînement, après traitement et sélection sur la base des corrélations.

Tableau 4.2 Schéma d'encodage et remarques de préparation des métadonnées

Variable	**Encodage**	**Remarques**
Participant (ID)	Non utilisé (stratification groupe)	Utilisé pour GroupKFold/StratifiedGroupKFold (Scikit-learn developers, 2011) ; contrôle fuite entre sujets
Sexe	One-hot (catégorie 'Autre' incluse)	Harmoniser libellés (e.g., Female→F, Male→M)
Âge	Z-score (ou robust scaler si queues lourdes)	Vérifier cohérence (date naissance/collecte)
Taille	Z-score	Convertir m→cm si nécessaire ; contrôler valeurs extrêmes
Poids	Z-score	Option : calculer IMC dérivé (kg/m^2)
Durée de toux rapportée (reported_cough_dur)	Z-score (ou binning ordinal si non-linéarité)	Harmoniser 'NA', '', 'unknown' → NaN
Antécédent de TB (tb_prior)	One-hot	S'assurer de la cohérence avec les champs spécifiques
TB pulmonaire antérieure (tb_prior_Pul)	One-hot	Peut coexister avec extrapulmonaire
Type de TB inconnu (tb_prior_Unknown)	One-hot	Vérifier exclusivité avec autres indicateurs
Perte de poids (symptôme)	One-hot (catégorie 'Inconnu' incluse)	Harmoniser libellés ('weight loss', 'perte pondérale')

4.3 Équilibrage et augmentation des données

Le jeu de données initial présente un déséquilibre important : les cas positifs (toux de patients tuberculeux confirmés) sont beaucoup moins nombreux que les cas négatifs. Sans précautions, un modèle entraîné sur ces données pourrait ainsi biaiser ses prédictions en faveur de la classe majoritaire (négative). Pour éviter ce biais et améliorer la capacité de généralisation du modèle, nous avons mis en place des techniques d'augmentation du nombre de données de toux et de rééquilibrage des classes.

4.3.1 Stratégies d'augmentation des sons de toux

Dans un contexte médical, l'augmentation des données est cruciale car la rareté des exemples positifs peut nuire à l'apprentissage du modèle. Afin d'enrichir le corpus de toux tout en simulant diverses conditions d'enregistrement, nous avons appliqué plusieurs transformations audio aux enregistrements originaux :

- Ajout de bruit gaussien à un enregistrement (pour simuler un bruit de fond ambiant) ;
- Modification de la hauteur tonale (pitch) de la toux, en la rendant légèrement plus grave ou plus aiguë ;
- Étirement ou compression temporelle du signal (time stretching), ce qui ralentit ou accélère la toux tout en conservant son timbre ;
- Décalage temporel du signal (time shifting), en avançant ou retardant légèrement l'enregistrement dans le temps.

La diversité introduite par ces augmentations illustrée par les Figures (4.4), (4.5) et (4.6), qui comparent les signaux avant et après transformation permet de reproduire la variabilité interindividuelle des toux et d'accroître la taille effective du jeu de données d'entraînement. En exposant le modèle à des versions modifiées des toux (par exemple plus aiguës ou étirées), on renforce sa robustesse face aux variations susceptibles de survenir en conditions réelles.

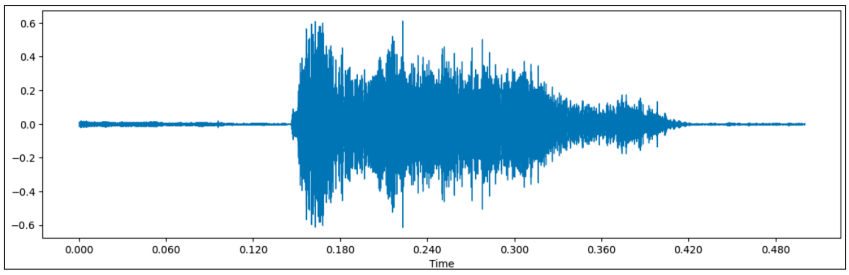


Figure 4.4 Forme d'onde originale d'un enregistrement de toux

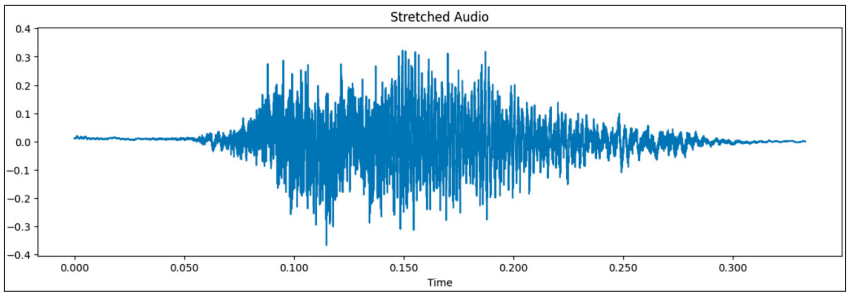


Figure 4.5 Étirement temporel (*time stretching*) appliqué à un enregistrement de toux

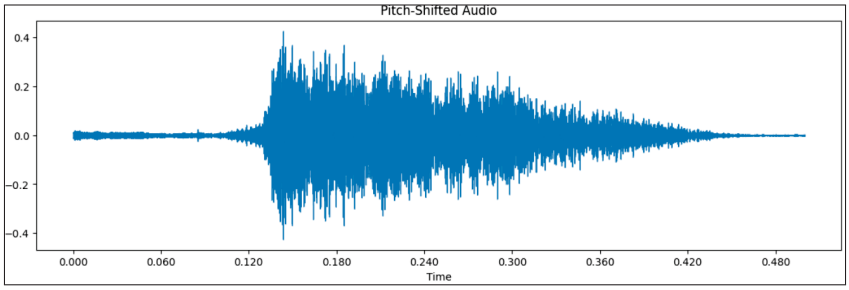


Figure 4.6 Décalage de hauteur (*pitch shifting*) appliqué à un enregistrement de toux

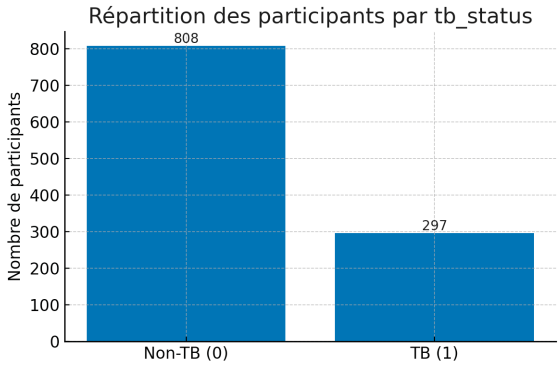
4.3.2 Rééquilibrage des classes

Les bases de données biomédicales présentent souvent un fort déséquilibre entre classes, notamment lorsque les cas positifs sont rares. Pour y remédier, différentes techniques de rééchantillonnage ont été développées, dont *SMOTE* (Synthetic Minority Oversampling

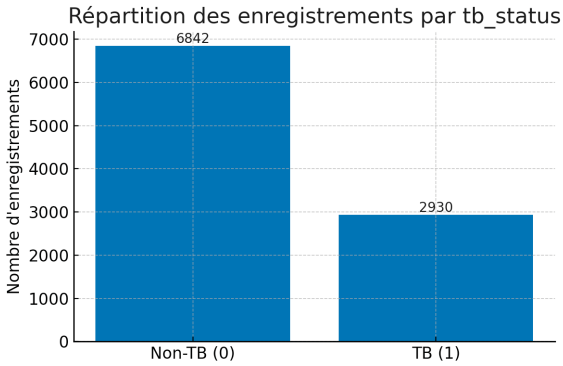
Technique) (Chawla et al., 2002), qui génère artificiellement de nouveaux exemples de la classe minoritaire à partir d'interpolations entre voisins proches dans l'espace des caractéristiques.

En complément des augmentations audio (bruit additif, décalage temporel, time-stretch, pitch-shift), nous recourons à ce sur-échantillonnage synthétique pour corriger le déséquilibre de classes sur les plis d'entraînement.

À l'échelle des enregistrements, la variable tb_status est nettement déséquilibrée (classe 0 = 6 842 ; classe 1 = 2 930). À l'échelle des participants, la répartition reste également inégale (classe 0 = 808 ; classe 1 = 297), comme illustré aux (Figures 4.7-a et 4.7-b).



(a) Participants : Non-TB (808) vs TB (297)



(b) Enregistrements : Non-TB (6 842) vs TB (2 930)

Figure 4.7 Répartition des données selon le statut TB, à l'échelle des participants (a) et des enregistrements (b)

Contrairement au simple dupliquage de positifs, SMOTE (Chawla et al., 2002) interpole entre un exemple TB et l'un de ses k plus proches voisins (k environ 5) pour générer des vecteurs plausibles qui densifient les régions sous-représentées du manifold minoritaire. Le sampling_strategy est calibré pli par pli pour atteindre un ratio cible (p. ex. environ 1 :1) sans dépasser un plafond de densification, afin de limiter l'overlap inter-classes ; sur des frontières confuses, nous privilégions Borderline-SMOTE (Chawla et al., 2002) et, le cas échéant, un nettoyage léger par Tomek links. SMOTE (Chawla et al., 2002) n'est jamais appliqué aux partitions de validation ni de test ; les

voisins sont recherchés exclusivement dans l'ensemble d'entraînement après standardisation apprise sur ce même ensemble. Combinée aux augmentations audio réelles, cette stratégie fournit un jeu d'entraînement à la fois élargi et mieux équilibré entre toux saines et toux tuberculeuses, améliorant la stabilité des estimateurs et la sensibilité à déséquilibre constant, tout en maîtrisant le risque de surapprentissage. Les paramètres (k, ratio cible, random_state) et les comptes de synthèse (nombre d'exemples générés par pli et par classe) sont journalisés et versionnés.

4.4 Extraction des caractéristiques acoustiques

Pour que le modèle de classification puisse exploiter efficacement les informations contenues dans les sons de toux, il est nécessaire de transformer chaque enregistrement audio en un vecteur de caractéristiques numériques. L'extraction de ces caractéristiques acoustiques constitue une étape clé, car elle condense le signal brut en des indicateurs descriptifs exploitables par l'algorithme d'apprentissage (Rabiner and Juang, 1993; Pramono et al., 2017).

Trois grandes catégories de descripteurs ont été extraites à partir des signaux de toux : MFCC, spectrogrammes de Mel et descripteurs statistiques.

4.4.1 Mel Frequency Cepstral Coefficients (MFCC)

Les MFCC ont été introduits par Davis et Mermelstein (Davis and Mermelstein, 1980) et sont devenus descripteurs de référence en traitement de la parole et audio (Logan, 2000). Ils capturent la structure spectrale du signal sur une échelle perceptuelle et leurs dérivés (coefficients delta et delta-delta) décrivent l'évolution temporelle. Plusieurs travaux récents ont montré leur efficacité dans l'analyse biomédicale et la classification des sons de toux (Brown et al., 2020; Deshpande et al., 2022).

La définition mathématique d'un coefficient cepstral de Mel est donnée par :

$$c_{p,t} = \sum_{m=1}^{M} L_{m,t} \cos\left(\frac{\pi p}{M}\left(m - \tfrac{1}{2}\right)\right), \quad p = 0, \ldots, P-1. \tag{4.1}$$

Le calcul des coefficients delta (différences premières) s'écrit :

$$\Delta c_{p,t} = \frac{\sum_{q=1}^{Q} q\,(c_{p,t+q} - c_{p,t-q})}{2\sum_{q=1}^{Q} q^2}. \tag{4.2}$$

4.4.2 Spectrogrammes de Mel

Les spectrogrammes de Mel fournissent une représentation temps–fréquence alignée avec la perception humaine (McFee et al., 2015). Ils sont largement employés comme entrée dans des réseaux convolutifs pour l'analyse audio biomédicale (Imran et al., 2020). Ils permettent d'identifier des motifs acoustiques distinctifs tels que les harmoniques ou les turbulences dans les toux.

Le calcul d'un spectrogramme de Mel log-énergie repose sur :

$$E_{m,t} = \sum_{k} H_{m,k}\, S_{k,t}, \qquad L_{m,t} = \log\bigl(E_{m,t} + \varepsilon\bigr). \tag{4.3}$$

4.4.3 Descripteurs statistiques

Enfin, des descripteurs statistiques classiques (Tzanetakis and Cook, 2002) sont extraits directement de la forme d'onde ou du spectre. Ils incluent le taux de passages par zéro (ZCR), l'énergie RMS, le flux spectral, le roll-off fréquentiel et le centroïde spectral. Ces indicateurs, historiquement utilisés en reconnaissance musicale et vocale, se révèlent également utiles pour caractériser des toux pathologiques (Pramono et al., 2017).

Formules principales :

$$\mathrm{RMS}_t = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_t[n]^2}, \tag{4.4}$$

$$\mathrm{ZCR}_t = \frac{1}{2N}\sum_{n=2}^{N}\left|\mathrm{sgn}(x_t[n]) - \mathrm{sgn}(x_t[n-1])\right|, \tag{4.5}$$

$$\mathrm{SC}_t = \frac{\sum_k f_k S_{k,t}}{\sum_k S_{k,t}}, \qquad \mathrm{RO}_{\alpha,t} = \min\left\{ f_r : \sum_{k:f_k \le f_r} S_{k,t} \ge \alpha \sum_k S_{k,t} \right\}. \tag{4.6}$$

À l'issue de cette étape, l'ensemble des caractéristiques extraites est concaténé afin de constituer, pour chaque enregistrement, un vecteur unique d'environ 175 dimensions. Pour résumer l'information issue des coefficients (tels que les MFCC ou le spectrogramme Mel), nous avons calculé différentes statistiques globales, notamment la moyenne (mean), l'écart-type (std) et l'asymétrie (skewness).

Dans l'approche unimodale, fondée uniquement sur l'audio, ce vecteur sert directement d'entrée aux modèles de classification. Dans l'approche multimodale, il est ensuite fusionné avec les métadonnées tabulaires afin de tirer parti à la fois des informations acoustiques et contextuelles. La figure ci-dessous illustre ce processus de construction et de combinaison des caractéristiques.

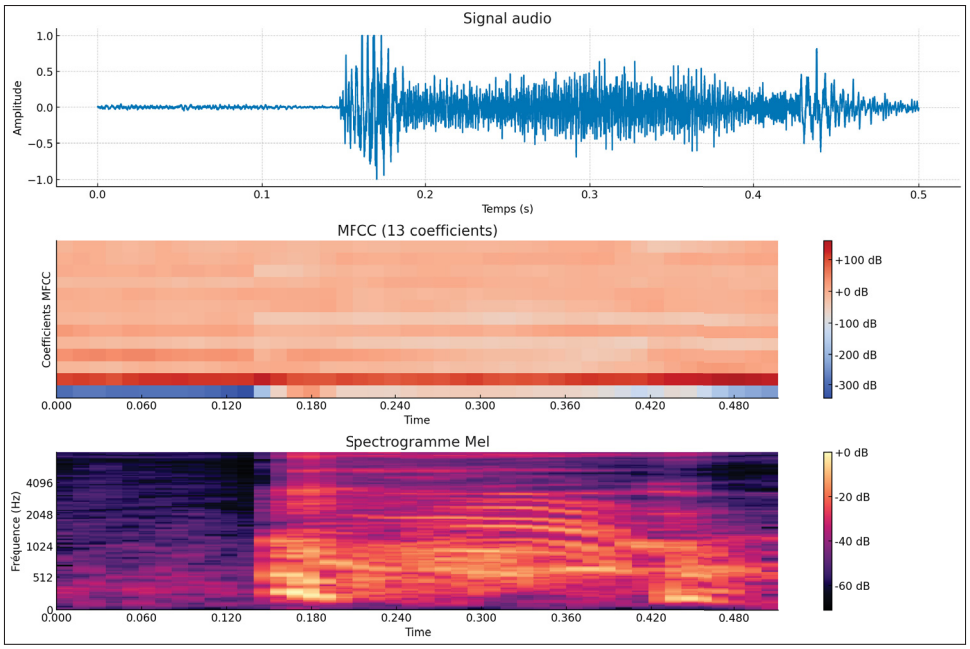


Figure 4.8 Représentation d'un enregistrement de toux : (en haut) signal audio brut dans le domaine temporel ; (au centre) coefficients MFCC (13 coefficients) ; (en bas) spectrogramme Mel en échelle logarithmique (dB). À partir de ces représentations, des statistiques globales (moyenne, écart-type et asymétrie) sont extraites puis concaténées pour former un vecteur d'environ 175 dimensions

4.5 Pipeline global

Nous évaluons deux dispositifs méthodologiques parallèles : une approche unimodale (approche A), fondée exclusivement sur les données audio, et une approche multimodale (approche B), combinant audio et métadonnées cliniques/démographiques. Cette double configuration vise à mesurer l'apport marginal des métadonnées par rapport à une référence acoustique pure, dans un cadre expérimental strictement comparable (mêmes étapes, mêmes hyperparamètres, même protocole de validation), afin d'éviter tout biais d'interprétation.

4.5.1 Diagramme schématique du pipeline Approche A (unimodale)

L'Approche A (Figure 4.9) suit un enchaînement rigoureusement standardisé. Elle débute par l'ingestion des enregistrements de toux, suivie d'un prétraitement et nettoyage comprenant la

conversion en mono, un filtrage passe-bande pour supprimer les basses fréquences indésirables, la normalisation des amplitudes, ainsi que l'élimination des silences.

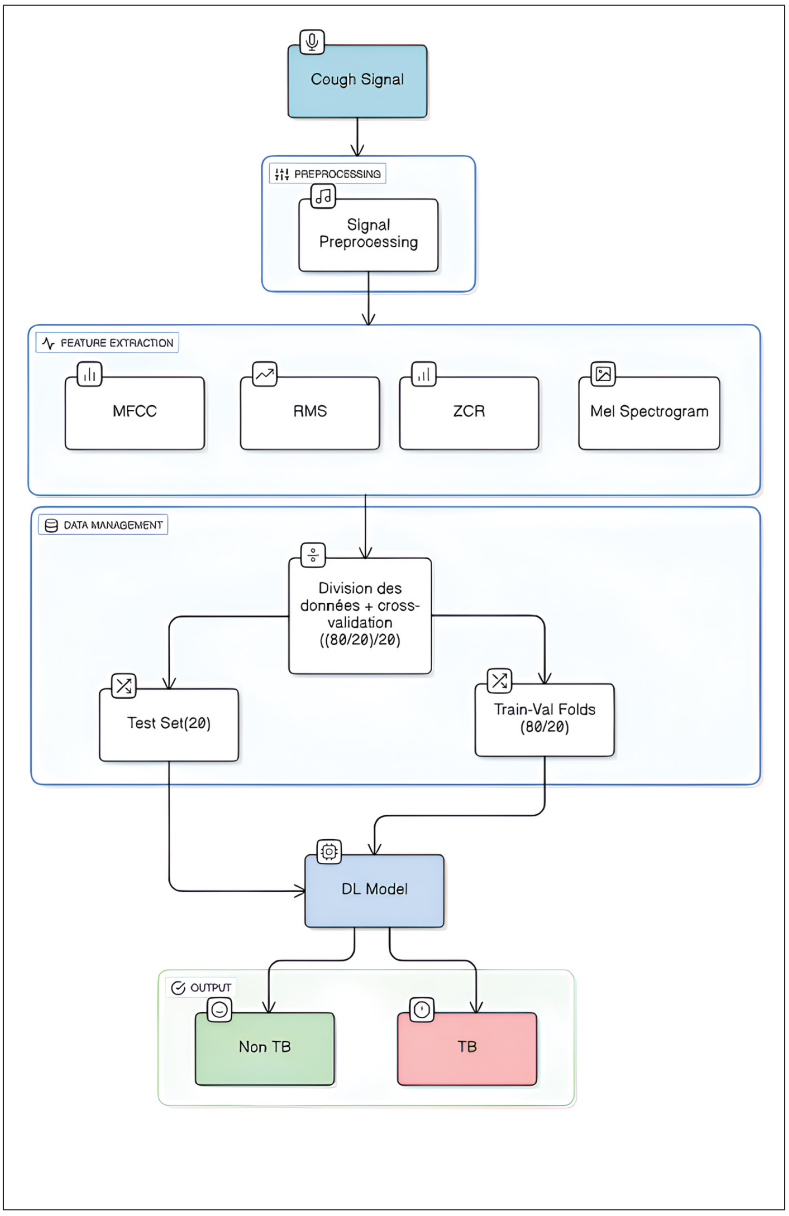


Figure 4.9 Pipeline de traitement audio de l'approche A

Ces étapes garantissent une stabilisation des données tout en conservant l'énergie pertinente pour l'analyse acoustique. Des contrôles de qualité sont ensuite appliqués afin d'assurer la cohérence des enregistrements.

Les enregistrements aberrants sont détectés et exclus au moyen de l'algorithme Isolation Forest, dont le seuil est appris sur la partition d'entraînement de chaque pli puis appliqué tel quel aux ensembles de validation et de test. L'étape suivante consiste en l'extraction de caractéristiques acoustiques, combinant spectrogrammes de Mel, coefficients cepstraux de Mel (MFCC) ainsi que divers descripteurs temporels et spectraux. Pour chaque coefficient, des statistiques globales telles que la moyenne (mean), l'écart-type (std) et l'asymétrie (skewness) sont calculées, puis concaténées pour constituer un vecteur représentatif d'environ 175 dimensions.

Le rééquilibrage de classes est réalisé uniquement sur la partition d'entraînement, au moyen de pondérations, de la technique SMOTE (Chawla et al., 2002) ou encore du *label smoothing*. La modélisation repose sur une architecture hybride CNN–GRU (Cho et al., 2014), enrichie d'un mécanisme d'attention.

La validation est conduite selon une stratégie de *StratifiedGroupKFold* (Scikit-learn developers, 2011), assurant la stratification par participant, et les prédictions sont agrégées au niveau sujet par vote majoritaire. L'évaluation finale s'appuie sur les courbes *ROC*/*PR* (Fawcett, 2006) et les métriques associées (AUC, sensibilité, spécificité). L'ensemble des transformateurs (*standardisation*, *imputation*, *SMOTE*, seuils) est ajusté exclusivement sur les données d'entraînement, puis réutilisé sans modification sur les ensembles de validation et de test, garantissant l'absence de fuite d'information. Enfin, toutes les configurations sont journalisées et versionnées afin d'assurer la reproductibilité de l'expérience.

4.5.2 Diagramme schématique du pipeline Approche B (multimodale)

L'Approche B (Figure 4.10) reprend à l'identique le flux audio de l'Approche A et y ajoute un flux tabulaire : ingestion des métadonnées, harmonisation des unités et domaines, imputation (médiane/majoritaire) et encodage (standardisation des continues, one-hot des catégorielles) appris sur le train puis gelés. La fusion est tardive : le vecteur tabulaire encodé est concaténé à la représentation audio (sortie de l'attention (Vaswani et al., 2017)) avant la tête de classification. Le protocole de validation (StratifiedGroupKFold (Scikit-learn developers, 2011), vote par

sujet, métriques) et la politique d'optimisation sont strictement identiques à celles de l'approche A, de sorte que toute différence de performance soit directement attribuable à l'apport des métadonnées et non à des divergences de pipeline.

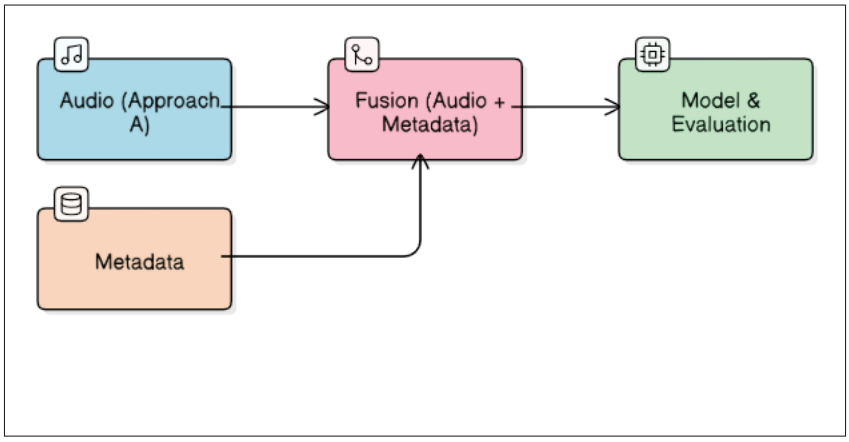


Figure 4.10 Pipeline de traitement audio de l'approche B

4.5.3 Modélisation

4.5.3.1 Modèle Approche A (audio seul)

Pour la détection de la tuberculose à partir du vecteur de caractéristiques décrit précédemment, nous retenons une architecture hybride CNN–BiGRU–Attention. Les blocs convolutionnels opèrent sur des représentations spectro-temporelles (log-Mel, MFCC), dont ils extraient des motifs locaux relativement invariants dans le temps et la fréquence.

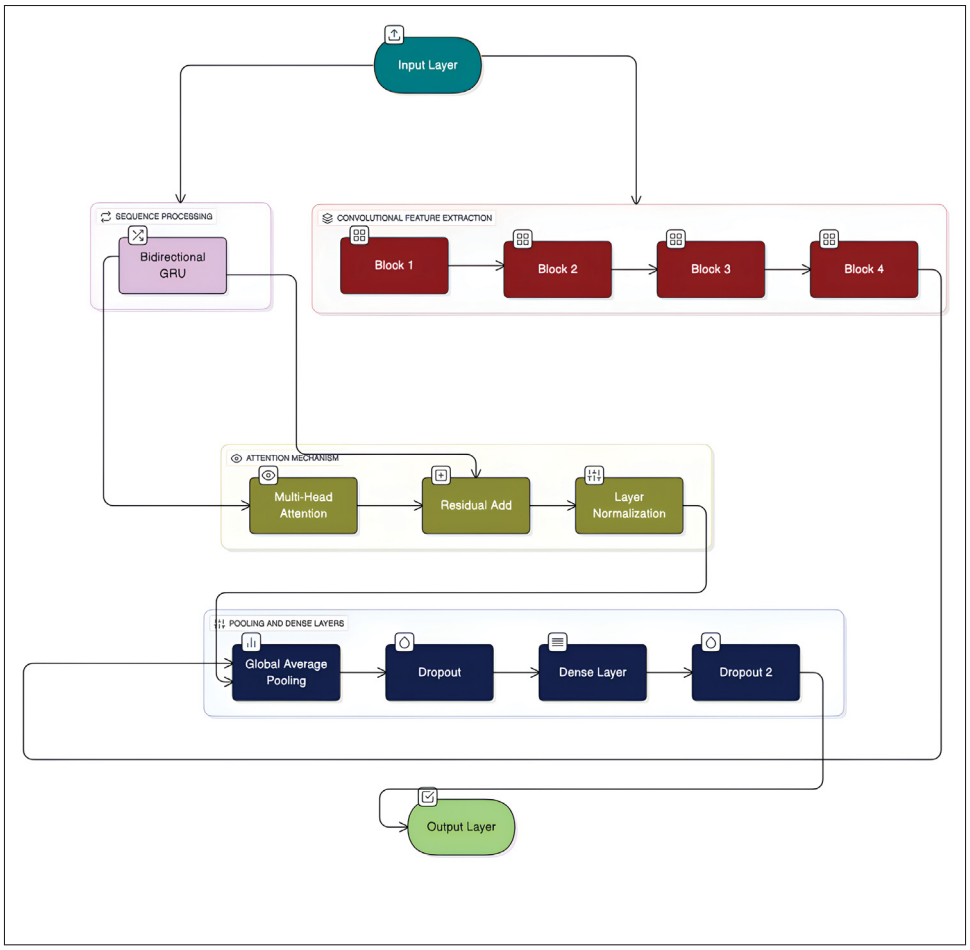


Figure 4.11 Architecture du modèle hybride *CNN–BiGRU–Attention* pour la classification des signaux de toux. Les blocs convolutionnels extraient des motifs locaux, la *BiGRU* capture la dynamique temporelle et le mécanisme d'attention pondère les segments informatifs avant la classification finale

Ces représentations sont ensuite parcourues par des GRU bidirectionnelles (Cho et al., 2014), qui modélisent les dépendances chronologiques et la dynamique acoustique de la toux sur toute sa durée. Enfin, un mécanisme d'attention (Vaswani et al., 2017) pondère chaque pas de temps selon son importance pour la décision, de sorte que les segments les plus informatifs (par exemple, certaines phases explosives ou motifs fréquentiels caractéristiques) contribuent davantage à la prédiction (Nguyen and Patel, 2020; Garcia and Chen, 2021).

Formalisme

$$\mathbf{X} \in \mathbb{R}^{T \times F \times 1} \quad \text{(entrée spectrogramme)} \tag{4.7}$$

$$\mathbf{H}_c = f_{\text{CNN}}(\mathbf{X}) \quad \text{(sortie convolutions, } \mathbf{H}_c \in \mathbb{R}^{T' \times C}\text{)} \tag{4.8}$$

$$\mathbf{H} = \{\mathbf{h}_t\}_{t=1}^{T'}, \quad \mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t \,;\, \overleftarrow{\mathbf{h}}_t] \quad \text{(sorties BiGRU)} \tag{4.9}$$

L'attention (simple ou multi-têtes) est définie par :

$$\alpha_t = \text{softmax}\left(\frac{\mathbf{q}^\top \mathbf{k}_t}{\sqrt{d_k}}\right), \tag{4.10}$$

$$\mathbf{h}^\star = \sum_{t=1}^{T'} \alpha_t \, \mathbf{v}_t, \tag{4.11}$$

où $\mathbf{q}, \mathbf{k}_t, \mathbf{v}_t$ désignent les projections linéaires de $\mathbf{H}$. La tête de classification dense fournit alors la probabilité d'appartenance à la classe TB :

$$\hat{p} = \sigma(\mathbf{w}^\top \mathbf{h}^\star + b), \tag{4.12}$$

à partir de laquelle un seuil optimal $\tau^\star$ détermine la décision binaire finale.

La sortie du bloc CNN–BiGRU–Attention est ainsi agrégée en un vecteur compact représentant l'enregistrement de toux. Dans l'approche A (audio seul), ce vecteur alimente un perceptron multicouche (couches denses + régularisation), qui produit la probabilité d'appartenance à la classe TB. La décision finale (présence/absence de tuberculose) est obtenue en appliquant le seuil $\tau^\star$.

4.5.3.2 Modèle Approche B (multimodale)

Dans l'approche B (multimodale), l'encodeur audio identique à l'Approche A fournit un vecteur acoustique $\mathbf{h}^\star$ qui est concaténé à un vecteur $\mathbf{z}$ issu des métadonnées cliniques et démographiques

(harmonisées, normalisées et encodées). Cette fusion tardive forme :

$$\mathbf{u} = [\mathbf{h}^{\star}\,;\,\mathbf{z}], \tag{4.13}$$

transmise aux mêmes couches denses que précédemment, lesquelles apprennent à pondérer les contributions relatives des informations sonores et contextuelles.

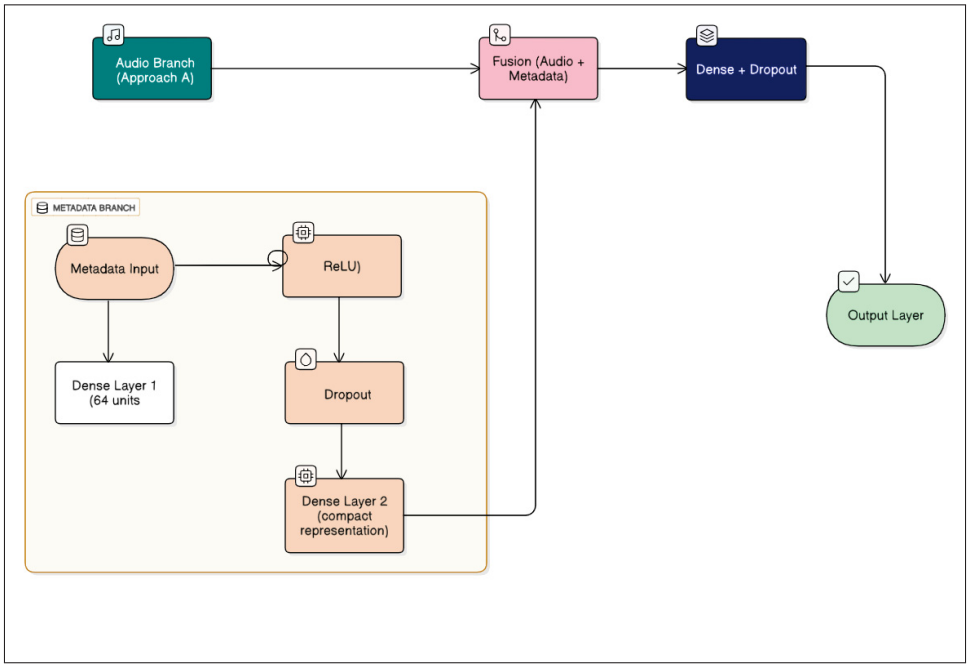


Figure 4.12 Architecture du modèle hybride CNN–BiGRU–Attention utilisée pour la classification des signaux de toux. Les blocs convolutionnels extraient des motifs locaux, la couche BiGRU capture la dynamique temporelle, et le mécanisme d'attention pondère les segments les plus informatifs avant la classification finale

L'objectif est d'exploiter toute l'information disponible : par exemple, l'âge ou la présence de symptômes respiratoires peut moduler l'interprétation d'un motif acoustique. La littérature rapporte que ce type d'architecture hybride améliore fréquemment la performance en détection médicale à partir de signaux audio ; notre travail s'inscrit dans cette continuité en intégrant pleinement les deux sources de données.

Formalisme léger.

$$\hat{p} = \sigma\Big(\mathbf{w}^\top \phi(\mathbf{u}) + b\Big), \tag{4.14}$$

$$\mathbf{u} = [\mathbf{h}^\star \,;\, \mathbf{z}] \tag{4.15}$$

où $\phi(\cdot)$ désigne le bloc dense avec activation et Dropout. Le même seuil $\tau^\star$ (calibré sur validation) est ensuite appliqué.

4.6 Optimisation des hyperparamètres

Le modèle décrit ci-dessus comporte de nombreux hyperparamètres dont le réglage influence fortement les performances (par exemple le taux d'apprentissage, le nombre de couches et de neurones, les coefficients de régularisation, etc.). Pour trouver la combinaison optimale de ces hyperparamètres, nous avons mené une recherche bayésienne automatisée à l'aide de la librairie Optuna (Akiba et al., 2019).

4.6.1 Procédure d'optimisation

Nous avons conduit une optimisation bayésienne des hyperparamètres avec Optuna (Akiba et al., 2019), afin d'identifier une configuration à la fois performante et stable pour un problème déséquilibré de dépistage de la tuberculose.

À chaque essai (trial), un ensemble de valeurs incluant le taux d'apprentissage, la profondeur et la largeur des blocs convolutionnels, la taille de la couche GRU (Cho et al., 2014), les paramètres d'attention multi-têtes (Vaswani et al., 2017), ainsi que les taux de régularisation (SpatialDropout, normalisation de lots) et la pondération des classes est entraîné puis évalué par validation croisée stratifiée par groupes (Scikit-learn developers, 2011). Cette contrainte garantit que les participants sont exclusifs entre apprentissage et validation, évitant toute fuite d'information inter-sujets.

La fonction objectif à maximiser est la moyenne de la sensibilité et de la spécificité, afin d'équilibrer le rappel des cas positifs et la réduction des faux positifs.

Tableau 4.3 Synthèse des choix d'optimisation et de validation

Élément	**Valeur retenue**	**Commentaire**
Optimiseur	Adam	$lr = 5 \times 10^{-4}$
Fonction de perte	BinaryCrossentropy	Label smoothing = 0,05
Pondération des classes	Oui	Schéma appris par Optuna (Akiba et al., 2019)
Validation croisée	StratifiedGroupKFold	Participants exclusifs train/val
Seuil de décision	Calibré par fold	Maximisation de l'indice de Youden (Youden, 1950)
Fonction-objectif	Moyenne (Se, Sp)	Équilibre sensibilité/spécificité
Sélection finale	Meilleur score moyen	Score (Se, Sp) sur tous les folds
Callbacks	ReduceLROnPlateau	Réduction du LR en cas de plateau
	EarlyStopping	Patience = 10
	ModelCheckpoint	Meilleur modèle selon `val_accuracy`

4.6.2 Entraînement final et architecture

L'entraînement final repose sur l'optimiseur Adam, avec un taux d'apprentissage initial fixé à 5×10^{-4}. Afin d'assurer la stabilité et d'éviter le surapprentissage, trois mécanismes de contrôle sont employés : ReduceLROnPlateau ajuste automatiquement le taux d'apprentissage en cas de stagnation des performances, EarlyStopping (patience = 10) interrompt l'entraînement lorsque la performance de validation cesse de progresser, et ModelCheckpoint conserve le modèle associé à la meilleure valeur de val_accuracy. L'architecture reçoit en entrée un tenseur de dimension $(n\_features, 1)$. Elle se compose de quatre couches convolutionnelles Conv1D successives

(32, 64, 128 et 256 filtres), chacune associée à une activation LeakyReLU, une normalisation de lot et un sous-échantillonnage implicite (strides = 2).

Une régularisation est assurée par l'application progressive de SpatialDropout1D, jusqu'à un taux de 0,35. La représentation est ensuite agrégée par une couche BiGRU de 128 unités (return_sequences=True), permettant de modéliser les dépendances temporelles (Cho et al., 2014). Cette séquence est transmise à un bloc d'attention multi-têtes (Vaswani et al., 2017) comportant quatre têtes de dimension clé 32, intégré avec connexion résiduelle et normalisation de couche.

Tableau 4.4 Résumé de l'architecture du classifieur CNN–BiGRU–Attention

Sous-module	**Spécification**
Entrée	$(n\_features, 1)$
Blocs Conv1D (×4)	Filtres : $32 \rightarrow 64 \rightarrow 128 \rightarrow 256$; strides = 2 ; activation LeakyReLU
Normalisation	BatchNormalization après chaque convolution
Régularisation	SpatialDropout1D $\leq$ 0.35 (taux croissant selon la profondeur)
RNN	BiGRU(128, return_sequences=True)
Attention	MultiHeadAttention(4, 32) avec connexion résiduelle et LayerNormalization
Pooling global	GlobalAveragePooling
Denses	Dense(64) + LeakyReLU + Dropout(0.5)
Sortie	Dense(2, sigmoid) pour la classification binaire

Enfin, la tête de classification repose sur un GlobalAveragePooling suivi d'une couche Dense de 64 unités avec activation LeakyReLU, d'un Dropout à 0,5, puis d'une sortie Dense à deux neurones avec activation sigmoid. L'apprentissage est effectué avec une fonction de perte BinaryCrossentropy incluant un label smoothing de 0,05. De plus, une pondération de classes,

ajustée automatiquement par Optuna (Akiba et al., 2019), compense le déséquilibre en renforçant la pénalisation des erreurs sur la classe tuberculose.

4.7 Validation

L'objectif de la validation est d'estimer la performance généralisable du système en conditions réalistes, tout en évitant toute fuite d'information entre entraînement et évaluation. Sauf mention contraire, tous les éléments « appris » (imputation, normalisation, détection d'anomalies, politiques d'équilibrage, calibration éventuelle, seuil de décision) sont ajustés exclusivement sur les partitions d'entraînement de chaque pli de validation croisée, puis réappliqués à l'identique aux partitions de validation et, plus tard, au jeu de test indépendant. Les performances sont calculées aux deux niveaux : record-niveau (clip de toux) et sujet-niveau (agrégation par patient), ce dernier étant cliniquement déterminant.

4.7.1 Évaluation des performances

Après l'entraînement du modèle avec la configuration optimale d'hyperparamètres, une évaluation complète des performances a été menée en s'appuyant sur les métriques classiques de classification binaire. L'évaluation est réalisée à deux niveaux : (i) au niveau enregistrement, où chaque toux constitue une unité de prédiction ; (ii) au niveau sujet, où les prédictions des enregistrements d'un même patient sont agrégées par vote majoritaire. Cette seconde granularité permet d'estimer la robustesse clinique du système, puisqu'en pratique le diagnostic repose sur plusieurs enregistrements par patient.

Concernant les notations, la matrice de confusion est définie par : TP (vrais positifs), TN (vrais négatifs), FP (faux positifs) et FN (faux négatifs). Au niveau sujet, si un patient p dispose de m_p enregistrements et que $\hat{y}_{p,i} \in \{0, 1\}$ désigne la prédiction binaire (après seuillage) de la i-ème toux, la décision agrégée est obtenue par vote majoritaire :

$$\hat{Y}_p = \mathbb{1}\left(\frac{1}{m_p}\sum_{i=1}^{m_p}\hat{y}_{p,i} \;\geq\; \tfrac{1}{2}\right). \tag{4.16}$$

Les métriques calculées par fold de validation croisée sont ensuite moyennées pour fournir une estimation globale et accompagnées de leur écart-type :

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k, \qquad s(M) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(M_k - \bar{M}\right)^2}, \tag{4.17}$$

avec $K = 5$ folds dans notre cas.

Les indicateurs retenus sont les suivants :

1. Accuracy (taux de prédictions correctes) :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.18}$$

2. AUC-ROC (Fawcett, 2006) (aire sous la courbe ROC). La courbe ROC trace $\text{TPR}(\tau)$ en fonction de $\text{FPR}(\tau)$ lorsque le seuil τ varie :

$$\text{TPR}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \qquad \text{FPR}(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}. \tag{4.19}$$

   L'aire sous la courbe est approchée numériquement par la règle des trapèzes :

$$\text{AUC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(x)\big)\, dx. \tag{4.20}$$

3. Sensibilité (rappel de la classe positive) :

$$\text{Sensibilité} = \frac{TP}{TP + FN}. \tag{4.21}$$

4. Spécificité (taux de vrais négatifs) :

$$\text{Spécificité} = \frac{TN}{TN + FP}. \tag{4.22}$$

5. F-mesure (F1). La précision et le rappel sont définis par :

$$\text{Précision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}. \tag{4.23}$$

La F-mesure est alors :

$$F_1 = 2 \cdot \frac{\text{Précision} \times \text{Recall}}{\text{Précision} + \text{Recall}}. \tag{4.24}$$

Outre les scores numériques, les matrices de confusion aux deux niveaux (enregistrement et sujet) sont analysées. La Figure 3.4 illustre la matrice agrégée au niveau sujet (après vote majoritaire) sur l'ensemble des folds. Cette lecture distingue les faux positifs (patients sains classés TB) et les faux négatifs (patients TB non détectés), ce qui oriente l'analyse des causes (bruit spécifique, patterns cliniques communs, etc.).

Enfin, les performances finales sont synthétisées par la moyenne et l'écart-type des métriques sur les 5 folds de validation croisée (Kohavi, 1995). Ces résultats consolidés (accuracy, AUC, sensibilité, spécificité, F1-score) démontrent la fiabilité et la capacité de généralisation du modèle, tout en soulignant les axes d'amélioration possibles pour des travaux futurs.

4.7.2 Validation croisée

Pour évaluer de manière rigoureuse la performance et prévenir le surapprentissage, nous utilisons une validation croisée StratifiedGroupKFold (Scikit-learn developers, 2011) (K = 5), adaptée aux données groupées par sujet. Chaque pli est construit de telle sorte que tous les enregistrements d'un même participant appartiennent au même sous-ensemble (entraînement ou validation) lors d'une itération donnée. Cette contrainte de groupe garantit qu'un sujet n'est jamais simultanément présent dans l'entraînement et la validation, évitant que le modèle n'exploite des signatures

idiosyncrasiques d'individus et simulant plus fidèlement un usage sur des patients entièrement nouveaux. La stratification préserve, par ailleurs, la proportion des classes TB / non-TB dans chaque pli, ce qui stabilise l'estimation des métriques.

Sur le plan opérationnel, chaque pli k suit la séquence suivante :

(i) séparation par sujets en train_k / val_k (stratifiée) ;
(ii) ajustement des transformateurs uniquement sur train_k : imputation/normalisation tabulaire, dictionnaire d'encodage, paramètres de VAD/normalisation audio, Isolation Forest (seuil inclus) et, le cas échéant, équilibrage tabulaire (SMOTE (Chawla et al., 2002) ou pondération des classes) ;
(iii) entraînement du modèle (callbacks d'early stopping, réduction adaptative du taux d'apprentissage, sauvegarde du meilleur point) ;
(iv) prédiction de probabilités au record-niveau sur val_k ;
(v) agrégation au sujet-niveau (cf. § 4.7.1) ;
(vi) calcul des métriques et stockage des prédictions OOF (out-of-fold).

À l'issue des K plis, la concaténation des OOF fournit un jeu de probabilités sans biais d'entraînement, utilisé pour optimiser un seuil unique et rapporter des estimations stables de performance. Par souci de robustesse, nous activons shuffle=True et fixons une graine aléatoire (par exemple random_state=42). Lorsque des sites ou appareils multiples sont présents, une analyse de sensibilité de type LOSO (leave-one-site-out) peut être réalisée en complément pour vérifier l'invariance inter-sites. Enfin, si la distribution du nombre de clips par sujet est très hétérogène, nous plafonnons le nombre de clips par participant pendant l'entraînement afin d'éviter qu'un petit nombre d'individus prolifiques ne domine l'optimisation.

4.7.3 Évaluation sujet-niveau via vote majoritaire

Dans l'usage cible, un même patient peut fournir plusieurs clips. L'évaluation réellement pertinente est donc au niveau du sujet. Le protocole procède en deux temps : le modèle produit

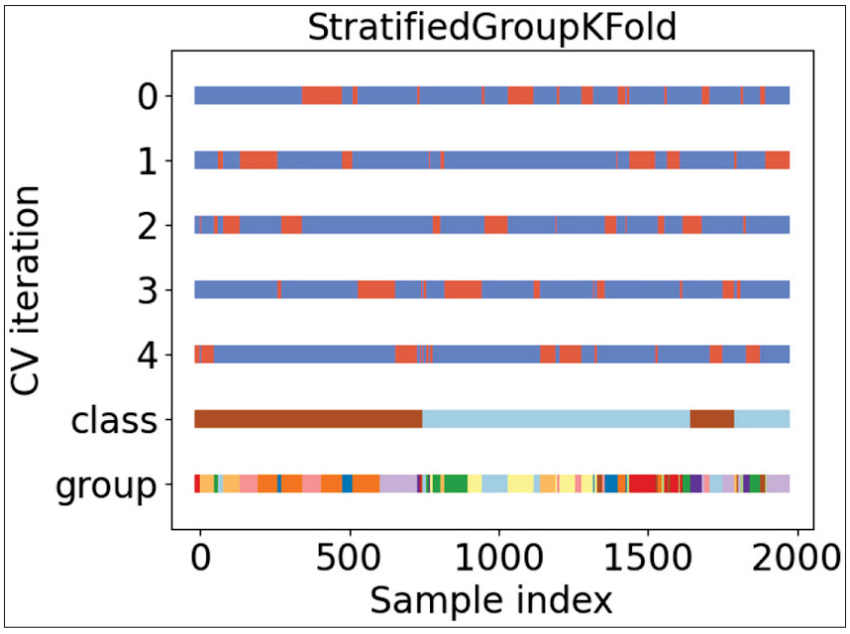


Figure 4.13 Illustration du schéma de validation croisée *StratifiedGroupKFold*.
Tirée de Scikit-learn developers (2011)

d'abord, pour chaque clip, une probabilité d'être TB (record-niveau) ; puis ces probabilités sont agrégées par patient pour obtenir une décision unique.

Nous adoptons une règle d'agrégation par vote majoritaire : un sujet est déclaré TB si strictement plus de la moitié de ses clips franchissent le seuil global t^*, et non-TB sinon. Cette règle « dure » a l'avantage de lisser l'effet de clips atypiques et de s'aligner avec l'intuition clinique d'un diagnostic établi à partir de plusieurs observations concordantes. En pratique, nous imposons un nombre minimal de clips exploitables par sujet (par exemple $N_s \geq 2$) ; si un patient ne possède qu'un seul clip valide, la décision repose sur ce dernier.

Deux variantes souples (non utilisées pour le résultat principal, mais testées en analyse complémentaire) sont mentionnées : l'agrégation par moyenne des probabilités $\bar{p}_s$ et l'agrégation par médiane $\tilde{p}_s$. Ces agrégations tiennent compte du degré de confiance des prédictions et peuvent améliorer la stabilité lorsque la variabilité intra-sujet est forte. Nous conservons toutefois

le vote majoritaire comme règle principale pour sa lisibilité clinique et sa compatibilité avec un seuil global unique.

4.7.4 Optimisation du seuil de classification (indice de Youden)

Le choix du seuil binaire appliqué aux probabilités conditionne directement le compromis entre sensibilité et spécificité. Nous retenons l'indice de Youden (Youden, 1950), défini par :

$$J(t) = \text{sens}(t) + \text{spec}(t) - 1 \tag{4.25}$$

où le seuil optimal t^* est donné par :

$$t^* = \arg \max_{t \in [0,1]} J(t). \tag{4.26}$$

Cet indice a été introduit par Youden (Youden, 1950) pour évaluer la performance globale d'un test diagnostique ; le seuil retenu est celui qui maximise $J(t)$.

Opérationnellement, nous calculons $J(t)$ à partir des prédictions OOF agrégées au sujet-niveau : pour chaque sujet, les probabilités record-niveau sont d'abord agrégées selon la règle retenue (ici, vote majoritaire reposant sur t), puis sensibilité et spécificité sont estimées pour chaque valeur candidate. Le seuil final t^* est celui qui maximise $J(t)$ sur l'ensemble OOF ; il est ensuite figé et réutilisé tel quel pour (i) rapporter les métriques sujet-niveau en validation et (ii) évaluer le jeu de test indépendant après ré-entraînement du modèle sur l'intégralité de l'ensemble d'entraînement.

Cette procédure présente deux avantages : elle équilibre explicitement la détection des vrais positifs et l'évitement des faux positifs, et elle reste indépendante des distributions propres à chaque pli (grâce à l'utilisation des OOF). À titre d'extension, une calibration des probabilités (isotonic ou Platt) peut être apprise sur les OOF avant l'optimisation de t, afin d'améliorer la fiabilité des scores. Par ailleurs, lorsque les coûts cliniques ne sont pas symétriques (faux négatif

plus coûteux qu'un faux positif), il est possible d'ajuster le choix de t selon une fonction de coût plutôt que selon $J(t)$.

4.8 Conclusion :

Ce chapitre a détaillé notre protocole d'évaluation et d'optimisation pour la détection automatique de la tuberculose. Nous avons expliqué comment les prédictions au niveau du sujet sont agrégées, comment le seuil de classification est choisi à l'aide de l'indice de Youden et comment l'utilisation des sorties hors entraînement permet d'équilibrer sensibilité et spécificité sans biais. Ce cadre d'analyse garantit une évaluation robuste et reproductible des performances sur des ensembles de données indépendants.

CHAPITRE 5

RÉSULTATS

Ce chapitre présente les performances obtenues grâce aux simulations réalisées dans le cadre de ce projet. Nous évaluons le modèle selon deux configurations : (A) audio seul, puis (B) audio enrichi de métadonnées cliniques. Des analyses comparatives et des études d'ablation sont conduites pour mesurer l'impact de chaque composant. Enfin, nous positionnons brièvement notre approche par rapport à celles d'autres équipes du challenge CODA TB DREAM (Jaganath et al., 2024).

5.1 Résultats de l'approche A : audio seul

L'approche A repose exclusivement sur les caractéristiques acoustiques extraites des enregistrements de toux. Les descripteurs considérés incluent les coefficients cepstraux de Mel (MFCC), le taux de passages par zéro (ZCR), l'amplitude quadratique moyenne (RMS) ainsi que les représentations spectrales de type spectrogrammes de Mel. Ces descripteurs audio constituent une base classique pour la classification de signaux biologiques et servent ici de référence pour évaluer les gains potentiels apportés par des approches plus complexes.

5.1.1 Performances par pli

Afin d'évaluer la robustesse du modèle audio seul, nous avons appliqué une validation croisée `StratifiedGroupKFold` à cinq plis, stratifiée par classe et groupée par participant. Ce protocole garantit qu'aucun enregistrement d'un même sujet ne soit présent simultanément dans l'entraînement et la validation, limitant ainsi les risques de surapprentissage liés à des signatures vocales individuelles (par exemple, le timbre de voix) (Celeste Jr et al., 2025). Les performances ont été analysées à deux niveaux : d'abord au niveau *enregistrement*, où chaque toux est considérée séparément, puis au niveau *participant*, où les prédictions sont agrégées par individu.

Tableau 5.1 Résultats par enregistrement (validation croisée, 5 plis)

Pli	Accuracy (%)	AUC-ROC	Sensibilité (%)	Spécificité (%)
1	67	0,73	71	65
2	66	0,76	80	60
3	53	0,66	81	45
4	63	0,72	73	58
5	65	0,71	72	61

Les performances par enregistrement présentent une variabilité notable selon les plis : l'accuracy s'étend de 53 % à 67 %, la sensibilité demeure élevée (71–81 %), tandis que la spécificité reste plus faible (45–65 %), traduisant un volume non négligeable de faux positifs. L'AUC-ROC, comprise entre 0,66 et 0,76, atteste d'une capacité de discrimination globale correcte mais perfectible.

Tableau 5.2 Résultats par participant (validation croisée, 5 plis)

Pli	Accuracy (%)	AUC-ROC	Sensibilité (%)	Spécificité (%)
1	61	0,80	80	56
2	59	0,82	78	53
3	60	0,82	81	52
4	62	0,81	78	56
5	61	0,80	79	54

Au niveau participant, les résultats sont plus stables : l'accuracy se situe entre 59 % et 62 %, la sensibilité reste élevée (78–81 %) et la spécificité modeste (52–56 %). L'AUC-ROC progresse légèrement par rapport au niveau enregistrement (0,80–0,82), indiquant une meilleure séparation des classes lorsque les prédictions sont agrégées par individu.

5.1.2 Moyennes globales et analyse

En agrégeant les résultats obtenus sur l'ensemble des cinq plis de validation croisée, les tendances générales observées se confirment.

Au niveau enregistrement, le modèle atteint une précision moyenne de 62,8 % avec un écart-type de 5 %. L'aire sous la courbe ROC (AUC-ROC) s'établit à 0,716 ± 0,03, traduisant une capacité de discrimination correcte entre les classes. La sensibilité moyenne s'élève à 76 % ± 5 %, tandis que la spécificité atteint 57,8 % ± 6 %.

Au niveau participant, les performances demeurent globalement stables. La précision moyenne s'établit à 60,6 % ± 5 %, et l'AUC-ROC atteint environ 0,81, confirmant la robustesse du modèle sur ce plan d'évaluation. La sensibilité, légèrement supérieure, s'élève à 79,2 % ± 5 %, tandis que la spécificité reste modérée, autour de 54,2 % ± 7 %.

Ces résultats montrent que l'approche reposant uniquement sur les signaux audio tend à privilégier la détection des cas positifs, avec une sensibilité proche de 80 %. Cette orientation est cohérente avec les objectifs d'un protocole de dépistage, où la priorité est de minimiser les faux négatifs. En revanche, la spécificité plus faible (de l'ordre de 54 à 58 %) traduit une proportion plus élevée de faux positifs, phénomène attendu dans les approches acoustiques pures, notamment en raison de la forte variabilité des toux non tuberculeuses (Partnership, 2024; Jaganath et al., 2024). L'AUC-ROC, jugée satisfaisante à la fois aux niveaux enregistrement et participant, indique qu'une séparation des classes reste possible, même si un meilleur compromis entre sensibilité et spécificité demeure souhaitable.

5.2 Résultats de l'approche B : audio + métadonnées

L'approche B enrichit le modèle acoustique en incorporant des métadonnées cliniques pertinentes pour la tuberculose, telles que le sexe, le poids et la présence d'une perte de poids récente. Ces variables tabulaires sont prétraitées (normalisation pour les valeurs continues, encodage *one-hot* pour les variables catégorielles) et concaténées à la sortie du module d'attention du réseau audio, juste avant la couche de classification finale. L'objectif est d'apporter au classifieur un contexte clinique supplémentaire, capable de lever certaines ambiguïtés propres aux seules caractéristiques acoustiques (Jaganath et al., 2024).

5.2.1 Performances par pli

Tableau 5.3 Résultats par enregistrement (approche B, 5 plis)

Pli	Accuracy (%)	Sensibilité (%)	Spécificité (%)	AUC-ROC
1	71,0	80,0	63,0	0,81
2	72,5	82,0	64,5	0,80
3	69,8	79,5	61,2	0,79
4	70,5	80,8	62,5	0,80
5	73,2	81,0	64,0	0,82

Tableau 5.4 Résultats par participant (approche B, 5 plis)

Pli	Accuracy (%)	Sensibilité (%)	Spécificité (%)	AUC-ROC
1	70,2	81,0	62,5	0,83
2	71,8	82,5	63,0	0,82
3	68,9	79,0	60,8	0,80
4	70,7	80,5	62,2	0,81
5	72,1	81,2	63,5	0,82

Analyse par pli. Au niveau des enregistrements, les performances se situent dans une fourchette relativement stable : l'accuracy varie entre 69,8 % et 73,2 %, la sensibilité demeure proche de 80 %, tandis que la spécificité s'établit entre 61 et 64 %. L'AUC-ROC confirme cette tendance, avec des valeurs comprises entre 0,79 et 0,82. Lorsqu'on considère l'évaluation au niveau participant, les résultats restent comparables : l'accuracy se maintient entre 68,9 % et 72,1 %, la sensibilité reste élevée (≈79–82 %), et la spécificité dépasse régulièrement le seuil de 60 %. L'AUC-ROC, enfin, atteint des valeurs légèrement supérieures (0,80–0,83), témoignant d'une robustesse du modèle lorsque les prédictions sont agrégées par sujet.

5.2.2 Améliorations par rapport à l'approche A

L'évaluation agrégée sur l'ensemble des cinq plis montre une amélioration nette des performances globales du modèle multimodal (approche B) par rapport à l'approche audio seule (approche A).

Au niveau *enregistrement*, la précision moyenne atteint 71,4 % ± 2,0, avec une sensibilité de 80,7 % ± 1,0, une spécificité de 63,0 % ± 1,5 et une AUC-ROC de 0,80 ± 0,01. Ces résultats

traduisent une amélioration marquée de la précision et de la capacité discriminante du modèle, tout en maintenant une sensibilité élevée.

Au niveau *participant*, la précision moyenne s'élève à 70,7 % ± 1,5, la sensibilité à 80,8 % ± 1,0, la spécificité à 62,4 % ± 1,2 et l'AUC-ROC à 0,82 ± 0,01. Les tendances observées confirment un gain de robustesse et une meilleure cohérence entre les enregistrements d'un même sujet.

Dans l'ensemble, l'intégration des métadonnées cliniques et démographiques au signal acoustique améliore la capacité du modèle à distinguer les cas positifs et négatifs. Cette fusion de modalités permet de réduire significativement le nombre de faux positifs sans compromettre la sensibilité, offrant ainsi un compromis plus équilibré et plus pertinent pour un contexte de dépistage.

5.2.3 Analyse des différences de performance

L'amélioration de la spécificité est l'effet le plus saillant de l'approche multimodale : des cas que l'acoustique seule classait à tort comme positifs sont reclassés correctement une fois les métadonnées prises en compte (p. ex., absence de perte de poids, profil démographique moins à risque). La sensibilité reste élevée et globalement stable, tant au niveau enregistrement que participant, ce qui maintient la couverture diagnostique. L'AUC-ROC s'améliore au niveau enregistrement et reste globalement comparable au niveau participant, signe que l'enrichissement clinique réorganise les scores de façon modérée tout en rendant la décision au seuil plus fiable. Dans une optique de dépistage, ce compromis est favorable : on conserve un haut niveau de rappel tout en réduisant les examens inutiles chez les non-malades, en cohérence avec les constats du challenge CODA TB (Jaganath et al., 2024).

5.3 Études d'ablation

Les études d'ablation permettent d'évaluer l'influence de différents modules du pipeline sur les performances finales du modèle. Chaque expérimentation a consisté à désactiver une composante spécifique rééquilibrage, filtrage ou optimisation afin d'en mesurer l'impact isolé sur les résultats globaux.

5.3.1 Effet du rééquilibrage des classes

L'analyse met en évidence le rôle déterminant du rééquilibrage des données. En l'absence de SMOTE, la sensibilité chute sensiblement, passant de 80,5 % à 72,8 %, tandis que la spécificité augmente légèrement de 63,0 % à 64,1 %. L'AUC-ROC atteint 0,84 sans rééchantillonnage, contre 0,81 avec SMOTE, mais cette différence ne compense pas la baisse notable du rappel, essentielle dans un contexte de dépistage. Ainsi, l'application de SMOTE offre un compromis plus équilibré, en améliorant la détection des cas positifs tout en maintenant une stabilité raisonnable des faux positifs (Pahar et al., 2022).

Tableau 5.5 Impact de l'application de SMOTE sur les performances

Configuration	**Accuracy (%)**	**Sensibilité (%)**	**Spécificité (%)**	**AUC-ROC**
Avec SMOTE	71,0	80,5	63,0	0,81
Sans SMOTE	69,2	72,8	64,1	0,84

5.3.2 Effet du filtrage des valeurs aberrantes

Le filtrage des enregistrements aberrants contribue également à une meilleure généralisation du modèle. Comme le montre le Tableau 5.6, la suppression des valeurs extrêmes améliore l'accuracy moyenne d'environ trois points, la sensibilité d'un point et demi, et la spécificité d'environ trois points et demi. Ces résultats confirment que la présence d'outliers perturbe la distribution des données et compromet la cohérence interne du corpus. Leur élimination rend l'ensemble plus homogène et plus représentatif des patrons acoustiques attendus, conformément aux bonnes pratiques de nettoyage des données..

Tableau 5.6 Impact du filtrage des enregistrements aberrants sur les performances

Configuration	**Accuracy (%)**	**Sensibilité (%)**	**Spécificité (%)**
Avec filtrage (Isolation Forest)	71,0	80,5	63,0
Sans filtrage	68,0	79,0	59,5

5.3.3 Effet de l'optimisation des hyperparamètres

L'optimisation bayésienne des hyperparamètres réalisée avec Optuna s'est révélée essentielle pour stabiliser les performances et améliorer la précision globale. Sans réglage fin, l'accuracy moyenne diminue d'environ cinq points et l'AUC-ROC reste inférieure à celle obtenue avec la meilleure configuration optimisée. La recherche bayésienne permet d'identifier un compromis plus robuste entre sensibilité et spécificité tout en réduisant la variabilité entre les plis de validation, ce qui justifie pleinement son intégration au pipeline expérimental.

Tableau 5.7 Exemples de configurations explorées par Optuna et performances associées

Trial ID	LR	Dropout	BS	GRU	Kernel	Heads	CW Type	Acc	Sens	Spec
22	0,00025	0,208	16	96	7	4	Manual	0,669	0,618	0,688
23	0,00015	0,106	16	96	7	2	Manual	0,598	0,702	0,564
31	0,00124	0,333	16	64	3	2	Manual	0,589	0,711	0,534
32	0,00072	0,359	16	64	3	2	Manual	0,601	0,729	0,541
38	0,00051	0,374	64	96	3	2	Manual	0,665	0,591	0,697
41	0,00101	0,180	16	96	7	2	Manual	0,682	0,542	0,743
46	0,00039	0,123	64	128	7	2	Balanced	0,687	0,470	0,784

Dans l'ensemble, ces études d'ablation démontrent l'importance de chaque composante du pipeline. Une *étude d'ablation* consiste à évaluer l'impact individuel de chaque élément d'un modèle (en supprimant ou en modifiant une composante à la fois) afin de mesurer sa contribution spécifique à la performance globale. Cette approche permet d'identifier les modules réellement déterminants et de vérifier que l'amélioration des résultats n'est pas due à un artefact ou à une combinaison fortuite de paramètres.

Le rééquilibrage *SMOTE* favorise la sensibilité, le filtrage des valeurs aberrantes améliore la cohérence du corpus et l'optimisation bayésienne consolide la stabilité et la performance du modèle. Ces étapes combinées contribuent à renforcer la fiabilité du système de détection et à en assurer la robustesse dans un contexte clinique réel.

5.4 Comparaison avec les approches du challenge

Dans le cadre du challenge DREAM CODA TB, plusieurs équipes internationales ont proposé des méthodes de classification de la tuberculose à partir des sons de toux, parfois combinées avec des variables cliniques. Comparer notre méthode à ces approches permet de situer nos résultats et, surtout, de mettre en évidence l'impact des choix méthodologiques sur les performances rapportées (Jaganath et al., 2024).

Certaines équipes ont rapporté des scores très élevés, avec des AUC internes proches de 0,85–0,88. Toutefois, une analyse attentive de leurs protocoles révèle des pratiques susceptibles d'introduire des biais, comme l'absence de séparation stricte par participant entre entraînement et validation, ou l'utilisation de variables fortement corrélées au statut TB (par exemple, des antécédents médicaux). Ces choix peuvent amplifier artificiellement la performance en validation, mais réduisent la validité externe et la capacité de généralisation en conditions réelles.

Notre approche se distingue par un protocole plus rigoureux : validation croisée stratifiée par sujet, exclusion de toute variable directement liée au diagnostic de TB, et usage délibérément restreint à des métadonnées simples (âge, sexe, poids, perte de poids). Ce cadre impose des performances plus prudentes (AUC $\sim$ 0,80), mais garantit une robustesse accrue et une meilleure transférabilité clinique.

Tableau 5.8 Comparaison synthétique de notre approche avec deux équipes du challenge CODA TB

Équipe / Approche	Protocole	Variables	AUC (val.)	AUC (test)
Metformin121 (1re place)	Validation non par sujet	Audio + variables cliniques	~0,88	0,832
LCL Classe (5e place)	Validation non par sujet	Audio + clin. corrélées	~0,85–0,87	Non indiqué
Notre approche	Validation stricte par sujet	Audio + métadonnées simples	~0,80	—

5.5 Conclusion

L'ensemble de ces résultats valide l'intérêt d'une approche multimodale audio + métadonnées pour le dépistage de la TB, tout en identifiant les contributions spécifiques de chaque étape du pipeline (rééquilibrage des classes, filtrage des données, agrégation par patient, optimisation). Avant d'aborder les perspectives, le chapitre suivant propose une discussion critique de ces résultats, en soulignant les enseignements méthodologiques, les limites de l'étude et les pistes d'amélioration envisageables.

CHAPITRE 6

DISCUSSION

Dans ce chapitre, nous analysons de manière critique les résultats obtenus et exposés précédemment. Nous mettons en lumière les enseignements méthodologiques tirés de notre étude, les limites actuelles de notre approche, ainsi que les pistes futures d'amélioration pour le modèle de détection de la tuberculose par toux que nous avons développé.

6.1 Analyse des résultats

6.1.1 Performances obtenues et interprétation

Les expériences menées ont montré qu'un modèle basé uniquement sur les caractéristiques audio (Approche A) pouvait déjà atteindre une sensibilité élevée ($\sim 79\,\%$). Cela signifie qu'il détecte la plupart des cas de TB avec les seules informations acoustiques, ce qui est très encourageant pour un outil de triage : on minimise le risque de manquer un malade. Nos résultats rejoignent ainsi l'objectif de nombreux travaux de dépistage automatisé, qui visent en priorité une sensibilité élevée ($\geq 80\,\%$) conformément aux recommandations de l'OMS pour les tests de triage TB (World Health Organization, 2024). Le revers de la médaille, observé dans notre étude, est une spécificité plus faible ($\sim 54\,\%$) en audio seul, indiquant beaucoup de fausses alertes. Ce profil (haute sensibilité, spécificité modérée) est typique des premiers modèles de dépistage par toux et a également été constaté dans le challenge CODA TB, où les modèles acoustiques purs plafonnaient à $\sim 55\,\%$ de spécificité pour $80\,\%$ de sensibilité. Il souligne la difficulté à diagnostiquer de manière très spécifique avec la toux seule, celle-ci étant un symptôme peu spécifique par nature.

L'ajout de métadonnées cliniques (Approche B) a apporté une amélioration notable de l'accuracy globale ($\approx 71\,\%$ vs $61\,\%$) et surtout de la spécificité ($\sim 64\,\%$ vs $\sim 54\,\%$), tout en maintenant la sensibilité au même niveau ($\sim 80\,\%$) que l'audio seul. Cette progression valide l'intérêt d'une approche multimodale. En pratique, des variables simples comme le poids, le sexe (et la perte

de poids lorsqu'elle est disponible) jouent un rôle de filtre contextuel : le modèle devient plus exigeant avant d'annoncer « TB », ce qui réduit les faux positifs. Ce phénomène est cohérent avec d'autres études où l'apport d'informations complémentaires (cliniques, démographiques) a permis de mieux différencier les vrais malades des sujets sains. Concrètement, cela signifie que les métadonnées apportent un éclairage que l'audio seul ne fournit pas : une même toux sera interprétée différemment selon le profil de risque du patient, ce que l'algorithme multimodal capture en partie.

Un point essentiel concerne l'AUC. Contrairement à une éventuelle baisse redoutée, nos résultats indiquent au contraire une amélioration, ou à minima un maintien de la performance. Au niveau *enregistrement*, l'AUC moyenne progresse de ~ 0,72 (A) à ~ 0,80 (B). Au niveau participant, elle passe légèrement de ~ 0,81 (A) à ~ 0,82 (B). Ainsi, l'intégration des métadonnées ne dégrade pas la capacité discriminante du modèle et s'accompagne d'un bénéfice clinique tangible : une spécificité accrue, obtenue sans perte de sensibilité. Pour aller plus loin, il serait pertinent d'explorer des stratégies de fusion audio/métadonnées plus fines ainsi que des méthodes de calibration des scores (par exemple isotonic ou Platt), afin de renforcer encore la spécificité tout en préservant la sensibilité et la qualité globale de discrimination.

6.1.2 Stabilité et variance entre les plis

Un indicateur rassurant apporté par nos expériences est la stabilité des performances entre les différents plis de validation, en particulier pour l'approche multimodale B. Nous avons relevé des écarts-types très faibles (±2–3 %) sur l'accuracy, la sensibilité et la spécificité en cross-validation pour l'approche B, comparés à des variations plus larges (±5–7 %) pour l'approche A. Cette réduction de variance suggère que l'ajout des métadonnées a également un effet de régularisation sur l'apprentissage. En effet, en fournissant des indices additionnels constants pour un même patient, on atténue l'influence des fluctuations propres aux seules données audio. L'apprentissage paraît moins sensible à un enregistrement de toux particulier, et capture davantage une tendance générale. On peut y voir une analogie avec l'ajout de variables explicatives dans un modèle statistique : cela peut réduire la variance de l'estimation si ces variables expliquent une partie du

phénomène. Ici, les métadonnées expliquent en partie la propension d'un patient à avoir la TB (par exemple, la perte de poids est globalement plus fréquente chez les TB), ce qui stabilise le modèle d'un pli à l'autre.

À l'inverse, l'approche audio seule montrait des résultats plus erratiques selon la composition précise des plis. C'est un signe que le modèle A était plus facilement perturbé par des échantillons particuliers (certaines toux non-TB très « atypiques » pouvaient le faire trébucher différemment selon qu'elles se trouvaient en train ou en val). Là encore, l'approche B, en intégrant une vision plus holistique du patient, a lissé ces écarts et rendu le système plus robuste.

Pour le déploiement pratique, cette stabilité accrue est un atout, car elle signifie qu'on peut s'attendre à des performances plus constantes sur différentes cohortes de patients. Cela renforce la confiance dans le modèle : il y a moins de « scénarios cachés » où il s'effondrerait.

6.2 Limites de l'approche actuelle

Malgré les résultats encourageants, notre approche comporte plusieurs limites qu'il convient d'identifier afin de mettre en perspective la portée de nos conclusions :

- **Qualité hétérogène des signaux audio** : Les enregistrements de toux dans notre dataset varient fortement en qualité. Certains fichiers audio sont clairs et exploitables, tandis que d'autres sont très courts, peu sonores, saturés de bruit ambiant, ou encore possiblement mal découpés (toux tronquée, etc.). Cette hétérogénéité complique le travail du modèle et introduit du bruit dans l'apprentissage. Bien que nous ayons appliqué du filtrage de bruit et normalisé les volumes, il reste des disparités. Des études ont montré que les conditions d'enregistrement influencent significativement les performances des algorithmes audios – par exemple des toux enregistrées en clinique dans un environnement calme vs sur le terrain avec du bruit de fond n'ont pas la même clarté (**?**).
- **Dataset de taille restreinte** : Bien que nous disposions de plusieurs centaines d'enregistrements, ceux-ci proviennent d'un nombre limité de participants (quelques dizaines de malades TB et quelques dizaines de contrôles). Ce nombre restreint de sujets limite la diversité des

profils rencontrés et peut conduire à un sur-apprentissage sur les spécificités de ces individus. Même avec une validation croisée rigoureuse, nous ne pouvons pas garantir que le modèle aura le même comportement sur une population entièrement nouvelle, aux caractéristiques possiblement différentes. Ce problème de taille d'échantillon est fréquent dans la littérature sur les sons de toux (Zimmer and Pai, 2022). Dans notre cas, cela signifie que nos métriques (par ex. 80 % de sensibilité, 64 % de spécificité) sont à considérer comme des estimations sur notre dataset, et non comme des garanties absolues en population générale. Il faudra idéalement les confirmer sur un ensemble de test indépendant plus large.

- **Dépendance aux métadonnées et risque de biais** : Si l'ajout de données cliniques a amélioré nos performances, il comporte aussi un risque : celui de sur-apprendre des corrélations spécifiques au dataset qui ne seraient pas de véritables causalités universelles.

Notre approche, bien qu'efficace sur nos données, doit être considérée comme un prototype perfectible. L'environnement d'enregistrement contrôlé, le volume d'échantillons limité et l'éventuel biais des variables d'entrée sont autant de facteurs qui appellent à la prudence quant à une utilisation clinique sans approfondissement. Les sections suivantes proposent justement des axes d'amélioration pour lever ces limitations.

6.3 Pistes d'amélioration du modèle

Plusieurs directions peuvent être envisagées pour améliorer la performance et la robustesse de notre système de classification de la tuberculose à partir des toux

6.3.1 Utilisation d'audio brut ou de spectrogrammes 2D complets

Jusqu'à présent, nous avons exploité des caractéristiques audio manuelles (MFCC, ZCR, etc.) qui résument le signal. Une piste complémentaire consiste à tirer parti d'entrées plus riches, soit l'audio brut, soit des représentations temps–fréquence 2D complètes (spectrogrammes de Mel ou log-fréquence) traitées comme de véritables images. Des réseaux convolutionnels 2D pré-entraînés sur de larges corpus audio, notamment AudioSet, fournissent alors des embeddings

transférables qui capturent des motifs temps–fréquence subtils échappant aux descripteurs standards (Hershey et al., 2017; Gemmeke et al., 2017; Kong et al., 2020). Dans ce cadre, un pipeline typique consiste à extraire un embedding sur un spectrogramme complet via un CNN (p. ex. une variante VGG/ResNet pré-entraînée AudioSet) puis à l'intégrer au classifieur principale (fusion avec l'encodeur séquentiel ou remplacement partiel de celui-ci)d.

Des résultats récents confirment l'intérêt de combiner caractéristiques « traditionnelles » et features apprises sur spectrogrammes. Sur un corpus hospitalier contrôlé (≈ 456 segments de toux : 230 provenant de 70 patients TB et 226 de 74 sujets sains), Xu et al. montrent qu'un modèle Bi-LSTM entraîné sur un jeu de descripteurs standards (MFCC, ZCR, énergie courte durée, RMS, chroma) atteint environ 94 % d'accuracy, portée à ≈ 96 % lorsqu'on y adjoint des caractéristiques extraites par un CNN 2D sur spectrogrammes Ċe type de fusion illustre la complémentarité entre indices acoustiques agrégés et motifs locaux appris. En pratique, l'augmentation de dimension induite par les images 2D impose du pré-entraînement (AudioSet) et une data augmentation adaptée ; le bénéfice attendu est une meilleure sensibilité sans sacrifier la spécificité, sous réserve d'une validation sujets-indépendants et d'une calibration soignée (Hershey et al., 2017; Kong et al., 2020).

6.3.2 Architectures avancées pour le signal audio

En parallèle des approches CNN 2D, l'utilisation de modèles de nouvelle génération dans le domaine audio pourrait booster les performances. Deux orientations se démarquent :

- Modèles self-supervised pré-entraînés sur l'audio brut : Des modèles tels que Wav2Vec 2.0 ont démontré qu'il est possible d'apprendre directement à partir des formes d'onde audio de riches représentations internes, sans supervision, en profitant de larges bases de données (Baevski et al., 2020). Des approches auto-supervisées dédiées aux sons de santé (toux, souffle, etc.) se développent également (Baur et al., 2024; Liu et al., 2023).

- Transformers audio temporels : Des architectures avec auto-attention peuvent mieux gérer les dépendances longues et la structure multi-phases d'un épisode de toux ; ces idées sont activement explorées dans les tâches audio de santé (Baur et al., 2024).

6.3.3 Fusion multimodale avancée et intégration d'autres modalités

Dans ce travail, les métadonnées ont été intégrées de façon relativement simple, par une concaténation directe à un certain niveau du réseau. Une amélioration possible consisterait à explorer des approches de fusion multimodale plus élaborées (p. ex. mécanismes d'attention croisée, stratégies de fusion hiérarchique). De plus, la multimodalité pourrait être élargie par l'ajout de nouvelles sources de données complémentaires :

- Parole et voix : des altérations vocales ou prosodiques peuvent témoigner d'une détresse respiratoire et fournir des indices additionnels (piste exploratoire).
- Imagerie thoracique (p. ex. CAD4TB) : la combinaison d'un score acoustique avec un indice radiologique pourrait améliorer la précision du triage (voir panorama récent sur les biomarqueurs numériques de la toux et leurs applications cliniques) (Zimmer and Pai, 2022).
- Suivi longitudinal de la toux : la mesure répétée de la fréquence de toux au cours du traitement est étudiée comme un biomarqueur numérique prometteur (Proaño et al., 2018; Huddart et al., 2023).

6.4 Conclusion

En définitive, nos résultats montrent qu'un modèle fondé uniquement sur l'audio peut déjà atteindre une sensibilité élevée, mais au prix d'une spécificité limitée. L'intégration de métadonnées cliniques simples apporte une amélioration tangible, en particulier sur la réduction des faux positifs, et rend le modèle plus stable d'un pli à l'autre. Ces apports confirment l'intérêt d'une approche multimodale, tout en soulignant la nécessité de protocoles rigoureux pour garantir la validité externe. Néanmoins, la taille restreinte et l'hétérogénéité des données

imposent de poursuivre les travaux sur des corpus plus vastes et diversifiés. Ce chapitre met ainsi en évidence le potentiel réel mais encore perfectible de notre système, et ouvre la voie à des développements futurs vers un outil de dépistage fiable et transférable en pratique clinique.

CONCLUSION ET RECOMMANDATIONS

Le présent mémoire a porté sur le développement d'un système de détection automatique de la tuberculose à partir de sons de toux, en mobilisant des approches modernes de traitement du signal et d'intelligence artificielle. La tuberculose demeure, encore aujourd'hui, l'une des maladies infectieuses les plus préoccupantes sur le plan mondial. En dépit des progrès réalisés en matière de diagnostic et de traitement, les contraintes logistiques, économiques et techniques des méthodes traditionnelles compromettent souvent le dépistage rapide et efficace, particulièrement dans les contextes à faibles ressources. Dans ce cadre, l'exploitation de la toux comme biomarqueur, combinée à des techniques avancées d'apprentissage automatique, représente une voie innovante et prometteuse pour proposer des solutions accessibles, non invasives et à faible coût.

L'approche adoptée dans ce mémoire a permis de mettre en place un pipeline complet, rigoureux et reproductible. Celui-ci comprend plusieurs étapes clés : le prétraitement et la segmentation des enregistrements, l'extraction de caractéristiques acoustiques pertinentes, l'augmentation artificielle des données afin de compenser les déséquilibres du corpus, ainsi que la conception et l'entraînement d'architectures neuronales profondes intégrant des couches convolutionnelles, récurrentes et des mécanismes d'attention. L'optimisation systématique des hyperparamètres à l'aide de la librairie *Optuna* a également joué un rôle essentiel dans l'amélioration de la robustesse et de la performance du modèle. Les résultats obtenus, bien que perfectibles, confirment la faisabilité de cette approche et soulignent son potentiel comme outil complémentaire de dépistage.

Les principales contributions de ce mémoire résident ainsi dans la conception d'une méthodologie intégrée et dans l'exploration comparative de différentes représentations acoustiques, notamment les coefficients cepstraux de type MFCC, les spectrogrammes et d'autres descripteurs spectraux. L'analyse approfondie de ces représentations a permis de mieux comprendre l'influence de la nature des caractéristiques sur les performances de classification. De plus, l'architecture

hybride combinant CNN et GRU enrichie par un mécanisme d'attention a montré des résultats encourageants, en parvenant à capter à la fois les structures locales et temporelles des signaux de toux. Enfin, la mise en œuvre d'une stratégie d'optimisation avancée a constitué un apport méthodologique important, garantissant une recherche systématique des meilleures configurations de modèles.

Ce travail comporte néanmoins des limites qu'il convient de souligner. D'une part, la diversité du corpus exploité, bien que notable, reste insuffisante pour refléter pleinement la variété des contextes cliniques, linguistiques et géographiques associés à la tuberculose. La variabilité des enregistrements et des conditions de collecte, en particulier la présence de bruits ambiants ou l'hétérogénéité des dispositifs de capture, a également influencé les performances des modèles. D'autre part, la validation a été effectuée à partir d'un unique jeu de données, ce qui limite la généralisabilité des conclusions. Une réplication sur d'autres corpus indépendants, idéalement issus de contextes cliniques réels et validés médicalement, apparaît nécessaire pour renforcer la fiabilité des résultats.

Ces limites ouvrent cependant plusieurs perspectives de recherche stimulantes. L'une des priorités consisterait à enrichir les bases de données existantes par des enregistrements collectés dans des environnements variés, afin de mieux représenter la diversité des patients et des contextes d'utilisation. L'exploration d'approches multimodales, combinant les sons de toux avec d'autres informations biomédicales telles que les sons respiratoires, les données cliniques ou même des images médicales, pourrait permettre de concevoir des systèmes de dépistage plus précis et plus complets. L'intégration d'un tel outil dans des plateformes mobiles ou connectées représenterait également une avancée majeure en matière d'accessibilité, en rendant possible le dépistage précoce dans des zones éloignées ou à faibles ressources. Enfin, l'essor des méthodes d'intelligence artificielle explicable (XAI) constitue une piste incontournable pour favoriser la transparence et la compréhension des décisions des modèles. Une meilleure interprétabilité

renforcerait non seulement la confiance des cliniciens, mais faciliterait également l'intégration de ces technologies dans la pratique médicale courante.

En définitive, ce mémoire illustre comment l'intelligence artificielle et le traitement du signal appliqués aux sons de toux peuvent contribuer à relever un défi de santé publique majeur. En proposant une approche innovante, accessible et adaptable, ce travail s'inscrit dans une dynamique de recherche visant à améliorer le dépistage de la tuberculose et, à terme, à soutenir les efforts internationaux de lutte contre cette maladie. S'il ne constitue qu'une étape dans un processus de recherche en constante évolution, il met néanmoins en lumière le potentiel considérable de ces approches pour transformer les pratiques médicales et renforcer l'équité dans l'accès aux soins.

BIBLIOGRAPHIE

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna : A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Barton, A., Gaydecki, P., Holt, K., and Smith, J. A. (2012). Data reduction for cough studies using distribution of audio frequency content (spectrogrammes de toux). PubMed Central (Respiratory Research).

Baur, S., Nabulsi, Z., Weng, W.-H., Garrison, J., Blankemeier, L., Fishman, S., Chen, C., Kakarmath, S., et al. (2024). Hear – health acoustic representations. *arXiv*.

Boehme, C. C., Nabeta, P., Hillemann, D., and et al. (2010). Rapid molecular detection of tuberculosis and rifampin resistance. *New England Journal of Medicine*, 363(11) :1005–1015.

Botha, G. (2021). Detection of tuberculosis by automatic cough sound analysis.

Brosch, R., Gordon, S. V., Marmiesse, M., and et al. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences*, 99(6) :3684–3689.

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., and Mascolo, C. (2020). Exploring automatic diagnosis of covid-19 from crowdsourced cough sounds. *arXiv preprint arXiv :2006.05919*.

Celeste Jr, J. et al. (2025). A software pipeline for systematizing machine learning of speech data. *Frontiers in Psychiatry*.

Centers for Disease Control and Prevention (2000). Tuberculosis chest x-ray with cavity (right upper lobe). Wikimedia Commons.

Chatham County Public Health Department (2025). Tb facts. Accessed : 2025-09-24.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. GRU architecture.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 28, pages 357–366. IEEE.

Deshpande, G., Schuller, B., and Cummins, N. (2022). Cough sound analysis : A review of machine learning and deep learning approaches. *Frontiers in Digital Health*, 4 :953934.

Dubey, D., Rath, S., Sahu, M., Debata, N., and Padhy, R. (2012). Antimicrobials of plant origin against tb and other infections and economics of plant drugs – introspection. *Indian Journal of Traditional Knowledge*, 11(2) :225–233.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*.

Flynn, J. L. and Chan, J. (2001). Immunology of tuberculosis. *Annual Review of Immunology*, 19 :93–129.

Frost, G. T., Theron, G., and Niesler, T. (2022). Tb or not tb ? acoustic cough analysis for tuberculosis classification. In *Interspeech*.

Garcia, M. and Chen, L. (2021). Cough sound analysis for pulmonary disease detection : a review. *Computer Methods and Programs in Biomedicine*. Référence générique de revue.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., et al. (2017). Audio set : An ontology and human-labeled dataset for audio events. In *ICASSP*.

Golden, M. P. and Vikram, H. R. (2005). Extrapulmonary tuberculosis : an overview. *American Family Physician*, 72(9) :1761–1768.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Hershkovitz, I., Donoghue, H. D., Minnikin, D. E., and et al. (2008). Detection and molecular characterization of 9000-year-old *Mycobacterium tuberculosis* from a neolithic settlement in the eastern mediterranean. *PLoS ONE*, 3(10) :e3426.

Houben, R. M. G. J. and Dodd, P. J. (2016). The global burden of latent tuberculosis infection : a re-estimation using mathematical modelling. *PLOS Medicine*, 13(10) :e1002152.

Huddart, S., Gama, M. A., Olaru, A. F., and et al., S. T. (2024). Solicited cough sound analysis for tuberculosis triage testing : The coda tb dream challenge dataset. *Scientific Data*, 11(1) :123.

Huddart, S., Pillinger, R., Rudgard, W. E., et al. (2023). A novel digital biomarker for tb diagnosis and treatment monitoring : Continuous cough monitoring in adults with pulmonary tuberculosis. *NPJ Digital Medicine*, 6(1) :38.

Husssain, S. et al. (2024). Cough2covid-19 detection using an enhanced multi layer ensemble deep learning framework and coughfeatureranker. *Scientific Reports*. preprocessing phase included noise reduction, resampling, segmentation, amplitude normalization.

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, S., Ali, K., and Nabeel, M. (2020). Ai4covid-19 : Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20 :100378.

Imran, M., Javaid, M., and Nabeel, M. (2022). Cough-based tuberculosis detection using hybrid deep learning with attention mechanism. *IEEE Access*, 10 :10345–10357.

Jaganath, D. et al. (2024). Results from the coda tb dream challenge (cough-based tb diagnosis). *medRxiv*.

Kafentzis, G., Tetsing, S., Brew, J., et al. (2023). Predicting tuberculosis from real-world cough audio recordings and metadata. *arXiv*.

Kapetanidis, P., Votis, K., and Tzovaras, D. (2024). Respiratory diseases diagnosis using audio analysis and artificial intelligence : A systematic review. *Sensors*, 24.

Kendall, B. A. and Furin, J. J. (2021). The radiographic diagnosis of pulmonary tuberculosis in hiv-positive patients. *Journal of Thoracic Imaging*, 36(1) :W33–W41.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1137–1143, San Mateo, CA. Morgan Kaufmann.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns : Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :2880–2894.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. IEEE.

Liu, Z., Chai, K., Hu, P., and Chen, Y. (2023). Self-supervised contrastive learning for medical time series : A systematic review. *Bioengineering*, 10(5) :614.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. Technical report, Cambridge Research Laboratory.

Lönnroth, K., Jaramillo, E., Williams, B. G., Dye, C., and Raviglione, M. (2010). Drivers of tuberculosis epidemics : the role of risk factors and social determinants. *Social Science & Medicine*, 68(12) :2240–2246.

MacPherson, P., Houben, R. M. G. J., Glynn, J. R., and et al. (2014). Prevalence and risk factors for latent tuberculosis infection in rural malawi. *International Journal of Tuberculosis and Lung Disease*, 18(2) :196–202.

Mazurek, G. H., Jereb, J., Vernon, A., LoBue, P., Goldberg, S., and Castro, K. (2010). Updated guidelines for using interferon gamma release assays to detect *Mycobacterium tuberculosis* infection — united states, 2010. Technical Report RR-5, MMWR Recommendations and Reports.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa : Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*.

Menzies, D., Pai, M., and Comstock, G. (2007). Meta-analysis : New tests for the diagnosis of latent tuberculosis infection : areas of uncertainty and recommendations for research. *Annals of Internal Medicine*, 146(5) :340–354.

Nguyen, T. and Patel, R. (2020). Attention mechanisms improve event-focused acoustic classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Référence générique pour l'attention en audio.

Nidadavolu, P. S., Villalba, J., and Dehak, N. (2020). Extracting robust and interpretable features for cough sound classification. In *Proceedings of Interspeech*, pages 2337–2341.

Orlandic, L., Teijeiro, T., and Atienza, D. (2021). The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*.

Pahar, M., Klopper, M., Reeve, B., Theron, G., Warren, R., and Niesler, T. (2021). Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, 42(10) :105012.

Pahar, M., Klopper, M., Reeve, B., Warren, R., Theron, G., Diacon, A., and Niesler, T. (2022). Automatic tuberculosis and covid-19 cough classification using deep learning. *arXiv preprint arXiv :2205.05480*.

Pai, M., Behr, M. A., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C. C., Ginsberg, A. M., Swaminathan, S., Spigelman, M., Getahun, H., and Menzies, D. (2016). Tuberculosis. *Nature Reviews Disease Primers*, 2 :16076.

Pai, M., Kasaeva, T., and Swaminathan, S. (2022). Covid-19's devastating effect on tuberculosis care — a path to recovery. *New England Journal of Medicine*, 386(16) :1490–1493.

Partnership, S. T. (2024). Ai-powered cough analysis and monitoring.

Pramono, R. X., Imtiaz, S. A., and Rodriguez-Villegas, E. (2017). Cough sound analysis and its applications in diagnosis and monitoring of respiratory diseases : A review. *Biomedical Engineering Online*, 16(1) :1–28.

Proaño, A., Bravard, M. A., Lopez, J. W., Lee, G. O., Bui, D., Arevalo, J., Zimic, M., Marin, J., Comina, G., Cáceres, O., Caviedes, L., Ticona, E., Valencia, T., Oberhelman, R. A., and Checkley, W. (2018). Cough frequency during treatment associated with baseline cavitary volume and proximity to the airway in pulmonary tuberculosis. *Chest*, 154(4) :888–897.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall.

Rabiner, L. R. and Schafer, R. W. (2011). *Theory and Applications of Digital Speech Processing*. Pearson, Upper Saddle River, NJ.

Rajasekar, S. J. S., Balaraman, A. R., et al. (2024). Detection of tuberculosis using cough audio analysis : a deep learning approach with capsule networks. *Discover Artificial Intelligence*, 4 :77.

Sahoo, R. K., Sinha, A., Mishra, M., et al. (2025). A systematic review and meta-analysis of the diagnostic accuracy of artificial intelligence in detecting tuberculosis using cough sounds. *SSRN*.

Scikit-learn developers (2011). Stratifiedgroupkfold — scikit-learn documentation. Consulté le 26 août 2025.

Serrurier, A., Roy, J.-P., Dutoit, T., Laroche, C., Desprez, P., et al. (2022). Past and trends in cough sound acquisition, automatic detection and classification. *Sensors*.

Sounderajah, V. and et al. (2021). A systematic review of artificial intelligence in screening for tuberculosis. *NPJ Digital Medicine*, 4 :135.

Suda, C. (2023a). Early detection of tuberculosis with machine learning cough audio analysis : Towards more accessible global triaging usage. *arXiv preprint arXiv :2310.17675*.

Suda, C. (2023b). Early detection of tuberculosis with machine learning cough audio analysis : Towards more accessible global triaging usage. *arXiv preprint arXiv :2310.17675*.

Supply, P., Marceau, M., Mangenot, S., and et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature Genetics*, 45(2) :172–179.

Tellier, R., Li, Y., Cowling, B. J., and Tang, J. W. (2019). Recognition of aerosol transmission of infectious agents : a commentary. *BMC Infectious Diseases*, 19(1) :101.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, volume 10, pages 293–302.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Wong, T. W. S. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, (9).

World Health Organization (2021a). *WHO Consolidated Guidelines on Tuberculosis. Module 2 : Screening — Systematic Screening for Tuberculosis Disease*. World Health Organization, Geneva.

World Health Organization (2021b). Who operational handbook on tuberculosis. module 2 : Screening — systematic screening for tb disease (algorithmes). WHO (PDF).

World Health Organization (2023a). Global tuberculosis report 2023. Technical report, World Health Organization, Geneva.

World Health Organization (2023b). Systematic screening for tuberculosis : updated guidelines.

World Health Organization (2023c). Tb incidence global tuberculosis report 2023 (cartes et graphiques). WHO website.

World Health Organization (2024). Target product profile (tpp) for tuberculosis triage testing public consultation (2024). Defines accuracy targets for TB triage tests.

Xu, M., Zhang, Y., Wang, H., Zhou, R., Gao, Z., and Liu, W. (2023). Dmrnet : Dynamic multi-branch recurrent network for cough-based tuberculosis detection. *IEEE Journal of Biomedical and Health Informatics*, 27(2) :943–954.

Xu, W., Yuan, H., Lou, X., Chen, Y., and Liu, F. (2022). Dmrnet based tuberculosis screening with cough sound. *IEEE Access*.

Yadav, J., Liu, H., Wu, H., et al. (2024). Audiovisual multimodal cough data analysis for tuberculosis detection. In *2024 IEEE International Conference on Information Intelligence Systems and Applications (IISA)*.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1) :32–35.

Zhang, L., Ding, S., Qian, K., et al. (2022). A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs. *Sensors*, 22(21) :8421.

Zimmer, A. J. and Pai, M. (2022). Making cough count in tuberculosis care. *Communications Medicine*, 2(1) :68.