Exploration des modèles d'apprentissage statistique profonds couplés à la spectrométrie de masse pour améliorer la surveillance épidémiologique des maladies infectieuses

Noshine Mohammad

Sorbonne Université

Institut Pierre-Louis Epidémiologie et Santé Publique
UMR S 1136

Ecole Doctorale 393 Pierre Louis de Santé Publique à Paris

Épidémiologies et Sciences de l'information biomédicale

Équipe de recherche Maladies Transmissibles : Surveillance et Modélisation

Exploration des modèles d'apprentissage statistique profonds couplés à la spectrométrie de masse pour améliorer la surveillance épidémiologique des maladies infectieuses

Par **Noshine Mohammad**

Thèse de doctorat de Biostatistique et Biomathématiques

Présentée et soutenue publiquement le 7 Décembre 2023

Devant un jury composé de :

Raphaelle Metras, Chargée de recherche et HDR, Sorbonne Université, **Présidente du Jury**
Guillaume Desoubeaux, Professeur des Universités-Praticien Hospitalier, Université de Tours, **Rapporteur**
Arthur Tenenhaus, Professeur des Universités et Directeur de recherche, Centrale Supelec - Paris Saclay, **Rapporteur**
Laurence Lachaud, Professeure des Universités-Praticien Hospitalier, Université de Montpellier, **Examinatrice**
Antonin Lamazière, Professeur des Universités-Praticien Hospitalier, Sorbonne Université, **Examinateur**
Renaud Piarroux, Professeur des Universités-Praticien Hospitalier, Sorbonne Université, **Directeur de thèse**

Remerciements

Je voudrais exprimer ma gratitude envers toutes les personnes qui m'ont accompagné et soutenu durant ces trois années de thèse, sans lesquelles l'aboutissement de cette aventure n'aurait pas été possible. Un merci tout particulier à mes proches pour leur soutien infaillible.

J'adresse mes remerciements à mon directeur de thèse, le Professeur Renaud Piarroux, Chef du Service de Parasitologie-Mycologie de l'Hôpital de la Pitié-Salpêtrière à Paris.

Je tiens à exprimer ma sincère gratitude envers le Professeur Xavier Tannier, mon encadrant en informatique, pour son soutien continu tout au long de ma thèse. Je le remercie d'avoir accepté de superviser mes travaux, d'avoir été constamment disponible, et de m'avoir fait confiance pour mener à bien ces recherches. Son encadrement, ses précieux conseils, et son expertise ont grandement contribué à la mise en place des méthodes de recherche. Les réunions hebdomadaires ont joué un rôle essentiel dans le développement des travaux que nous présentons dans ce manuscrit.

Je remercie chaleureusement Cécile Nabet pour notre collaboration fructueuse dans la réalisation des publications de ce manuscrit. Je lui suis reconnaissante pour ses précieuses connaissances en biologie, ainsi que nos échanges passionnants et réguliers dans le domaine de la recherche.

Je remercie Alexandre Godmer pour son soutien et son aide constants tout au long de ces trois années de thèse. Je le remercie pour notre collaboration sur des projets de recherche, ainsi que pour l'opportunité qu'il m'a offerte de participer à des activités de vulgarisation de la recherche et de m'impliquer dans l'enseignement. Je le remercie pour nos échanges réguliers et passionnants, ainsi que pour son enthousiasme, ses conseils, et sa disponibilité.

Je remercie Jean Yves Brossas pour avoir partagé son expertise précieuse en spectrométrie de masse et pour son soutien actif dans les différentes études sur lesquelles j'ai eu l'opportunité de travailler.

J'adresse mes remerciements envers toutes les personnes qui ont contribué aux études présentées dans ce manuscrit de thèse.

J'adresse mes remerciements à l'équipe du Service de Parasitologie-Mycologie de l'Hôpital de la Pitié-Salpêtrière, avec laquelle j'ai eu le plaisir de travailler et d'échanger au cours de ces trois années de thèse.

Je remercie l'Institut Pierre-Louis d'Épidémiologie et de Santé Publique pour m'avoir accueilli au sein de leur école doctorale. Mes remerciements vont également à la Région Île-de-France et à l'entreprise Cerba Health Care pour avoir cofinancé ma thèse.

Je tiens à exprimer ma profonde gratitude envers mes amis pour leur soutien constant. En particulier, je remercie chaleureusement Diego Belliard, un ami proche, pour sa présence continue tout au long de cette thèse, son soutien infaillible, et son aide précieuse dans la rédaction de ce manuscrit. Je tiens également à remercier Laura Couché-Puliccani pour son soutien indéfectible.

Enfin, je tiens à remercier infiniment mes proches et ma famille pour m'avoir soutenu tout au long de ces années d'études, pour leur amour inconditionnel, et leur encouragement constant.

À l'attention du lecteur.../Note to readers...

Je tiens à exprimer ma gratitude envers toutes les personnes qui prendront le temps de parcourir mon manuscrit. Pour toute question, information supplémentaire ou demande, je vous invite à me contacter par courriel à l'adresse suivante : noshine.mohammad@gmail.com, ou à me suivre sur ma page LinkedIn via le lien suivant : www.linkedin.com/in/noshine-mohammad-phd.

Pour toute consultation des codes sources utilisés dans cette étude, je vous encourage à consulter mon profil GitHub à l'adresse suivante : https://github.com/NoshineMo. Vous y trouverez l'ensemble des codes qui ont permis de générer les résultats présentés dans ce manuscrit.

English version

I would like to express my sincere thanks to all of you who will take the time to read my manuscript. Please feel free to email me at noshine.mohammad@gmail.com or follow me on my LinkedIn page at www.linkedin.com/in/noshine-mohammad-phd for any questions, additional information or requests.

I encourage you to visit my GitHub profile at the following address to view the source code used in this study : https://github.com/NoshineMo. There you will have access to all the code for the generation of the results presented in this manuscript.

Table des matières

Résumé **22**

1 Introduction **25**

1.1 Contexte . . . 25

1.2 Objectif . . . 26

1.3 Plan du mémoire . . . 27

Abstract **24**

2 État de l'art **29**

2.1 La spectrométrie de masse par MALDI-TOF . . . 29

2.1.1 Principes et appareillage . . . 29

2.1.2 Comment identifier des micro-organismes à l'aide des banques de données ? . . . 30

2.1.3 Les caractéristiques du spectre . . . 31

2.2 Le traitement des spectres . . . 33

2.2.1 Lissage . . . 34

2.2.2 Soustraction de la ligne de base . . . 35

2.2.3 Normalisation . . . 36

2.2.4 Détection des pics . . . 36

2.2.5 Alignement des pics . . . 37

2.3 Les transformées du signal . . . 38

2.3.1 La Transformée de Fourier . . . 38

2.3.2 La Transformée en Ondelette . . . 39

2.3.3 Discussion générale sur le traitement des spectres . . . 41

2.4 Spectres MALDI-TOF, approches statistiques et apprentissage automatique . . . 41

2.4.1 Approches statistiques classiques . . . 42

2.4.2 Apprentissage automatique supervisé . . . 42

2.4.3 Qu'en est-il de l'apprentissage profond ? . . . 44

2.5 Principes d'évaluations . . . 45

2.6 Conclusion . . . 47

3 Modèle d'apprentissage profond couplé à la protéomique pour détecter les clones d'épidémies de levure dans plusieurs hôpitaux **49**

3.1 Contexte . . . 49

3.2 Objectif . . . 50
3.3 Acquisition et pré-traitement des données . . . 51
3.3.1 Isolats . . . 51
3.3.2 Diversité génétique et sensibilité au fluconazole . . . 51
3.3.3 Acquisition des spectres de masse MALDI-TOF . . . 52
3.3.4 Analyse des données de spectrométrie de masse MALDI-TOF . . . 52
3.3.5 Alignement . . . 53
3.4 Apprentissage automatique profond, évaluation et conception de l'étude . . . 53
3.4.1 Apprentissage automatique . . . 53
3.4.2 Méthode et mesures d'évaluation . . . 54
3.4.3 Conception de l'étude . . . 55
3.5 Résultats . . . 55
3.5.1 Diversité génétique . . . 55
3.5.2 Sensibilité au fluconazole . . . 56
3.5.3 Impact de la machine et de l'alignement avec MSIWarp . . . 56
3.5.4 Impact du milieu de culture par machine . . . 57
3.5.5 Impact de l'âge de la culture sur milieu de Sabouraud . . . 58
3.6 Discussion . . . 59
3.6.1 Points forts et points faibles de l'étude . . . 59
3.6.2 Observation générale . . . 60
3.6.3 Comparaison avec des études antérieures associant la technologie MALDI-TOF à l'apprentissage automatique . . . 61
3.6.4 Perspectives . . . 61
3.7 Conclusion . . . 62
3.8 Études associées . . . 62
3.8.1 Identification de clones d'*Aspergillus flavus* par MALDI-TOF et apprentissage profond . . . 62
3.8.2 Transmission nosocomiale d'*Aspergillus flavus* en unité néonatale : persistance et détection via MALDI-TOF et CNN . . . 62

4 Estimation précise de l'âge des moustiques anophèles par régression de réseau de neurones pour améliorer la surveillance épidémiologique de la transmission du paludisme 65
4.1 Contexte . . . 65
4.2 Objectif . . . 66
4.3 Collecte et acquisition des données . . . 66
4.3.1 Collecte sur le terrain et élevage de moustiques . . . 66
4.3.2 Préparation des échantillons pour la SM MALDI-TOF et acquisition des spectres de masse 67
4.3.3 Profilage protéique et préparation des données pour l'analyse de l'apprentissage profond . 67
4.4 Méthodes d'apprentissage profond et évaluation . . . 68
4.4.1 Modèles d'apprentissage profond . . . 68
4.4.2 Méthodes d'apprentissage profond pour la prédiction de l'âge . . . 70
4.4.3 Performances et mesures de prédiction . . . 70

4.4.4 Modélisation de la structure d'âge des populations de moustiques . . . 71
4.5 Résultats . . . 71
4.5.1 Validation de l'approche MALDI-TOF-DL pour les moustiques anophèles collectés sur le terrain . . . 71
4.5.2 Optimiser la précision à l'aide de nouvelles méthodes de prédiction. . . . 75
4.5.3 Généraliser notre modèle d'apprentissage profond à de nouvelles populations cibles . . . 79
4.5.4 Simulation de populations de moustiques anophèles sauvages . . . 83
4.6 Discussion . . . 85
4.6.1 Points forts et perspectives de cette étude . . . 85
4.6.2 Comparaison avec les autres approches méthodologiques . . . 86
4.6.3 Limites de l'étude . . . 86
4.7 Conclusion . . . 87

5 Exploration novatrice des modèles d'apprentissage statistique profonds pour l'analyse des spectres MALDI-TOF en épidémiologie des maladies infectieuses 89

5.1 Contexte . . . 89
5.2 Objectif . . . 90
5.3 Données . . . 91
5.3.1 Constitution des cohortes et définition des tâches . . . 91
5.3.2 Acquisition des spectres de masse pour les différentes cohortes . . . 93
5.4 Modèles avec différents contextes d'entrée . . . 94
5.4.1 Les réseaux de neurones à convolution 1D . . . 94
5.4.2 Les réseaux de neurones récurrents . . . 95
5.4.3 Les réseaux de neurones à convolution 2D pour les spectrogrammes . . . 97
5.4.4 Les réseaux de neurones à convolution 2D pour les scalogrammes . . . 99
5.4.5 Les Autoencodeurs . . . 101
5.4.6 Détails d'implémentation . . . 103
5.5 Résultats . . . 104
5.5.1 Performances d'identification et de prédiction . . . 104
5.5.2 Temps d'execution et consommation en énergie lors de l'entraînement des modèles . . . 109
5.6 Discussion . . . 110
5.6.1 Points forts de l'étude . . . 110
5.6.2 Comparaison avec des récentes études associant la technologie MALDI-TOF à l'apprentissage automatique . . . 111
5.6.3 Limites de l'étude . . . 112
5.6.4 Perspectives de l'étude . . . 112
5.7 Conclusion . . . 113

6 Conclusion et Perspectives 115

6.1 Synthèse des études . . . 115
6.2 Originalité de cette thèse . . . 116
6.3 Limites . . . 117

6.4 Perspectives . . . 117
6.4.1 Les spectres de masse MALDI-TOF . . . 117
6.4.2 Les modèles d'apprentissage statistiques profonds . . . 118
6.4.3 Les données . . . 118
6.4.4 Les applications . . . 118

Bibliographie **134**

Production scientifique dans le cadre de la thèse **137**

Publications issues de la thèse et présentations de posters **138**

Annexe A **138**

Annexe B **138**

Annexe C **138**

Table des figures

2.1 Appareil MALDI-TOF . . . 29

2.2 Principe de la spectrométrie de masse MALDI-TOF - désorption et ionisation assistée par une matrice avec détection en temps de vol. . . . 30

2.3 Création de banques d'identification avec la spectrométrie de masse. . . . 31

2.4 Exemple d'un spectre de moustique à 28 jours de l'espèce *An. stephensi* de la banque âge obtenue par MALDI-TOF. . . . 32

2.5 Visualisation des pics 8125 et 8257 du spectre précédent obtenue par MALDI-TOF. . . . 32

2.6 Forme et résolution d'un pic. . . . 33

2.7 Récapitulatif des spécifications techniques abordées dans ce chapitre. . . . 48

3.1 Visualisation de quatre spectres pré-traités de Clones et de Non-Clones de *Candida parapsilosis*. 51

3.2 Prétraitement étape par étape des spectres, du spectre brut aux spectres traités, avant leur utilisation dans la phase d'apprentissage automatique. . . . 52

3.3 Architecture du modèle CNN créé et entraîné avec un ensemble de données. . . . 54

3.4 Organigramme de la conception de l'étude. . . . 55

3.5 Dendrogramme des 96 isolats de *Candida parapsilosis* . . . 56

4.1 Architectures des modèles d'apprentissage profond pour prédire l'âge des moustiques anophèles à partir des spectres de masse MALDI-TOF. . . . 69

4.2 Validation de la SM MALDI-TOF couplée à l'apprentissage profond pour prédire l'âge des moustiques anophèles de terrain . . . 73

4.3 Application de nouvelles techniques de prédiction par apprentissage profond pour améliorer la précision de la prédiction de l'âge des moustiques anophèles. . . . 76

4.4 Comparaison des performances des techniques de prédiction et des modèles de réseaux de neurones pour la prédiction de l'âge des moustiques anophèles. . . . 78

4.5 Prédiction des distributions d'âge de populations simulées de moustiques anophèles sauvages avant et après l'intervention d'une moustiquaire imprégnée d'insecticide. . . . 84

5.1 Pré-traitement des spectres sur CNN et TCN. . . . 95

5.2 Architectures du CNN (gauche) et du TCN (droite) . . . 95

5.3 Pré-traitement des spectres sur RNN-BiGRU et TCN. . . . 96

5.4 Architectures du RNN-BiGRU (gauche) et du ESN (droite) . . . 97

5.5 Pré-traitement des spectres sur 2DCNN spectrogram et 2DCNN BiGRU - Hybrid spectrogram. . 98

5.6 Architectures du 2DCNN spectrogram (haut) et du 2DCNN BiGRU - Hybrid spectrogram (bas) 99

5.7 Pré-traitement des spectres sur 2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram. . . . 100

5.8 Architectures du 2DCNN scalogram (haut) et du 2DCNN BiGRU - Hybrid scalogram (bas) . . . 101

5.9 Pré-traitement des spectres sur les Autoencodeurs DAE et DCAE. . . . 102

5.10 Architectures des Autoencodeurs DAE fc (haut) et DCAE (bas) 103
5.11 Temps de calculs des modèles 109
5.12 Consommation en énergie en Kilowatt-heure (kWh) des modèles 110

Liste des tableaux

3.1 Impact de la machine et de l'alignement avec MSIWarp. 57
3.2 Impact du milieu de culture par machine. 58
3.3 Impact de l'âge de la culture. 59

4.1 Performance de la prédiction de l'âge par MALDI-TOF MS couplée au CNN et à une classification conventionnelle ou cohérente avec les rangs, à partir des spectres de masse du thorax. 74
4.2 Performances de la prédiction de l'âge et capacité de généralisation en utilisant la SM par MALDI-TOF couplé à un CNN et à une régression, à partir des spectres de masse des pattes, de la tête et du thorax. 80
4.3 Performances de la prédiction de l'âge et capacité de généralisation en utilisant la SM par MALDI-TOF couplé au TCN et à une régression, à partir des spectres de masse des pattes, de la tête et du thorax. 82

5.1 Informations générales sur les cohortes. 93
5.2 Performances des modèles sur les différentes cohortes. 106
5.3 Performances de l'Autoencodeur DAE fc. 107
5.4 Performances de l'Autoencodeur DCAE. 108

Liste des abréviations

1D : One Dimension

1DCNN : One Dimensional Convolution Neural Network

2D : Two Dimensions

2DCNN : Two Dimensions Convolutional Neural Network

2DCNN BiGRU-Hybrid (spectrogram or scalogram) : Two Dimensions Convolutional Neural Network Bidirectional Gated Recurrent Unit Hybrid model for spectrogram or scalogram

ACP : Analyse en Composante Principale

ADN : Acide désoxyribonucléique

AG : Algorithme Génétique (Machine learning model)

API : Application Programming Interface ou « interface de programmation d'application »

ARN : Acide ribonucléique

ANOVA : Analysis of variance ou « analyse de la variance »

BACT-PSL : Laboratoire de Bactériologie de l'Hôpital Universitaire de la Pitié-Salpêtrière à Paris, France

BCH : Hôpital Bichat Claude Bernard, Paris, France

BD : Becton Dickinson (société)

Bruker MBT Compass : Société Bruker Corporation, Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry Biotyper

CC : Coefficient de corrélation

CE-IVD : Conformité Européenne pour Dispositif Médical de Diagnostic In Vitro

CHR : $CHROMagar^{TM}$

CMI : Concentration Minimale Inhibitrice

CNN : Convolutional Neural Network

COH : Columbia agar with horse blood

CORN : Conditional Ordinal Regression for Neural networks

CWT : Continuous Wavelet Transform

CV : Cross Validation

DCAE : Deep Convolutional Auto-encoder

DAE : Deep Auto-encoder

DAE fc : Fully Connected Auto-encoder

DFT : Discrete Fourier Transform

DL : Deep Learning

DT : Decision Trees (Machine learning model)

DDT : Dichloro-Diphenyl-Trichloroethane

DWT : Discrete Wavelet Transform

ESN : Echo State Network

EUCAST : European Committee on Antimicrobial Susceptibility Testing

FFT : Fast Fourier Transform

FN : Faux Négatif

FP : Faux Positif

FT : Fourier Transform

FWHM : Full-Width at Half-Maximum

GRU : Gated Recurrent Unit

HCA : Hierarchical Clustering Analysis

HC : Hierarchical Clustering

HCCA : acide alpha cyano-4-hydroxycinnamique (matrice)

IA : Intelligence Artificielle

IGS (marker) : Intergenic spacer

IRD : Institut de Recherche pour le Développement

KNN : K Nearest Neighbors (Machine learning model)

LDA : Linear Discriminant Analysis (Machine learning model)

LC-MS/MS : Liquid Chromatography coupled to tandem Mass Spectrometry

LightGBM : Light Gradient-Boosting Machine (Machine learning model)

LSV : Log Score Value

LR : Linear Regression (Machine learning model)

MABSc : Complex *Mycobacterium Abscessus*

MAE : Mean Absolute Error (métrique)

MALDI-TOF : Matrix-Assisted Laser Desorption Ionization-Time Of Flight

MALDI-TOF-DL : Matrix-Assisted Laser Desorption Ionization-Time Of Flight Deep Learning model

MI-CLAIM : Minimum information about clinical artificial intelligence modeling

MIRS : Mid-InfRared spectroscopy

ML : Machine Learning

MRR : Marquage Relâchement Recapture

MS : Mass Spectrometry

MSE : Mean Squared Error (métrique)

MSP : Main Spectrum Profile

MTA : Material Transfer Agreement

MYCO-PSL : Laboratoire de Mycologie de l'Hôpital Universitaire de la Pitié-Salpêtrière à Paris, France

NB : Naïve Bayes (Machine learning model)

PCA : Principal Component Analysis (Machine learning model)

PIE : Période d'incubation extrinsèque

PLS-DA : Partial Least-Squares Discriminant Analysis (Machine learning model)

PMF : Peptide Mass Fingerprinting

PQN : Probabilistic Quotient Normalisation

PReLU : Parametric Rectified Linear Unit

PSL : Hôpital Universitaire de la Pitié-Salpêtrière à Paris, France

ReLU : Rectified Linear Unit

REIMS : Rapid Evaporative Ionization Mass Spectrometry

RF : Random Forest (Machine learning model)

RFE : Recursive Feature Elimination

ROC AUC or AUROC : Area Under the Receiver Operating Characteristic (métrique)

RNN : Recurrent Neural Network

RNN-BiGRU : Recurrent Neural Network Bidirectional Gated Recurrent Unit

RNN-GRU : Recurrent Neural Network Gated Recurrent Unit

SAB-CG : Sabouraud Chloramphenicol Agar

SAS : Sociétés par actions simplifiées

SM : Spectrométrie de Masse

SM MALDI-TOF : Spectrométrie de Masse de type désorption-ionisation laser assistée par matrice - en anglais, Matrix-Assisted Laser Desorption Ionization-Time Of Flight

SNIP : Statistics-sensitive Non-linear Iterative Peak-clipping

SNR : Signal-To-NoIse Ratio

SVM : Support Vector Machine (Machine learning model)

STFT : Short-Time Fourier Transform

TCN : Temporal Convolutional Neural Network

TIC : Total-Ion Current

UDWT : Undecimated Discrete Wavelet Transform

UPGMA : Unweighted Paired Group Mean Arithmetic

VPP : Valeur Prédictive Positive

VP : Vrai Positif

VN : Vrai Négatif

RVB : Rouge Vert Bleu

QC : Quick Classifier (Machine learning model)

Résumé

Titre : Exploration des modèles d'apprentissage statistique profonds couplés à la spectrométrie de masse pour améliorer la surveillance épidémiologique des maladies infectieuses

La spectrométrie de masse de type MALDI-TOF (matrix assisted laser desorption and ionisation time of flight) est une méthode de diagnostic en microbiologie rapide et robuste, permettant d'identifier les espèces de micro-organismes grâce à leur empreinte protéique constituée par le spectre de masse. Cependant, les applications clinico-épidémiologiques de cette technologie demeurent limitées par les outils bio-informatiques à disposition. Cette thèse se focalise sur l'application de modèles d'apprentissage statistique profonds aux données de spectrométrie de masse de type MALDI-TOF dans un but de surveillance épidémiologique des maladies infectieuses. Elle inclut la surveillance des épidémies de champignons et de mycobactéries en milieu hospitalier, ainsi que la caractérisation des anophèles vecteurs du paludisme.

Nous avons examiné l'impact des méthodes de préparation des échantillons et de l'analyse informatique des spectres de masse sur l'amélioration de l'apprentissage, afin d'identifier les clones fongiques épidémiques en milieu hospitalier et prévenir leur propagation. Notre étude a montré que le réseau de neurones à convolution (CNN) a un potentiel élevé pour identifier les spectres de clones spécifiques de *Candida parapsilosis*, atteignant une précision de 94 % en optimisant des paramètres essentiels (milieux de culture, temps de croissance, et la machine d'acquisition des spectres). Pour détecter des clones épidémiques d'*Aspergillus flavus* dans des cohortes hospitalières multicentriques, le CNN a également réussi à classer correctement la plupart des isolats, atteignant une précision supérieure à 93 % pour deux des trois appareils utilisés. Nous avons aussi montré qu'en utilisant des modèles d'apprentissage profond optimisés, tels qu'un CNN et un réseau de neurones à convolution temporelle (TCN), nous pouvons prédire l'âge des moustiques avec une précision moyenne de deux jours (meilleure erreur absolue moyenne : 1,74 jours). Cette approche permettrait ainsi de surveiller efficacement la structure de l'âge des populations de moustiques anophèles sauvages et de mieux les cibler par des mesures de contrôle. Enfin, nous avons démontré les performances de diverses architectures de réseaux de neurones et de différentes méthodes de représentation des spectres de masse, en utilisant différentes cohortes couvrant diverses problématiques épidémiologiques telles que la prédiction de l'âge, l'identification d'espèces étroitement apparentées des moustiques anophèles, la distinction entre sous-espèces proches, ainsi que la détection de la résistance chez *Mycobacterium abscessus*. L'étude a montré que parmi les différents modèles évalués, les modèles les plus performants, tels que les TCN et un réseau de neurones récurrents, pouvaient obtenir des résultats notables, atteignant une précision d'identification de 93 % pour les espèces d'anophèles étroitement liées et de 95 % pour les sous-espèces de *Mycobacterium abscessus*. De plus, l'utilisation de CNN et de TCN a permis de détecter les souches résistantes chez *Mycobacterium abscessus* avec une précision dépassant 97 %.

Cette thèse met en lumière l'utilisation de l'apprentissage profond en conjonction avec le MALDI-TOF, une approche jusqu'ici peu explorée. Avec la généralisation des instruments MALDI-TOF et la possibilité de coupler les analyses à des applications en ligne utilisant l'apprentissage profond, cette approche semble prometteuse, ouvrant la voie à d'autres applications épidémiologiques au-delà de la simple identification d'espèce, telles que la détection de clusters épidémiologiques de microorganismes résistants aux médicaments, la surveillance de la transmission des maladies bactériennes et fongiques, et l'évaluation de l'efficacité des interventions ciblées de lutte antivectorielle.

Mots clés :

Spectrométrie de Masse, Apprentissage Profonds, Surveillance, Épidemies, Infections, Prédictions

Abstract

Title: Exploring deep statistical learning models coupled with mass spectrometry to improve epidemiological monitoring of infectious diseases

MALDI-TOF (matrix assisted laser desorption and ionisation time of flight) mass spectrometry is a rapid and robust diagnostic method for microbiology, enabling microorganism species to be identified on the basis of their protein fingerprint in the mass spectrum. However, the clinical and epidemiological applications of this technology remain limited by the bioinformatics tools available. This thesis focuses on the application of deep statistical learning models to MALDI-TOF mass spectrometry data for the purpose of epidemiological surveillance of infectious diseases. This includes the monitoring of fungal and mycobacterial epidemics in hospitals, as well as the characterisation of Anopheles vectors of malaria.

We examined the impact of sample preparation methods and computer analysis of mass spectra on improving learning, in order to identify epidemic fungal clones in hospitals and prevent their spread. Our study showed that the convolution neural network (CNN) has a high potential for identifying the spectra of specific clones of *Candida parapsilosis*, achieving an accuracy of 94 % by optimising key parameters (culture media, growth time, and the spectra acquisition machine). For the detection of epidemic clones of *Aspergillus flavus* in multicentre hospital cohorts, the CNN was also able to correctly classify most isolates, achieving an accuracy of over 93 % for two of the three instruments used. We have also shown that by using optimised deep learning models, such as a CNN and a temporal convolution neural network (TCN), we can predict the age of mosquitoes with an average accuracy of two days (best mean absolute error: 1.74 days). This approach will enable us to effectively monitor the age structure of wild Anopheles mosquito populations and target them more effectively with control measures. Finally, we demonstrated the performance of various neural network architectures and mass spectra representation methods, using different cohorts covering various epidemiological issues such as age prediction, identification of closely related species of *Anopheles* mosquitoes, distinction between closely related subspecies, and detection of resistance in *Mycobacterium abscessus*. The study showed that of the different models evaluated, the best performing models, such as TCNs and a recurrent neural network, were able to achieve notable results, reaching an identification accuracy of 93 % for closely related *Anopheles* species and 95 % for *Mycobacterium abscessus* subspecies. In addition, the use of CNN and TCN enabled the detection of resistant strains in *Mycobacterium abscessus* with an accuracy in excess of 97 %.

This thesis highlights the use of deep learning in conjunction with MALDI-TOF, a hitherto little explored approach. With the widespread availability of MALDI-TOF instruments and the possibility of coupling analyses to online applications using deep learning, this approach looks promising, paving the way for other epidemiological applications beyond simple species identification, such as the detection of epidemiological clusters of drug-resistant microorganisms, monitoring the transmission of bacterial and fungal diseases, and evaluating the effectiveness of targeted vector control interventions.

Keywords:

Mass Spectrometry, Deep Learning, Monitoring, Outbreak, Infection, Prediction

Chapitre 1

Introduction

1.1 Contexte

Les maladies infectieuses ont un impact significatif sur la santé publique et l'économie mondiale depuis des siècles. Encore maintenant, elles constituent l'une des principales causes de décès et de handicap, représentant un défi croissant pour la sécurité sanitaire et le progrès humain. Dans ce contexte, les méthodes de diagnostic microbiologique comme (i) l'identification d'espèces (ii) la détection de clones responsables d'épidémies et (iii) la caractéristisation d'espèces vectrices représentent un défi majeur (Nii-Trebi et al., 2017). Ainsi, de nouvelles technologies appliquées aux méthodes diagnostiques en microbiologie ont émergés ces dernières années et parmi elles, la spectrométrie de masse (SM) permet de déterminer avec une précision et sensibilité extrême un paramètre intrinsèque des molécules : leur masse (Glish et Vachet, 2003). La SM est une méthode d'analyse qui consiste à transformer des molécules dans leur état naturel en ions à l'état gazeux puis à les séparer en fonction de leur rapport masse/charge noté m/z. L'application de cette technologie à l'étude de composés biologiques est relativement récente puisqu'elle date de la fin du XXème siècle.

Depuis trois décennies, la SM s'est érigée en une méthode fréquemment employée dans la recherche et la caractérisation des composés biologiques, notamment pour des applications telles que la détermination de la structure des protéines et la caractérisation de certaines modifications chimiques, telles que les liaisons disulfures, les glycosylations et les phosphorylations (Redeker et al., 1998, Parker et al., 2010). Une autre utilisation de la SM est l'analyse de mélanges de peptides ou de protéines plus ou moins complexes tel que l'étude d'un protéome spécifique (Aebersold et Mann, 2003, Parker et al., 2010). La spectrométrie de masse est une méthode précise d'analyse quantitative capable de détecter des quantités extrêmement faibles de composés chimiques simples, se rapprochant de l'ordre du femtomol, soit 10^{-15} moles. La gamme des spectromètres de masse regroupe un panel assez large d'appareils. Ils sont désormais plus accessibles, en termes de moyens financiers, d'entretien et d'utilisation principalement grâce à l'émergence d'instruments compacts (voir le Microflex de Bruker, Bilecen et al., 2015).

Il y a une dizaine d'années, la spectrométrie de masse de type (SM) de type MALDI-TOF (désorption-ionisation laser assistée par matrice - en anglais, matrix-assisted laser desorption ionization-time of flight) s'est implantée comme une méthode d'identification des micro-organismes dans les laboratoires de microbiologie. Cette technologie est devenue un outil incontournable pour le diagnostic microbiologique.

L'identification des micro-organismes et joue un rôle essentiel dans le diagnostic des infections et dans l'adaptation des traitements anti-infectieux en fonction des résistances naturelles et acquises des microorganismes (Hou et al., 2019, Zhu et Girault, 2023). Avant les années 2000, les techniques utilisées pour l'identification des bactéries reposaient sur les caractéristiques biochimiques des souches, à l'aide de galeries API ou d'automates d'identification biochimique. L'identification d'un micro-organisme nécessitait alors environ 24h. À partir des années 2010, la SM de type MALDI-TOF s'est implantée dans les laboratoires de microbiologie et a permis d'obtenir une identification plus rapide en seulement quelques minutes. Cette technologie est également déployée dans les laboratoires vétérinaires et certains laboratoires agroalimentaires, où elle est utilisée notamment pour détecter la présence de pathogènes alimentaires (Hou et al., 2019 ; Han et al., 2021).

Le MALDI-TOF est une méthode qui génère un profil protéique, agissant comme une empreinte digitale spécifique pour chaque micro-organisme. Elle permet une identification rapide d'un large éventail de micro-organismes

en comparant leurs profils avec une base de données de référence. Le MALDI-TOF s'est avéré plus précis que les méthodes d'identification traditionnelles basées sur des critères morphologiques (Patel, 2019 ; Gautier et al., 2014). Cependant, l'exploitation complète de ces données complexes nécessite des approches analytiques avancées, ce qui n'est pas le cas des systèmes commerciaux utilisés en routine (Torres-Sangiao et al., 2021). En effet, les données générées par la spectrométrie de masse MALDI-TOF peuvent être volumineuses, avec des dizaines à des centaines de spectres, chacun composé de milliers de mesures de masse et d'intensité. Ces données brutes nécessitent un traitement préliminaire pour éliminer le bruit et identifier les signaux pertinents (Armananzas et al., 2011). Un prétraitement inadéquat peut biaiser les données et entraver la découverte d'informations biologiques significatives (Coombes et al., 2007). Cela est dû à la présence de signaux réels des protéines ainsi que de divers types de bruit, tels que des facteurs chimiques ou électroniques. L'analyse et le traitement appropriés de ces données importantes peuvent fournir des informations au-delà de la simple identification des espèces.

Avec l'avènement de l'intelligence artificielle (IA), son intégration dans les institutions de santé promet de révolutionner nos connaissances et nos capacités. Parmi les disciplines de l'IA, l'apprentissage automatique permet aux ordinateurs d'apprendre des modèles mathématiques et de proposer des prédictions ou des décisions, à partir des données disponibles. Malgré ses débuts peu répandus, l'apprentissage automatique est de plus en plus utilisé en microbiologie diagnostique, facilitant le traitement de vastes ensembles de données complexes. Des chercheurs ont développé des algorithmes d'apprentissage automatique pour interpréter les cultures bactériennes, analyser les images en vue de détecter les micro-organismes et prédire les sensibilités aux antimicrobiens (Theodosiou et Read, 2023). Ainsi, l'apprentissage automatique pourrait contribuer à répondre à la demande croissante de résultats plus rapides et plus précis.

L'apprentissage statistique profond (ou apprentissage automatique profond, Deep Learning) est une branche de l'apprentissage automatique qui utilise des réseaux de neurones artificiels, initialement inspirés des neurones biologiques, comportant plusieurs couches pour traiter les données. Il peut adopter trois principales approches : l'apprentissage supervisé (où le modèle apprend avec des données étiquetées), l'apprentissage non supervisé (où le modèle explore des données non étiquetées pour détecter des motifs) et l'apprentissage par renforcement (où un agent apprend à optimiser son comportement en interagissant avec son environnement pour maximiser les récompenses à long terme). L'apprentissage profond excelle dans la résolution de tâches complexes impliquant des données de grande dimension.

Dans ce contexte, l'utilisation conjointe des avancées technologiques en SM de type MALDI-TOF et des modèles d'apprentissage statistique profonds ouvre de nouvelles possibilités pour améliorer la surveillance épidémiologique à partir de l'analyse des spectres. Les réseaux de neurones, sont particulièrement prometteurs pour extraire des informations pertinentes à partir des données de SM. Leur capacité à saisir des relations complexes et non linéaires peut considérablement contribuer à l'analyse et à l'interprétation des spectres protéiques, ouvrant ainsi de nouvelles perspectives pour la surveillance épidémiologique des maladies infectieuses.

1.2 Objectif

Cette thèse vise à explorer l'utilisation de modèles d'apprentissage statistique profonds en conjonction avec la SM de type MALDI-TOF pour relever les défis actuels de la surveillance épidémiologique. Nous proposons une approche interdisciplinaire qui intègre des concepts de biologie, de bioinformatique et d'apprentissage statistique profond pour développer des méthodes novatrices de détection précoce, de suivi et de caractérisation des agents pathogènes, ainsi que pour identifier des indicateurs prédictifs des épidémies. Tout d'abord, nous explorons la possibilité d'utiliser des techniques d'apprentissage statistique profond pour différencier efficacement, au sein d'un ensemble de spectres de masse, des organismes partageant une proximité taxonomique, tels que des espèces, des sous-espèces, des clones particuliers et des caractéristiques d'espèces vectrices, notamment pendant une épidémie. Également, nous cherchons à connaître la meilleure approche d'apprentissage automatique pour améliorer la capacité de la SM à individualiser des caractéristiques spécifiques au sein d'une espèce donnée. En outre, notre objectif est de voir dans quelle mesure l'application de l'apprentissage automatique profond couplé à la SM MALDI-TOF peut être utile à la surveillance épidémiologique des infections hospitalières.
Cette thèse vise à renforcer notre arsenal d'outils pour lutter contre les maladies infectieuses en termes d'identification, de détection de la résistance aux antimicrobiens et de clones responsables d'épidémies.

1.3 Plan du mémoire

Tout d'abord, dans un premier chapitre, nous présentons un état de l'art sur la SM de type MALDI-TOF, mettant en lumière la complexité de ses spectres en raison de leur composition et de leurs caractéristiques particulières. Nous abordons également les différentes méthodes de traitement préliminaire des spectres, essentielles pour en exploiter pleinement les informations. Ensuite, nous explorons les diverses applications de la SM MALDI-TOF dans le domaine de l'apprentissage automatique, y compris l'apprentissage profond, bien que cette dernière approche demeure relativement nouvelle et moins développée en raison de son émergence récente.

Dans le chapitre suivant, dans un contexte d'identification de clones fongiques en milieu hospitalier et pour suivre leur propagation, nous examinons l'impact des méthodes de préparation des échantillons et de l'analyse informatique des spectres de masse sur l'apprentissage d'un réseau de neurone, afin de connaître les mesures adéquates permettant à un réseau de neurones de détecter ensembles clonaux dans une population de champignons d'une même espèce.

Dans le chapitre qui suit, nous montrons qu'en utilisant des modèles d'apprentissage profond optimisés, nous pouvons prédire l'âge des moustiques anophèles avec une précision moyenne de deux jours. Cette approche permettrait ainsi de surveiller efficacement la structure de l'âge des populations de moustiques anophèles sauvages et de mieux évaluer l'efficacité des mesures de contrôle.

Enfin, dans un dernier chapitre méthodologique, nous explorons divers modèles de réseaux de neurones et différents types de représentation des spectres de masse pour évaluer leur applicabilité à la SM MALDI-TOF. Nous mettons en évidence leurs avantages et limites en les appliquant à diverses cohortes (jeux de données), notamment pour prédire l'âge des moustiques anophèles, identifier des espèces étroitement liées, distinguer des sous-espèces associées à la résistance au antibiotiques chez *Mycobacterium abscessus*.

Nous espérons que ces études contribueront à ouvrir de nouvelles perspectives dans la surveillance épidémiologique des maladies infectieuses.

Chapitre 2

État de l'art

2.1 La spectrométrie de masse par MALDI-TOF

2.1.1 Principes et appareillage

La SM MALDI-TOF est une méthode qui peut être appliqué à l'étude d'une large gamme de molécules d'intérêt biologique. Cette méthode permet d'analyser des protéines de gamme de masse variable de 1 à 300 000 Da, avec une bonne précision de mesure de masse, une préparation, une mise en œuvre relativement simples, une grande rapidité d'analyse et une relative tolérance vis-à-vis des tampons, des sels et de nombreux détergents. C'est une méthode de référence en SM pour la caractérisation et l'identification microbienne. Grâce à cette technologie, des empreintes spectrales de masses caractéristiques sont générées, fournissant des signatures uniques pour chaque micro-organisme. Une comparaison du spectre à identifier avec une banque de spectres de référence, permet une identification précise au niveau du genre et de l'espèce.

La SM MALDI-TOF (Figure 2.1) a été largement utilisée pour caractériser une grande variété de micro-organismes, incluant des bactéries et des champignons (Giebel et al., 2010 ; Marinach.et al, 2009 ; Desoubeaux et al., 2010 ; Singhal et al., 2015). Sa capacité à identifier rapidement les micro-organismes ouvre la voie à de nombreuses applications potentielles, notamment dans le domaine du diagnostic médical, de la surveillance de l'environnement et du contrôle de la qualité des aliments (Han et al, 2021). Le MALDI-TOF est particulièrement adapté pour une identification microbienne rapide et à haut débit.

L'échantillon de biomolécules destiné à l'analyse par MALDI-TOF est préparé par mélange avec une solution d'un composé organique qui absorbe l'énergie, appelé matrice. Ce mélange est déposé sur une surface métallique de la cible et après évaporation, on observe une co-cristalisation des protéines avec la matrice. La matrice permet de protéger les protéines lors de la phase d'ionisation.

FIGURE 2.1 – **Appareil MALDI-TOF** Source : A. Godmer

Une fois la cible introduite dans le spectromètre de masse, l'échantillon est soumis à l'action d'un rayon laser UV. Le rôle de la matrice est d'absorber l'énergie du laser, provoquant ainsi la vaporisation de l'échantillon et la formation d'ions de masses différentes. Ces ions, généralement de charge +1, sont ensuite accélérés sous l'action d'un puissant champ électrique. Les ions rentrent alors dans le tube de vol, libre de champ et sont séparés en fonction de leur rapport masse/charge (m/z). Un détecteur est placé à l'extrémité du tube de vol qui permet la détection des ions. Lors de l'analyse MALDI-TOF, le rapport m/z d'un ion est mesuré en déterminant le temps nécessaire pour qu'il parcoure la longueur du tube de vol. Les ions de plus petite taille atteignent le détecteur en premier. Une fois mesuré, le temps de vol (time of flight) est utilisé pour calculer la masse de chaque particule (Figure 2.2). La somme des ions analysés, forme un spectre caractéristique de l'échantillon.

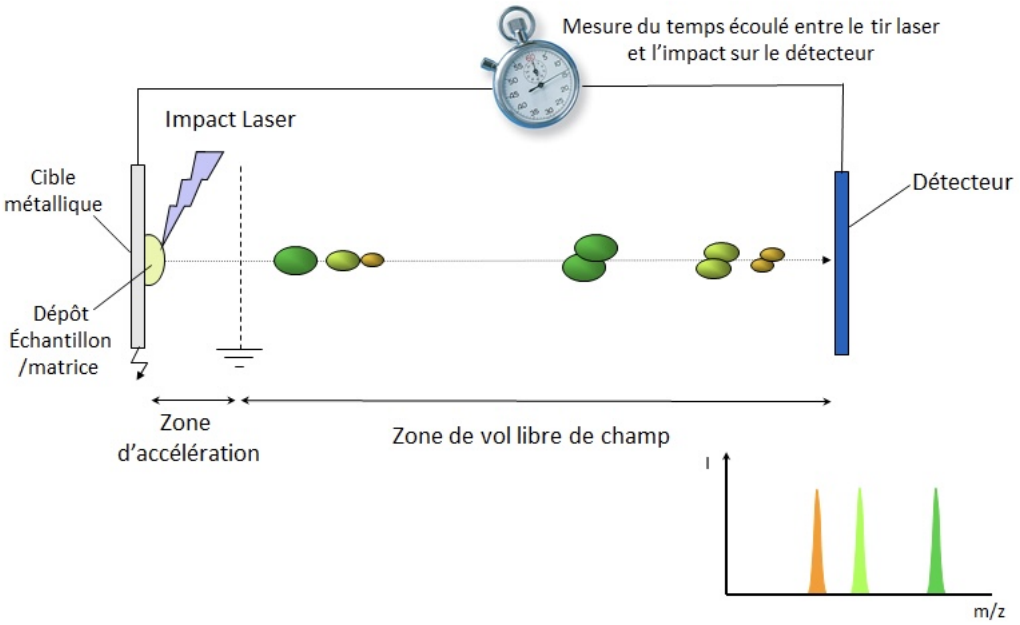


FIGURE 2.2 – **Principe de la spectrométrie de masse MALDI-TOF - désorption et ionisation assistée par une matrice avec détection en temps de vol.** Source : J-Y Brossas

Le temps que met un ion à passer dans le tube dépend du rapport entre sa charge et sa masse, c'est-à-dire, son rapport masse/charge, m/z. Ainsi, ce que le spectromètre observe réellement est un temps, le temps de vol d'un ion lorsqu'il voyage de l'entrée du tube de vol au détecteur. Cependant, ce temps est généralement converti par le logiciel du spectromètre en un rapport m/z, et c'est ce rapport qui est présenté à l'expérimentateur. Les ions mono-chargés sont en général majoritaires dans les spectres MALDI ce qui signifie que la masse m et le rapport m/z sont confondu.

En se basant sur les informations obtenues grâce au temps de vol (TOF), un spectre est généré pour les analytes présents dans l'échantillon. Cette approche offre une méthode précise et fiable pour l'identification et la caractérisation des micro-organismes à l'aide de la SM MALDI-TOF. Dans ces spectres, l'axe des abscisses correspond au rapport masse sur charge (m/z), tandis que l'axe des ordonnées représente l'intensité relative du signal.

2.1.2 Comment identifier des micro-organismes à l'aide des banques de données ?

L'identification des microorganismes par MALDI-TOF repose sur le fait que chaque micro-organisme est capable de produire un spectre de masse caractéristique d'un genre ou d'une espèce (Figure 2.3). Ainsi, en comparant un spectre de masse provenant d'un micro-organisme avec des spectres de masses préalablement caractérisés et regroupés dans une banque de données, il est possible d'obtenir une identification. Cela passe par la comparaison du profil de masse protéique (PMF : Peptide Mass Fingerprinting) correspondant au spectre de masse de l'organisme inconnu, avec ceux contenus dans une base de données.
Pour des analyses plus fines qui peuvent être destinées à d'autres objectifs que l'identification d'espèce (caractérisation de sous-espèces ou de clones par exemples) on procède par profilage protéique : on recherche des pics informatifs (appelés pics discriminants) afin de faire correspondre les masses des biomarqueurs de l'organisme inconnu avec celles de la base de données du protéome.

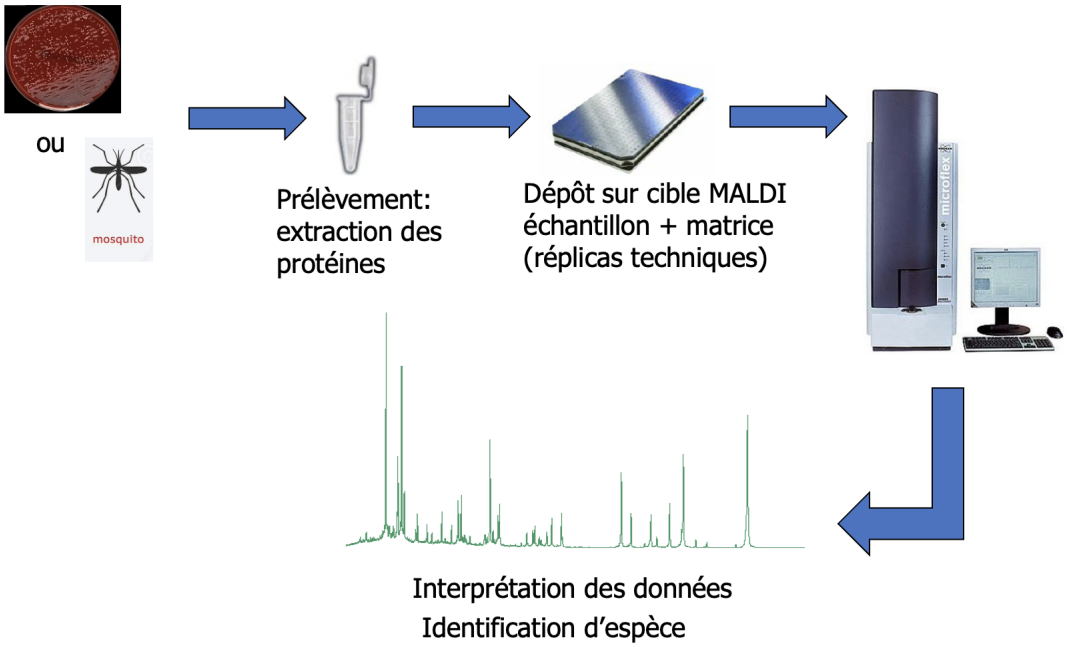


FIGURE 2.3 – **Création de banques d'identification avec la spectrométrie de masse.** Source : C. Nabet

Pour identifier un micro-organisme jusqu'au niveau de l'espèce, on utilise une gamme de masse typique allant de 2000 à 20 000 Da, principalement représentative des protéines ribosomales et d'autres protéines courantes (Angeletti, 2017). Les protéines ribosomales, qui constituent environ 60 à 70 % du poids sec d'une cellule microbienne, sont très abondantes dans cette gamme de masse et permettent d'établir le schéma caractéristique d'un micro-organisme particulier. Les autres protéines courantes sont des protéines structurelles, des régulateurs du stockage du carbone, des protéines de choc thermique, des protéines de liaison à l'ADN et des chaperonnes de molécules d'ARN. Ces protéines, souvent qualifiées de « ménage », sont omniprésentes et abondantes dans toutes les cellules, ce qui maintient la similarité des profils spectraux entre les organismes apparentés sur le plan phylogénétique ou taxonomique.

En comparant le schéma PMF inconnu avec les PMF connus dans une vaste base de données, on peut identifier le micro-organisme jusqu'au genre et souvent jusqu'à l'espèce. Cette approche est largement utilisée pour l'identification microbienne en raison de sa simplicité et de la disponibilité de nombreuses bibliothèques commerciales de PMF d'organismes, notamment par Bruker et Biomérieux (Cassagne et al., 2016).

Dans l'approche par pics discriminants, on suppose que la majorité des pics enregistrés correspondent aux protéines ribosomales, ce qui est source de stabilité et robustesse dans l'identification d'une espèce donnée. Le profilage protéique permet ainsi de caractériser les pics de masse appartenant à une même catégorie (genre, espèce, souche, état physiologique, etc.). Les masses protéiques enregistrées, exprimées en m/z (Da), peuvent servir de biomarqueurs potentiellement discriminants entre différentes catégories de spectres.

C'est par la combinaison de ces approches que l'identification des microbes par MALDI-TOF se révèle être un outil puissant et pratique dans les laboratoires de diagnostic microbien.

2.1.3 Les caractéristiques du spectre

Les pics dans un tracé de l'intensité en fonction du temps représentent les protéines ou les peptides présents dans l'échantillon (Figure 2.4). Un ensemble de données typique provenant d'une application clinique de la SM contient des dizaines ou des centaines de spectres ; chaque spectre contient plusieurs milliers de mesures d'intensité représentant un nombre inconnu de pics de protéines (Figure 2.5). Toute tentative d'interprétation de ce volume de données nécessite un traitement préliminaire afin d'identifier l'emplacement des pics et de quantifier leur taille avec précision (voir la section "Le traitement des spectres").

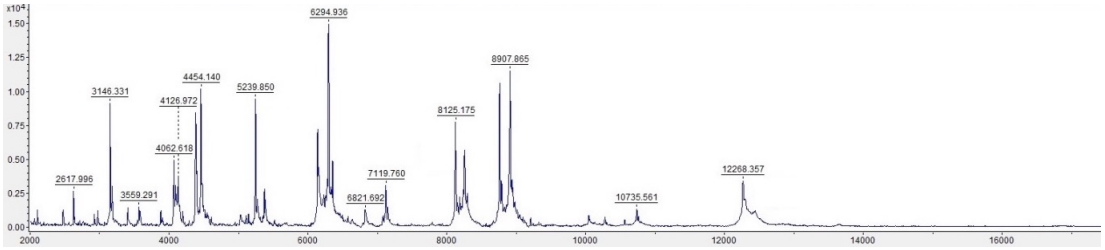


FIGURE 2.4 – **Exemple d'un spectre de moustique à 28 jours de l'espèce *An. stephensi* de la banque âge obtenue par MALDI-TOF.** Le spectre a été acquis avec le paramétrage par défaut du logiciel Microflex LT (Bruker France SAS) et est visualisé en utilisant l'option AutoXecute du logiciel FlexControl v3.4 software (Bruker France SAS). Le spectre a été acquis pour une fenêtre de masses variant de 2 à 20 kDa. Source : N. Shahmirian et C. Nabet

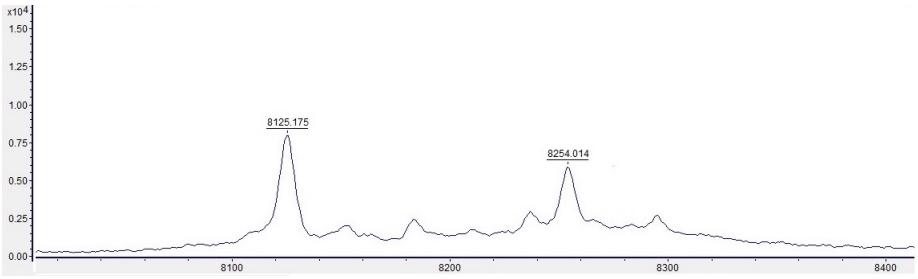


FIGURE 2.5 – **Visualisation des pics 8125 et 8257 du spectre précédent obtenue par MALDI-TOF.** Source : N. Shahmirian et C. Nabet

Dans un spectre de masse MALDI-TOF, la forme, la taille et la largeur des pics peuvent varier en fonction des conditions expérimentales spécifiques et des réglages de l'instrument. Voici un aperçu général de ces caractéristiques :

- Forme du pic : La forme du pic dans un spectre de masse MALDI-TOF est généralement gaussienne. Cependant, dans des situations réelles, d'autres facteurs peuvent influencer la forme des pics, entraînant des écarts par rapport à la distribution gaussienne idéale.
- Taille du pic : La taille du pic dans un spectre de masse MALDI-TOF correspond à l'intensité ou à l'abondance des ions détectés à un rapport masse/charge (m/z) spécifique.
- Largeur de pic : La largeur de pic fait référence à la largeur totale à mi-hauteur (FWHM : Full-Width at Half-Maximum) des pics dans le spectre de masse. Il s'agit d'une mesure de la résolution qui représente le plus petit écart de masse Δm qui peut être mesuré à une masse donnée. La largeur des pics sont donc la résolution est influencées par divers facteurs, notamment la distribution initiale des ions, la dispersion le long du tube de vol, la nature du détecteur et d'autres considérations instrumentales. En général, des pics plus larges peuvent résulter d'une ionisation moins contrôlée ou d'une dispersion accrue pendant le vol des ions (Figure 2.6).

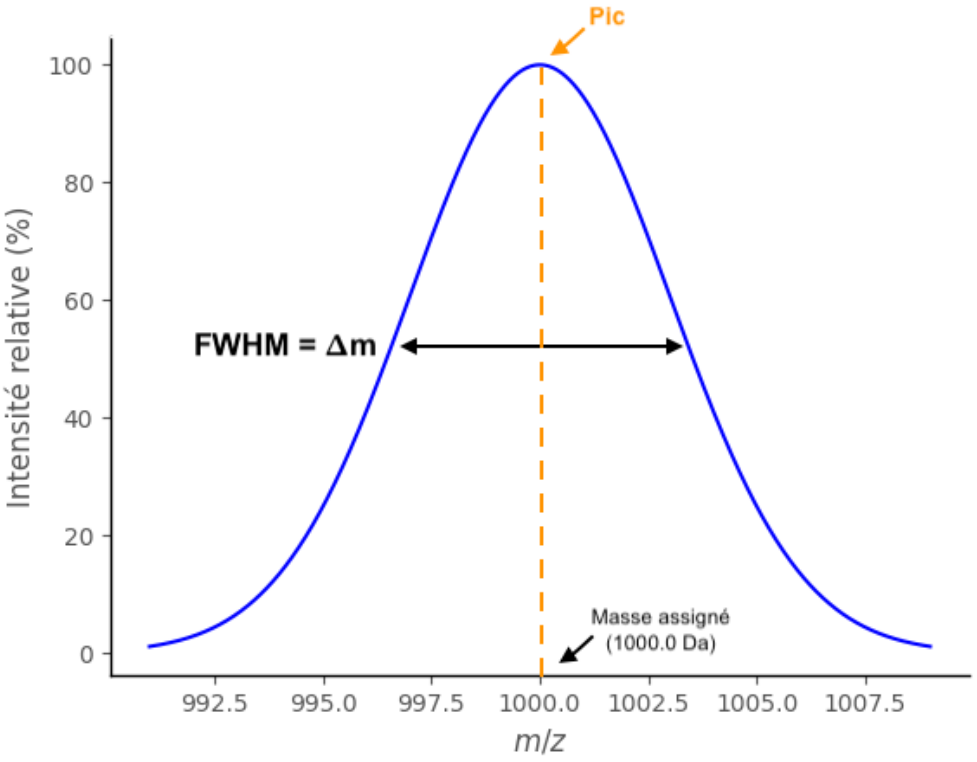


FIGURE 2.6 – **Forme et résolution d'un pic.** Source : Photothèque personnelle

Il est essentiel d'optimiser les paramètres expérimentaux et les réglages instrumentaux pour obtenir des pics bien résolus (c'est-à-dire des pics distincts et nets qui apparaissent dans le spectre de masse obtenu à partir de l'analyse d'un échantillon) dans un spectre de masse MALDI-TOF. Des pics avec une bonne résolution augmentent la précision de la masse, ce qui permet de sélectionner un plus grand nombre de pics pour l'analyse. Cependant, la forme, la taille et la largeur exactes des pics peuvent varier en fonction de la complexité de l'échantillon, de la matrice utilisée, des capacités de l'instrument et d'autres facteurs expérimentaux.

Les spectromètres de masse sont sujets à diverses sources de bruit. Les données brutes typiques présentent toujours des caractéristiques complexes, car les signaux des protéines, caractérisés par de "vrais" pics dans le spectre de masse, peuvent être contaminés par plusieurs processus chimiques et/ou physiques résultant de la procédure de mesure. Les sources de bruit comprennent : le bruit mécanique causé par les réglages de l'instrument, le bruit électronique provenant de la fluctuation d'un signal électronique et de la distance de déplacement du signal, le bruit chimique influencé par la préparation et la contamination de l'échantillon, la température dans le tube de vol et les erreurs de lecture du signal par le logiciel. Ils entraînent la présence de signaux parasites et une ligne de base complexe dans les spectres de masse (Malyarenko et al., 2005). Ces interférences compliquent l'identification des pics d'intérêt de faible intensité. La complexité des spectres de masse découle de la variabilité inter- et intra-échantillon, des processus chimiques et physiques de la mesure, ainsi que de la préparation et de la dégradation des échantillons.

Pour résoudre ce problème, un pré-traitement des spectres est nécessaire pour réduire le bruit et les biais systématiques avant une analyse plus approfondie des données MALDI-TOF (Coombes et al., 2007).

2.2 Le traitement des spectres

Les données de laboratoire de spectromètre de masse de type MALDI-TOF sont complexes, composées de dizaines voire de centaines de spectres. Chaque spectre contient des dizaines de milliers de mesures d'intensité représentant un nombre inconnu de pics protéiques. Toute tentative d'interprétation de ce volume de données nécessite un traitement préliminaire afin de nettoyer les données, de détecter les vrais signaux dans les spectres bruyants (Armananzas et al., 2011), d'identifier l'emplacement des pics et de quantifier leur taille avec préci-

sion. Le pré-traitement des données est une étape cruciale qui transforme les données brutes en données d'entrée appropriées pour une analyse ultérieure, telle que l'analyse et l'identification par apprentissage automatique ou la découverte de biomarqueurs. Les données brutes contiennent des signaux provenant des peptides/protéines réels et des signaux dérivés de plusieurs formes de bruit. Ainsi, des méthodes de pré-traitement inadéquates ou incorrectes peuvent entraîner un ensemble de données biaisées et empêcher de parvenir à des conclusions biologiques significatives (Coombes et al., 2007 ; Sorace et Zhan, 2003 ; Baggerly et al., 2004).
Le traitement de bas niveau des spectres implique un certain nombre d'étapes compliquées qui interagissent de manière complexe.

Les techniques de pré-traitement idéales en protéomique consistent à réduire au mieux tous les types d'incertitude dans les données brutes de SM de type MALDI-TOF afin de permettre la reproductibilité des données et la comparaison des spectres. Les principaux objectifs du pré-traitement sont (Eidhammer et al., 2008) d'éliminer le bruit sans rejeter aucune des valeurs m/z d'intérêt et de déterminer les valeurs m/z et d'intensité avec la meilleure précision. Le "nettoyage" des données brutes de spectromètre de masse passe par plusieurs étapes de pré-traitement faisant appel à différentes techniques appliquées à ces étapes.

Les techniques de pré-traitement les plus courantes sont :

- le lissage,
- la correction de la ligne de base,
- la normalisation,
- la détection des pics,
- et l'alignement des pics.

L'ordre séquentiel dans lequel ces étapes sont exécutées est crucial pour le déroulement ultérieur du processus. Il est essentiel de noter que dans la littérature spécialisée, les différentes méthodes de pré-traitement ne s'alignent pas nécessairement sur une séquence identique des étapes susmentionnées et peuvent omettre certaines d'entre elles.

2.2.1 Lissage

En général, les spectres sont irréguliers, ce qui rend difficile la détection des vrais pics parmi le bruit. C'est pourquoi un algorithme de lissage est généralement appliqué afin d'adoucir les spectres. Le lissage du signal est la première étape du pré-traitement des données et vise à supprimer le bruit des instruments dans les données et les variations stochastiques dans le signal du spectre.

Plusieurs méthodes ont été proposées avec succès dans la littérature pour corriger ce problème. Les techniques les plus simples sont basées sur l'utilisation d'une fenêtre coulissante, où l'intensité de chaque valeur m/z est ajustée en fonction de l'intensité des valeurs m/z voisines.

Pour un spectre d'entrée, nous le représentons par $(m/z, x)$, où le premier élément est le vecteur m/z et le second est le vecteur d'intensité (de même longueur). Afin de faciliter les descriptions lors du traitement des signaux, nous utilisons $x(t)$ pour désigner la forme continue du vecteur d'intensité et $x[n]$ pour désigner la forme discrète du vecteur d'intensité. Ici, t et n servent de variables d'indexation. Il est important de noter que le spectre d'entrée est toujours discret. Nous utilisons la forme continue par souci de cohérence avec la description originale. Dans les applications réelles, nous échantillonnons généralement le filtre continu pour obtenir sa forme discrète. Les valeurs de m/z peuvent être obtenues à partir du vecteur m/z en utilisant la variable d'indexation correspondante.

Un spectre après lissage peut être exprimé sous la forme :

$$y[n] = x[n] * w[n]$$

pour le cas discret, et :

$$y(t) = x(t) * w(t)$$

pour le cas continu, où $*$ désigne l'opération de convolution. Dans les équations ci-dessus, $w[n]$ et $w(t)$ sont respectivement un vecteur de poids et une fonction de poids. L'utilisation de $w[n]$ et $w(t)$ différents conduira à

des filtres différents.

Voici les techniques de lissages les plus utilisés (Yang et al., 2009 ; Coombes et al., 2005 ; Coombes et al., 2007) :

- Filtre à moyenne mobile (Moving average filter)
- Filtre de Savitzky-Golay (Savitzky-Golay filter)
- Filtre gaussien (Gaussian filter)
- Fenêtre de Kaiser (Kaiser window)
- Transformée en ondelettes continue (Continuous Wavelet Transform) [1]
- Transformée en ondelettes discrète (Discrete Wavelet Transform)
- Transformée en ondelettes discrète non décimée (Undecimated Discrete Wavelet Transform)

Beaucoup de librairies sont disponibles en ligne et proposent une (ou plusieurs) technique de lissage des spectres. La plupart d'entre elles utilisent les algorithmes présentés précédemment. MassUp propose deux méthodes de lissage : la fenêtre à moyenne mobile et la fenêtre de Savitzky-Golay, toutes deux issues de la bibliothèque MALDIquant (Gibb et Strimmer, 2012). La transformée en ondelettes discrète non décimée (UDWT) disponible dans la librairie Cromwell.

2.2.2 Soustraction de la ligne de base

La ligne de base est une forme spécifique de bruit principalement due à des perturbations chimiques, définie comme un décalage des intensités des pics qui dépend souvent de la valeur m/z, de sorte qu'elle est la plus élevée à de faibles valeurs m/z et qu'elle présente une décroissance exponentielle vers des masses plus élevées (Eidhammer et al., 2008). Pour la SM MALDI-TOF, la ligne de base est un biais décroissant de façon monoïque résultant des amas de matrice formés pendant l'ionisation (Shin et al., 2007 ; Sun et Markey, 2011).

La correction de la ligne de base est un problème difficile qui peut également introduire des artefacts (Antoniadis et al., 2010).
Il s'agit d'un processus en deux étapes :

1. l'estimation de la ligne de base, qui peut être omise. Elle consiste à identifier la base à soustraire.
2. la soustraction de la ligne de base du signal.

La soustraction de la ligne de base, consiste à éliminer le "lit" estimé sur lequel repose le profil spectral, composé d'un signal non biologique, par exemple le bruit chimique lié à la mise en suspension de la matrice ionisée. Elle élimine les artefacts systématiques, généralement attribués à des grappes de molécules de matrice ionisées qui frappent le détecteur pendant les premières parties de l'expérience, ou à une surcharge du détecteur.

Parmi les méthodes de correction de la ligne de base couramment utilisées, on peut citer :

- Minimum monotone (Monotone minimum)
- Interpolation linéaire (Linear interpolation)
- Loess
- Transformée en ondelettes continue (CWT, Continuous Wavelet Transform)
- Moyenne mobile des minima (Moving average of minima)
- Détection Sensible et Non Linéaire avec Élimination d'Amplitudes de Crêtes en Itératif (SNIP, Sensitive Nonlinear Iterative Peak-clipping) (Gibb et Strimmer., 2012)

Ces méthodes sont toutes disponibles sous forme de logiciel gratuit dans différents progiciels tels que Cromwell (Matlab) (Coombes et al., 2005), PROcess (R) de Bioconductor, MALDIquant (R) (Gibb et Strimmer, 2012) ou SpecAlign (Java) (Wong et al., 2005). Le logiciel Mass-Up (López-Fernández et al., 2015) permet à l'utilisateur d'utiliser toutes les méthodes de correction de la ligne de base fournies par MALDIquant (c'est-à-dire Top Hat, SNIP, Convex Hull et Median).

Le calcul de la correction de la ligne de base et la recherche des pics sont deux problèmes fortement liés, il est donc naturel de retrouver des méthodes qui effectuent ces deux opérations conjointement comme avec la CWT.

1. À noter que la CWT est un cas spécial utilisé pour le lissage et/ou la réduction de la ligne de base.

2.2.3 Normalisation

La normalisation vise à rendre les signaux proportionnels entre eux, corrigeant ainsi la variabilité de l'instrument tout en améliorant l'efficacité de l'ionisation de l'échantillon, ce qui a un impact sur le nombre d'ions peptidiques détectés. Une contrainte majeure du MALDI-TOF est que l'intensité des valeurs m/z est relative et peut varier entre les spots d'un même échantillon. C'est pourquoi on a généralement recours à la normalisation, qui permet de comparer les intensités de différents spectres. Les méthodes de normalisation les plus courantes sont le courant ionique total (TIC), la normalisation par quotient probabiliste (PQN), le score Z, la normalisation linéaire, la moyenne ou la médiane (López-Fernández et al., 2015). La normalisation corrige les différences systématiques dans la quantité totale de protéines désorbées et ionisées à partir de la plaque d'échantillon.

Cependant, cette étape de pré-traitement est souvent négligée, car elle est perçue comme non essentielle et ne conduit pas à une amélioration significative de la qualité des spectres pour certaines tâches spécifiques réalisées après cette phase de pré-traitement.

2.2.4 Détection des pics

Dans le domaine de la SM, une distinction claire est établie entre l'identification et la détection des pics. Les pics enregistrés par un SM demeurent essentiellement anonymes. L'unique information dont nous disposons à leur sujet concerne leur masse, laquelle demeure insuffisante pour une caractérisation exhaustive de la protéine ou du peptide à l'origine de chaque pic.

Le terme d'identification des pics fait référence au processus de détermination de l'espèce exacte de la molécule de protéine qui a provoqué la détection d'un pic. Ce processus implique généralement des expériences supplémentaires (souvent en transférant les molécules d'une masse cible dans un autre instrument où elles sont physiquement fragmentées le long des limites des acides aminés et envoyées dans un second spectromètre de masse pour déterminer la taille des fragments) et des recherches dans des bases de données pour comparer les résultats avec les schémas de fragmentation de protéines connues.

La détection des pics peut être définie comme le processus de sélection des valeurs d'intérêt, dit "vrais pics", (c'est-à-dire liés aux peptides/protéines) à partir d'un spectre donné. Elle est généralement appliquée après la correction de la ligne de base et le lissage. Cette étape consiste à détecter le signal des pics sous la forme de paires de masse et d'intensité des peptides. La détection des pics dans les spectres de masse est l'étape initiale de l'analyse des données MALDI-TOF. Il convient de veiller à ce que cette étape soit aussi précise que possible, car les erreurs qui s'y produisent affectent fortement les performances des étapes suivantes et peuvent éventuellement conduire à des conclusions erronées. Même si les pics d'intérêt correspondent aux biomolécules, ils apparaissent comme des maxima locaux dans un spectre. La détection de ces pics est un défi en raison de l'importance du bruit de fond.

Pour faciliter la détection de pics, plusieurs algorithmes ont été proposés, ces derniers utilisent un ou plusieurs des critères suivants pour identifier les vrais pics (Yang et al., 2009) :

- SNR (Signal to Noise Ratio) : le SNR est une mesure du signal par rapport au bruit de fond. Les pics sont sélectionnés si leur intensité est supérieure à un seuil donné pour exclure les valeurs de m/z de faible intensité. Ce seuil peut être défini de manière absolue, par exemple en utilisant une intensité minimale, ou de manière relative, comme le rapport signal/bruit.
- Seuil de détection/intensité (Detection/Intensity threshold) : Ce critère est utilisé pour éliminer les petits pics dans les régions plates. L'utilisation du rapport signal sur bruit (SNR) seul dans ces régions peut sélectionner des points bruyants comme pics, puisqu'ils peuvent avoir un SNR élevé.
- Pentes des pics (Slopes of peaks) : Dans le cadre de ce critère, la forme des pics est utilisée pour filtrer les fausses occurrences. Tout pic potentiel est éliminé si les pentes gauche et droite sont inférieures à un seuil préétabli. La limite est définie comme la moitié du niveau de bruit local (Coombes et al., 2003).
- Maximum local (Local maximum) : Selon ce critère, un pic est sélectionné s'il s'agit d'un maximum local de N points voisins.
- Rapport de forme (Shape ratio) : selon ce critère, un pic est sélectionné si son rapport de forme (Yang et al., 2009) dépasse un certain seuil.
- Lignes de crête (Ridge lines) : utilisées dans la méthode basée sur la CWT (Du et al., 2006) pour détecter les vrais pics. Par exemple, dans la librairie Cromwell, les pics sont sélectionnés s'ils sont des maxima locaux et s'ils sont plus grands qu'un SNR donné. Toutefois, dans le logiciel PROcess, outre le SNR, le seuil d'intensité et le critère du rapport de forme sont également utilisés.

— Critère basé sur un modèle (Model-based criterion) : ces méthodes utilisent une fonction modèle pour ajuster les pics.
— Largeur des pics (Peak width) : un pic est détecté comme vrai si sa largeur est comprise dans un intervalle donné.
— Dérivation du spectre : pour extraire les minimas locaux (examen des zéros de la dérivée). Avec la dérivée première on conserve un maximum de pics locaux au détriment d'une élimination maximale du bruit. Avec une dérivée seconde régularisée (Savitzky-Golay par exemple) on peut détecter, ad certum modum, les pics qui se chevauchent (Dubrovkin et al., 2014).

Notons un cas particulier : la transformée en ondelette continue (CWT : Continuous Wavelet Transform) ne comporte pas d'étapes distinctes de lissage et de correction de la ligne de base. Comme le souligne Du et al (2006), la CWT peut supprimer la ligne de base d'un spectre brut (i.e. un spectre sans lissage et suppression de ligne de base). Le principe est de suivre le maximum du module de l'ondelette donnant ainsi des crêtes (ridge) qui caractérisent la régularité du signal (Mallat et al., 1992). On utilise ces dernières pour détecter les pics.

2.2.5 Alignement des pics

Les erreurs liées à l'étalonnage ou aux limitations d'un spectromètre de masse peuvent entraîner des variations entre le vecteur m/z observé et le véritable temps de vol des ions. En conséquence, lorsqu'on répète des expériences, des décalages systématiques se produisent, ce qui signifie que deux protéines identiques obtenues dans des spectres différents peuvent afficher des valeurs m/z différentes. Ces erreurs systématiques peuvent être attribuées soit à un seul instrument, soit à l'utilisation de différents instruments. De plus, un détecteur à haut débit en spectrométrie de masse peut générer de nombreux spectres par patient, introduisant ainsi des variations indésirables dans les données collectées en raison de divers facteurs tels que la réponse non linéaire du détecteur, la suppression de l'ionisation, de légères modifications dans la composition de la phase mobile, et les interactions entre les composants analysés. En outre, la résolution des pics a tendance à varier d'une expérience à l'autre et évolue également vers la fin du spectre.

Pour résoudre ce problème, on a recours à un processus appelé alignement des pics, également connu sous le nom de "peak matching". Cette méthode vise à déterminer quels pics correspondent aux mêmes peptides ou protéines dans différents échantillons. Ces pics sont soumis à un processus d'alignement qui corrige de petites dérives dans leur position m/z, résultant de l'étalonnage nécessaire pour la SM à temps de vol. Cela garantit que les peptides communs à plusieurs spectres sont identifiés et comparés en utilisant la même valeur m/z, assurant ainsi une analyse précise.

Il existe diverses méthodes pour aligner les spectres MALDI-TOF :

— l'étalonnage fréquent (Chaurand et al., 2001),
— le regroupement ou un nouvel étalonnage (Ressom et al., 2007),
— les corrélations croisées (Malyarenko et al., 2005),
— la minimisation de l'entropie (Villanueva et al. 2005),
— l'utilisation de splines cubiques (Jeffries et al., 2005),
— l'interpolation lorsque les différences de position m/z entre les spectres sont subtiles et continues, car elle permet une estimation plus précise des positions m/z des pics pour un alignement précis (He et al., 2011 ; Eriksson et al., 2020),
— l'alignement des pics basée sur une fenêtre mobile utilisée avec succès dans des travaux antérieurs (Fernández et al., 2014),
— l'alignement d'un spectre brut en ajustant l'échelle m/z pour maximiser la corrélation avec un spectre synthétique. Pour créer ce spectre synthétique, la méthode des impulsions gaussiennes centrées sur les masses spécifiées par les pics de référence. Une fois qu'une nouvelle échelle m/z est obtenue la méthode calcule un nouveau spectre en utilisant une interpolation cubique par morceaux et en l'ajustant par rapport au vecteur m/z d'origine, préservant ainsi efficacement la forme des pics (Monchamp et al., 2006),
— l'ajustements pour déterminer une fonction de déformation nécessaire à l'alignement, avec la possibilité d'incorporer d'autres paramètres pour des déformations complexes.

Suite à l'exécution de l'algorithme d'alignement, tous les pics alignés ont les mêmes valeurs de masse dans tous les spectres. Ces valeurs de masse correspondent à la masse du pic virtuel. On peut améliorer les techniques précédemment évoquées en utilisant une méthode d'optimisation plus performante (Voir le tableau Annexe A

Table S1, dans la colonne "Peak alignment" les articles présentant/mentionnant des méthodes d'alignement des spectres). L'avantage principal de l'alignement réside dans sa capacité à résoudre efficacement le délicat problème de la correspondance entre les pics des spectres. Un léger désalignement est généralement tolérable car il ne fait que légèrement élargir les pics. Cependant, un désalignement significatif peut rendre les données inutilisables.

2.3 Les transformées du signal

2.3.1 La Transformée de Fourier

La *transformée de Fourier* (FT : Fourier Transform) est une méthode très connue pour traiter les signaux analytiques. Les techniques couramment utilisées permettant d'appliquer directement la transformée de Fourier sur des données sont : la*transformée de Fourier à court terme* (STFT : Short-Time Fourier Transform) et la *transformée de Fourier rapide* (FFT : Fast Fourier Transform).

Théorie

Une *transformée de Fourier à court terme* (STFT) est déterminée à partir d'un signal en effectuant une *transformée de Fourier discrète* (DFT) sur une fenêtre mobile de petite taille, parcourant l'intégralité de la durée du signal. L'emplacement de chaque entrée dans une DFT détermine son temps (axe x) et sa fréquence (axe y).

Il est important de noter que chaque cellule de notre STFT est **complexe**, ce qui signifie que chaque entrée contient à la fois une composante de magnitude et une composante de phase.
Dans la plupart des applications typiques, la STFT est réalisée sur un ordinateur à l'aide de *la transformée de Fourier rapide* FFT (Fast Fourier Transform), de sorte que les deux variables soient discrètes et quantifiées.

Mathématiquement, la DFT, la STFT à temps discret et la FFT sont défini comme suit :

Soit x une séquence (ou signal) de N échantillons (ou discrétisation) pour $n \in \{0, ..., N-1\}$. La DFT comprend des fréquences de k cycles sur N échantillons, $k = 0, ..., N-1$.

$$X_{DFT}(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk}$$

où $W_N = e^{\frac{-2i\pi}{N}}$

La STFT est une séquence de transformées de Fourier d'un signal fenêtré. Elle fournit des informations de fréquence localisées dans le temps pour les situations où les composantes de fréquence d'un signal varient dans le temps, alors que la FT fournit des informations de fréquence moyennées sur l'ensemble de l'intervalle de temps du signal.

Elle est définie comme suit :
Le m-ième bloc fenêtré du signal x est donné par $X_{STFT}(m, k)$:

$$X_{STFT}(m, k) = \sum_{n=0}^{L-1} x(n) g(n - m) W_L^{nk}$$

où x représente un signal discrétisé en N échantillons et g une fonction de fenêtre en L points.

La valeur absolue d'un pas (bin) de la transformé de Fourier, $|X(t, f)|$ au moment t et à la fréquence f, détermine la quantité d'énergie entendue à partir de la fréquence f au moment t.

La transformée de Fourier, comprend une version continue qui utilise une intégrale plutôt qu'une somme appelée *transformée de Fourier continue*. La version discrète de la transformée est généralement utilisée pour les opérations informatiques. Même sous la forme discrète, la plupart des ordinateurs n'ont pas la puissance de calcul nécessaire pour résoudre l'équation brute de la transformée. La *transformée de Fourier rapide* (FFT) possède

des propriétés intéressantes qui facilitent le calcul.

Elle est définie comme suit :

$$X_{FFT}(k) = \sum_{n=0}^{\frac{N}{2}-1} x(2n) W_{\frac{N}{2}}^{nk} + W_N \sum_{n=0}^{\frac{N}{2}-1} x(2n+1) W_{\frac{N}{2}}^{nk}$$

pour les mêmes hypothèses que celles définies précédemment.

Cette équation divise la somme des produits en deux : une le long des indices impairs et une autre le long des indices pairs. Cette procédure peut être modélisée de manière bien plus efficace à l'aide de processus informatiques plutôt qu'avec la DFT.

Utilisations sur les spectres du MALDI-TOF

La transformée de Fourier (FT) est un outil puissant pour analyser les spectres MALDI-TOF, en particulier pour extraire des informations sur les masses moléculaires, la composition chimique et d'autres caractéristiques des ions présents dans l'échantillon. Elle permet également d'améliorer la résolution spectrale et de faciliter l'interprétation des données spectrales.
Elle est couramment utilisée pour convertir un signal dans le domaine temporel en un signal dans le domaine fréquentiel. Dans le contexte de la SM MALDI-TOF, voici quelques utilisations possibles de la transformée de Fourier :

- Amélioration de la résolution spectrale : La FT peut être utilisée pour améliorer la résolution spectrale d'un spectre MALDI-TOF en augmentant le nombre de points dans le domaine fréquentiel. Cela peut permettre de distinguer plus facilement des pics qui pourraient se chevaucher dans le domaine tempore (Whistler et al. 2007 ; Wong et al. 2005).
- Identification des ions : En convertissant un spectre MALDI-TOF dans le domaine fréquentiel, on peut identifier les différentes fréquences (ou masses) des ions présents dans l'échantillon. Cela peut être utile pour déterminer la composition chimique de l'échantillon (Haegler et al. 2009).
- Filtrage et suppression du bruit : La FT peut être utilisée pour appliquer des filtres dans le domaine fréquentiel, ce qui peut permettre de supprimer le bruit indésirable ou de mettre en évidence des caractéristiques spectrales spécifiques (Shin et al. 2007 ; Zhang et al. 2007 ; Dubrovkin 2019 ; Conrad et al. 2017).
- Déconvolution : La FT peut également être utilisée pour effectuer des opérations de déconvolution sur les spectres MALDI-TOF, ce qui peut aider à séparer les signaux provenant de différentes espèces ioniques et à améliorer la précision de la mesure des masses (Dubrovkin 2019 ; Conrad et al. 2017).
- Analyse de la distribution de masse : En utilisant la FT, il est possible d'analyser la distribution de masse des ions présents dans l'échantillon, ce qui peut fournir des informations sur la variabilité des masses moléculaires (Rockwood et al. 1995 ; Rockwood et al. 2003 ; Shin et al. 2007).
- Étude des changements structuraux : La FT peut être utilisée pour suivre les changements structuraux dans les ions au fil du temps, ce qui peut être important dans des études cinétiques ou de réaction (Apicella et al. 2013).

2.3.2 La Transformée en Ondelette

Comme évoqué dans les méthodes de pré-traitement des spectres exposées dans les sections précédentes, il est envisageable d'appliquer la transformée en ondelette.

Les méthodes basées sur la transformation en ondelette sont souvent utilisées pour le débruitage des spectres (Du et al., 2006 ; Coombes et al., 2005). Il existe essentiellement deux types de méthodes basées sur la transformation en ondelette, à savoir *la transformation en ondelette discrète* (DWT : Discrete Wavelet Transform) et *la transformation en ondelette continue* (CWT : Continuous Wavelet Transform). Ces méthodes transforment les spectres de masse dans le domaine des ondelettes et les représentent en termes de coefficients d'ondelettes à plusieurs échelles.

Théorie

L'ondelette est une méthode de fenêtrage avec différentes résolutions pour les régions. La décomposition en ondelettes permet de représenter un signal dans un plan temporel en utilisant une échelle au lieu d'une fréquence. Une ondelette, une petite forme d'onde, est une onde localisée pendant une durée limitée. Si l'on compare l'ondelette à la transformée de Fourier, l'analyse de Fourier décompose le signal en sinusoïdes de différentes fréquences, tandis que l'ondelette décompose le signal en formes décalées ou mises à l'échelle à partir d'une ondelette mère $\psi(t)$.

La transformée en ondelettes continue (CWT) pour le signal $x(t)$ est définie comme l'intégration de $x(t)$ avec les formes décalées ou mises à l'échelle $\psi_{a,b}(t)$:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) * \psi(\frac{t-b}{a}) dt$$

où $a \in \mathbb{R}^+ \backslash \{0\}$, $b \in \mathbb{R}$

En d'autres termes, la CWT est la somme du signal multiplié par des formes décalées et mises à l'échelle à partir de ψ :

$$CWT(scale, position) = \int_{-\infty}^{+\infty} x(t) * \psi(scale, position, t) dt$$

L'ondelette de base originale $\psi(t)$ est appelée ondelette mère, et ses variations $\psi_{a,b}(t)$ sont appelées ondelettes filles. Le a est un facteur d'échelle pour la mise à l'échelle de la fonction $\psi(t)$, tandis que b est un facteur de décalage pour la translation de la fonction $\psi(t)$. Le résultat de la CWT est une matrice remplie de coefficients d'ondelettes localisés par échelle et par position.

La transformée en ondelette discrète calcule les échelles et les translations sur la base de la puissance de deux. La procédure de calcul de la transformée en ondelettes discrète est la suivante où $h[n]$ est un filtre passe-haut et $g[n]$ un filtre passe-bas et un niveau de niveau de coefficients fixe :

1. Le signal est décomposé simultanément par un filtre passe-bas $g[n]$ et un filtre passe-haut $h[n]$.
2. La sortie de $h[n]$ est ensuite échantillonnée par deux pour générer des coefficients de détail et la sortie de $g[n]$ est échantillonnée par deux pour générer des coefficients d'approximation. Les coefficients obtenus à partir de la sortie de $h[n]$ sont appelés coefficients de niveau 1.
3. La sortie de $g[n]$ passe par un autre groupe de filtres passe-haut et passe-bas. Les étapes 1. et 2. se poursuivent jusqu'à l'obtention du dernier niveau de coefficients.

L'avantage de la transformée en ondelette discrète par rapport à la transformée en ondelettes continue est son efficacité, car elle ne calcule que les échelles et les positions basées sur la puissance de deux, tandis que la redondance de la transformée en ondelettes continue facilite l'interprétation de la détection des pics (Du et al., 2006).

Utilisations sur les spectres du MALDI-TOF

La transformée en ondelettes continue (CWT) calcule les transformées en ondelettes à chaque échelle tout en capturant plus d'informations sur les pics du spectre de masse (Du et al., 2006). Elle est notamment utilisée pour la détection et sélection des pics. Ces dernières années, l'algorithme basé sur la transformée en ondelettes continue (CWT) a été largement étudié en raison de sa précision, de ses performances et de sa nature multi-échelle. Compte tenu de la variation de la forme des pics en fonction de la masse (qui deviennent plus larges et plus bas à des masses plus élevées), l'utilisation d'ondelettes, avec leur adaptabilité et leur capacité à traiter plusieurs échelles, s'avère un choix judicieux pour le nettoyage des spectres de masse. En effectuant une transformation du spectre dans l'espace des ondelettes, l'algorithme peut exploiter les informations supplémentaires contenues dans la forme des pics, ce qui permet de réduire les faux positifs. (Du et al., 2006) Cependant, malgré son statut de méthode de détection de pics privilégiée pour de nombreux chercheurs en spectrométrie de masse (Yang et al., 2009), l'algorithme basé sur la CWT présente des limitations lorsqu'il s'agit d'identifier des pics de faible amplitude ou des pics qui se superposent. La CWT est variable en fonction du décalage : un petit décalage dans la position de départ du spectre peut entraîner une baisse importante des performances.

Afin de surmonter cette limitation, Coombes et al. ont proposé la transformation en ondelettes discrète non décimée (UDWT), qui est une version améliorée de la DWT pour le débruitage des spectres (Lang et al., 1996 ; Coombes et al., 2005) en réduisant les artefacts dus au décalage du spectre sur l'axe du rapport masse/charge (c'est-à-dire l'erreur de masse). Dans les applications, il a été signalé qu'elle permettait un meilleur débruitage

qualitatif (Coombes et al., 2005). De par son invariance par rapport au décalage, la littérature semble converger vers l'utilisation des DWT dans le débruitage/ lissage (Coombes et al., 2007).

Comme mentionné dans les sections précédentes, elle est aussi utilisée afin de corriger la ligne de base. Elle est disponible dans les librairies gratuites dans différents progiciels tels que Cromwell (Matlab), PROcess (R), MALDIquant (R), SpecAlign (Java) ou MassSpecWavelet (R) (Du et al., 2006).

Il est possible d'utiliser les deux formes de transformée en ondelette pour traiter les spectres de MALDI-TOF. Cependant, il est important de noter que l'utilisation de cette transformée peut nécessiter une expertise en traitement du signal et en mathématiques avancées. De plus, le choix de l'ondelette et des paramètres de la transformée peut avoir un impact significatif sur les résultats, il est donc essentiel de bien paramétrer l'analyse en fonction des spécificités des données et des objectifs de recherche.

2.3.3 Discussion générale sur le traitement des spectres

Dans cette section intitulée "Le traitement des spectres", nous avons exposé les étapes conventionnellement adoptées pour le pré-traitement des spectres MALDI-TOF dans le dessein d'améliorer leur qualité. L'objectif de ces étapes consiste à rendre les spectres aussi exploitables que possible. Toutefois, il convient de noter que, bien que ces techniques soient largement employées, elles ne revêtent pas nécessairement le caractère optimal et appellent fréquemment des améliorations et des optimisations. De surcroît, il existe un éventail considérable de méthodes de pré-traitement des spectres qui n'ont pas été abordées ici. Des chercheurs en mathématiques ont exploré diverses approches pour résoudre des problèmes spécifiques liés au pré-traitement. Par exemple, Starostin et al. (2020) ont utilisé une approche géométrique pour détecter les pics, tandis que Zhang et al. (2008) ont proposé un modèle bayésien non paramétrique (Zhang et al., 2008 ; House et al., 2011). Conrad et al. (2017) ont utilisé la méthode de Compress Sensing pour détecter les pics discriminants dans leur algorithme.

Le pré-traitement des spectres reste un sujet de recherche dynamique, suscitant l'intérêt de nombreux chercheurs en SM qui s'efforcent constamment d'améliorer la qualité des spectres générés par le MALDI-TOF. Cependant, dans le cadre de cette thèse, notre objectif n'est pas d'évaluer les diverses techniques pour déterminer la meilleure, mais plutôt de choisir celle qui convient le mieux et qui est la plus efficace pour un traitement de base. Après avoir identifié le pré-traitement approprié, notre intention est d'appliquer des algorithmes d'apprentissage profond (Deep Learning) aux données traitées, sans nous engager dans la recherche d'une technique de sélection des pics discriminants, car nous souhaitons que le modèle s'occupe de cette étape. C'est pourquoi nous avons opté pour des techniques de pré-traitement simples, légères et efficaces.

En résumé, en ce qui concerne le traitement des spectres, notre objectif est de préserver autant d'informations que possible dans le spectre tout en éliminant le bruit évident, de manière à ce que les informations résiduelles puissent être utilisées par le modèle pour effectuer des prédictions.

Un tableau de synthèse en Annexe A Table S1 a été élaboré dans le but de répertorier les techniques de pré-traitement décrites dans la littérature, en incluant à la fois les méthodes couramment utilisées et les approches originales. Nous avons essayé d'inclure le plus grand nombre possible d'articles variés. Dans ce tableau, les étapes de pré-traitement ont été distinctement séparées pour mettre en évidence celles qui ont été employées et celles qui ne l'ont pas été dans chaque article répertorié. Pour chaque étape, les techniques utilisées ont été résumées.

2.4 Spectres MALDI-TOF, approches statistiques et apprentissage automatique

Au cours de la dernière décennie, le MALDI-TOF, grâce à des algorithmes automatisés d'analyse et d'interprétation des données, a révolutionné le diagnostic clinique en permettant une identification rapide, fiable et économique des espèces microbiennes, ainsi qu'une rationalisation potentielle des tests de sensibilité aux antimicrobiens (Croxatto et al., 2016 ; Moreno-Camacho et al., 2018 ; Vrioni et al., 2018). Elle surpasse largement les méthodes de diagnostic traditionnelles en termes de coût, de vitesse et de précision pour identifier les espèces microbiennes (Scola et al., 2009 ; Stevenson et al., 2010 ; Foster et al., 2013 ; Haigh et al., 2013 ; Tadros et al., 2013). Cet instrument est couramment utilisé en recherche et en clinique, principalement pour l'identification

des micro-organismes tels que les bactéries, les levures et les champignons filamenteux. De plus, des recherches explorent son utilisation pour diagnostiquer d'autres maladies que les infections microbiennes, dans le but de découvrir des biomarqueurs de maladies et de permettre des diagnostics précoces, rapides et non invasifs (Hou et al., 2019 ; López-Cortés et al., 2021).

Dans les paragraphes suivants, nous présentons les méthodes algorithmiques statistiques traditionnelles ainsi que les algorithmes d'apprentissage automatique couramment employés dans le cadre de l'analyse et de l'identification des spectres MALDI-TOF.

2.4.1 Approches statistiques classiques

Afin de repérer des pics significatifs, les chercheurs emploient une gamme d'approches statistiques, notamment le test de Student, le test ANOVA (Fernández et al. 2014 ; Delavy et al. 2020), le test de Fisher, ainsi que le test de Wilcoxon (Cuénod et al. 2021). Ils font également usage d'algorithmes de sélection de caractéristiques, tels que la méthode basée sur le gain d'information (information gain-based feature selection) et celle basée sur la corrélation (correlation-based feature selection) (Appavu et al., 2011 ; Balasubramanian et al., 2022).

Pour parvenir à un regroupement ou à une distinction des groupes d'empreintes de masse, des algorithmes statistiques tels que l'analyse en composantes principales (PCA : Principal Component Analysis) (Pais et al. 2019 ; Del Prete et al. 2021 ; Fernández et al. 2014 ; Mortier al. 2021 ; Deulofeu et al. 2023 ; Rodríguez-Temporal et al. 2023) et l'analyse de regroupement hiérarchique (HCA) (Del Prete et al. 2021 ; Pizzato et al. 2022) ont souvent été utilisés. L'analyse en composantes principales consiste à regrouper des objets de telle sorte que les objets d'un même groupe soient plus semblables les uns aux autres que ceux des autres groupes. L'ACP est une méthode de réduction de dimension qui explique la variation d'un grand nombre de réponses originales (comme les centaines de pics d'empreintes digitales de masse) à l'aide d'un plus petit nombre de facteurs appelés composantes principales. Le regroupement hiérarchique (HC : Hierarchical Clustering) est une méthode d'analyse de regroupement qui vise à construire une hiérarchie de classes (i.e. clusters), fréquemment utilisée après une ACP.

2.4.2 Apprentissage automatique supervisé

Traditionnellement, l'analyse des spectres de masse MALDI-TOF se limitait à quelques caractéristiques empiriquement associées aux espèces microbiennes, telles que la hauteur et l'aire sous les pics. Cependant, cette approche, bien qu'efficace pour identifier les espèces, sous-exploitait la richesse d'informations contenues dans ces spectres. Pour combler cette lacune, des chercheurs ont intégré des algorithmes d'apprentissage automatique dans leurs travaux.
Les méthodes d'apprentissage automatique, en explorant les dépendances statistiques et en prenant en compte les interactions non linéaires entre les caractéristiques, ont permis de révéler de nouvelles informations jusque-là inconnues dans les spectres MALDI-TOF. Ces informations se sont avérées particulièrement précieuses pour l'identification et la distinction d'espèces microbiennes, en particulier celles qui sont phylogénétiquement proches ou les sous-lignées d'espèces. De plus, il a été récemment reconnu que les données des spectres MALDI-TOF peuvent également être exploitées pour établir des profils de résistance aux antibiotiques. Ainsi, les méthodes d'apprentissage automatique ont contribué à améliorer l'identification des espèces microbiennes et à rationaliser l'évaluation de la résistance aux antimicrobiens, ouvrant ainsi de nouvelles perspectives dans la recherche en microbiologie (DeBruyne et al., 2011 ; Vervier et al., 2015 ; Fangous et al., 2014 ; Burckhardt et al., 2018 ; Florio et al., 2018 ; Sogawa et al., 2017 ; Mather et al., 2016).

Principes de base

L'apprentissage automatique est un domaine de l'intelligence artificielle qui se concentre sur la création de systèmes informatiques capables d'apprendre à partir de données. L'apprentissage automatique supervisé est l'une de ses sous-catégories les plus courantes. Dans l'apprentissage automatique supervisé, un modèle est formé à partir d'un ensemble de données étiquetées, où chaque exemple de données est associé à une étiquette ou une réponse correcte. Le modèle apprend à partir de ces exemples en identifiant des motifs et des relations entre les données d'entrée et les étiquettes correspondantes. Une fois que le modèle est entraîné, il peut être utilisé pour faire des prédictions ou des classifications sur de nouvelles données non étiquetées en se basant sur les schémas qu'il a appris pendant l'entraînement. En résumé, l'apprentissage automatique supervisé consiste à enseigner à

un modèle à partir d'exemples étiquetés pour qu'il puisse effectuer des tâches de prédiction ou de classification sur de nouvelles données.

Exploration des approches d'apprentissage automatique supervisé appliquées aux spectres MALDI-TOF

Plusieurs algorithmes d'apprentissage automatique supervisés présentent des approches variées. Pour les spectres MALDI-TOF, voici les algorithmes couramment employés :

- Support Vector Machine (SVM) : Les machines à vecteurs de support sont des algorithmes d'apprentissage supervisé qui identifient l'hyperplan séparateur à marge maximale entre les classes dans un espace de dimension supérieure, utilisant des fonctions noyaux telles que le noyau de base radiale ou polynomial. Le SVM est principalement utilisé comme algorithme de sélection des pics (caractéristiques) dans ClinProTools. Ensuite, la classification des nouvelles instances est réalisée en utilisant un algorithme kNN basé sur ces pics sélectionnés. Pour la classification de spectres par SVM, une étape préalable de sélection des pics est systématiquement effectuée avant l'utilisation du SVM sur les pics sélectionnées (Ressom et al. 2007 ; Fernández et al., 2014 ; Conrad et al. 2017 ; Mortier et al. 2021 ; Rodríguez-Temporal et al. 2023).
- Linear Discriminant Analysis (LDA) : L'analyse discriminante linéaire est une méthode de classification et de réduction de dimension qui projette les données dans un espace de dimension réduite tout en maximisant la séparation entre les classes (Delavy et al. 2020 ; Rodríguez-Temporal et al. 2023).
- Decision Trees (DT) : Les arbres de décision sont des modèles qui prennent des décisions en suivant une série d'étapes basées sur des caractéristiques des données (Fernández et al., 2014). La version la plus optimisé de cet algorithme qui est le gradient-boosted decision trees (LightGBM) est souvent utilisé pour la classification de spectres MALDI-TOF (Weis et al., 2022).
- Random Forests (RF) : Les forêts aléatoires sont des ensembles d'arbres de décision. Ils combinent plusieurs arbres pour améliorer la précision et la robustesse des prédictions (López-Fernández et al., 2015 ; Delavy et al. 2020 ; Mortier al. 2021 ; Rodríguez-Temporal et al. 2023).
- K-Nearest Neighbors (kNN) : les k plus proches voisins est un algorithme de classification qui attribue une étiquette à un point de données en se basant sur les étiquettes des k points les plus proches dans l'ensemble de données d'entraînement (López-Fernández et al., 2015 ; Mortier al. 2021 ; Rodríguez-Temporal et al. 2023).
- Logistic regression (LR) : La régression logistique est utilisée pour la classification binaire. Elle modélise la probabilité qu'une observation appartienne à une classe particulière en utilisant une fonction logistique (Delavy et al. 2020 ; Weis et al. 2021 ; Mortier al. 2021).
- Naïve Bayes (NB) : Le classificateur Naïve Bayes est basé sur le théorème de Bayes. Il est utilisé pour la classification et repose sur l'hypothèse que les caractéristiques sont indépendantes les unes des autres, d'où le terme "naïf" (Fernández et al., 2014, Gong et al., 2021, Li et al, 2022).
- Quick Classifier (QC) : Disponible sur ClinProTools, cet algorithme calcule la surface moyenne de chaque pic et fournit une valeur p par classe. Lors de la classification, les surfaces des pics sont triées par l'algorithme de tri univarié et une moyenne de tous les pics est calculée pour indiquer l'appartenance à une classe (Weis et al., 2020 ; Fiamanya et al., 2022).
- Algorithme génétique (AG) : Également disponible sur ClinProTools, ce sont des algorithmes d'optimisation qui s'inspirent des processus biologiques tels que la mutation, le croisement et la sélection. Ils améliorent progressivement une collection de solutions candidates pour atteindre une solution optimale. L'algorithme génétique est utilisé pour sélectionner une combinaison de pics qui séparent les classes, en utilisant une fonction de coût qui mesure la variance entre les classes. Lorsque l'algorithme génétique est sélectionné dans ClinProTools, il n'est utilisé que comme algorithme de sélection des pics. La classification des instances inédites est effectuée à l'aide d'un algorithme kNN basé sur les pics sélectionnés (Weis et al., 2020 ; Fiamanya et al., 2022).

L'empreinte de masse MALDI-TOF simplifie le processus d'identification des biomarqueurs en détectant les pics significatifs après une analyse comparative des motifs d'empreinte. Ces pics peuvent également être sélectionnés en utilisant des algorithmes d'apprentissage automatique tels que le moindre rétrécissement absolu (Lasso), l'opérateur de sélection, l'analyse discriminante des moindres carrés partiels (PLS-DA) (Rodríguez-Temporal et al. 2023), et la méthode d'élimination récursive des caractéristiques avec validation croisée (RFE).

Les plateformes et logiciels équipés d'algorithmes de traitement, d'analyse et de classification par apprentissage automatique et approches statistiques simplifient considérablement la tâche d'analyse et d'interprétation des données. Parmi les exemples de ces outils, on peut citer les ateliers d'apprentissage automatique Weka et

Scikit-Learn, les packages R comme MALDIquant, ainsi que les outils ClinProTools et FlexAnalysis de Bruker Daltonics, qui sont utilisés pour analyser les spectres de masse MALDI-TOF. D'autres options populaires incluent la plateforme Mass-Up, entre autres. Pour mener à bien leurs études, la plupart des chercheurs utilisent des logiciels tels que R, R Studio, MATLAB, Python, MALDI Biotools 3.0, Statistics Program for Social Sciences, Mathematica, ou une combinaison de ces outils.

Néanmoins, il est important de noter que les modèles d'apprentissage automatique (Machine Learning) supervisé, tels que les machines à vecteurs de support (SVM), les arbres de décision, celles basées sur les arbres de décision, les plus proches voisins ou les méthodes de régression ont généralement une structure de modèle relativement simple. Ils sont adaptés lorsque les données d'entrée sont moins complexes et que les relations entre les caractéristiques sont relativement linéaires. Ils peuvent fonctionner efficacement avec des ensembles de données de taille moyenne à grande, mais ils peuvent avoir du mal à extraire des informations significatives à partir de données massives sans une ingénierie de caractéristiques appropriée. En effet, généralement, dans l'apprentissage automatique supervisé, les caractéristiques doivent être extraites et sélectionnées manuellement à partir des données brutes. Cela nécessite une expertise dans ce domaine. Souvent, une phase préliminaire appelée "ingénierie des caractéristiques" ou "extraction des caractéristiques" (features engineering/extraction) est réalisée avant d'entrer les caractéristiques filtrées dans le modèle.

2.4.3 Qu'en est-il de l'apprentissage profond ?

Principes de base

L'apprentissage profond supervisé est une approche d'apprentissage automatique supervisé où un modèle, généralement un réseau de neurone, est entraîné à partir d'un ensemble de données étiquetées. Le modèle apprend à partir de ces étiquettes pour effectuer des prédictions ou des classifications sur de nouvelles données. Les réseaux de neurones inspirés du fonctionnement des neurones biologiques, et qui par la suite se sont rapprochés des méthodes statistiques, sont composés de couches de neurones interconnectés qui apprennent à partir des données. Il est particulièrement adapté pour extraire des caractéristiques complexes et abstraites à partir de données brutes, comme des images, du texte ou des séquences, ce qui en fait une méthode puissante pour un large éventail de tâches, de la reconnaissance d'images, à la traduction automatique. Sa capacité à capturer des motifs complexes et à extraire des caractéristiques abstraites en fait un outil puissant pour l'analyse et l'interprétation des données.

Réseaux de neurones et spectres MALDI-TOF

Les réseaux de neurones sont conçus pour capturer des modèles non linéaires et des caractéristiques abstraites à partir des données. Ils sont particulièrement adaptés aux données complexes, telles que les images, les sons et les séquences, mais peuvent également être utilisés pour les spectres MALDI-TOF. Les modèles d'apprentissage profond ont tendance à exceller lorsque les données sont volumineuses, car ils peuvent exploiter la capacité de leurs nombreuses couches pour extraire automatiquement des caractéristiques pertinentes à partir de données brutes. Les réseaux de neurones peuvent apprendre à partir des données brutes sans nécessiter une ingénierie de caractéristiques intensive. Ils sont capables d'apprendre des caractéristiques hiérarchiques à partir des données, ce qui peut être particulièrement utile pour les spectres MALDI-TOF, où la structure des données peut être complexe.

Contrairement aux modèles évoqués dans la section précédente, l'apprentissage profond, plus précisément l'utilisation des réseaux neuronaux, offre la possibilité d'intégrer à la fois la sélection des caractéristiques et la prédiction au sein d'un même modèle. Un modèle unique peut accomplir ces deux tâches si ses architectures sont correctement conçues. Cette approche ne supprime pas complètement les biais, mais les réduit au minimum en utilisant un mécanisme unifié.

Voici ce que l'on peut dire sur son utilisation :

1. Amélioration de la précision de l'identification : L'apprentissage profond, en particulier les réseaux de neurones, peut être utilisé pour extraire des caractéristiques complexes et non linéaires à partir des spectres MALDI-TOF. Cela permet d'améliorer la précision de l'identification des espèces microbiennes, notamment des bactéries, des levures et des champignons.
2. Identification multiniveaux : Les réseaux de neurones peuvent être formés pour identifier les micro-organismes à différents niveaux, y compris le genre, l'espèce et même la souche. Cela permet une identi-

fication plus détaillée et précise.

3. Extraction de biomarqueurs : L'apprentissage profond peut être utilisé pour extraire des biomarqueurs à partir des spectres MALDI-TOF, ce qui peut être précieux pour la recherche en biomédecine et la découverte de nouvelles informations sur les échantillons.
4. Réduction du bruit : Les réseaux de neurones sont capables de réduire le bruit dans les spectres, ce qui peut améliorer la qualité des données et rendre les analyses plus robustes.
5. Automatisation : L'utilisation de l'apprentissage profond permet d'automatiser de nombreuses étapes du processus d'analyse des spectres MALDI-TOF, ce qui peut accélérer considérablement les flux de travail en laboratoire.

Il convient de noter que l'utilisation de l'apprentissage profond (Deep Learning) nécessite généralement de grandes quantités de données d'entraînement et une expertise en modélisation. De plus, l'interprétation des modèles d'apprentissage profond peut être complexe, ce qui peut poser des défis en termes de transparence et d'explicabilité des résultats. Néanmoins, l'application de l'apprentissage profond aux spectres MALDI-TOF a révélé un fort potentiel pour améliorer la précision des analyses et la capacité à extraire des informations pertinentes à partir de ces spectres. Il est important de noter qu'à l'heure actuelle, le nombre de publications portant sur l'utilisation de réseaux neuronaux pour analyser les spectres MALDI-TOF est inférieur à celui des publications traitant de l'application d'algorithmes d'apprentissage automatique à ces spectres. De plus, la plupart des réseaux neuronaux utilisés, se limitent généralement à des modèles simples tels que les perceptrons multi-couches (Tong et al., 2011 ; Weis et al. 2021 ; Deulofeu et al. 2023) ou les réseaux de neurones à convolution à une dimension (Mortier al. 2021). Cependant, des recherches récentes indiquent que des bio-informaticiens et des bio-mathématiciens se penchent sur l'exploration de modèles plus complexes dans le but d'améliorer la précision des prédictions (Li et al. 2022 ; Merchan et al. 2023).

Suite à cette observation, j'ai décidé d'entreprendre des recherches dans le domaine de l'apprentissage profond. Mon objectif est d'explorer divers modèles de réseaux de neurones afin d'évaluer leur potentiel pour être intégrés aux spectres de masse MALDI-TOF, avec, pour finalité, d'améliorer la précision des prédictions dans le contexte de la surveillance épidémiologique des maladies infectieuses.

2.5 Principes d'évaluations

Les métriques permettant d'évaluer les performances des modèles, dans les différents projets exposés dans ce mémoire, sont définis dans cette section.

Accuracy (Exactitude)

Définition : L'accuracy mesure la proportion de prédictions correctes parmi l'ensemble des prédictions effectuées.
Formule :

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

où :
VP (True Positives) : Nombre de vrais positifs VN (True Negatives) : Nombre de vrais négatifs FP (False Positives) : Nombre de faux positifs FN (False Negatives) : Nombre de faux négatifs

Precision (Précision)

Définition : La précision mesure la proportion de vrais positifs parmi les prédictions positives.
Formule :

$$\text{Precision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Recall (Rappel)

Définition : Le rappel mesure la proportion de vrais positifs parmi toutes les occurrences réelles de la classe positive.
Formule :

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

F1 Score

Définition : Le F1 Score est une métrique qui combine à la fois la précision et le rappel en une seule mesure, utile lorsque les classes sont déséquilibrées.
Formule :

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

Spécificité (Specificity)

Définition : La spécificité mesure la proportion de vrais négatifs parmi toutes les occurrences réelles de la classe négative.
Formule :

$$\text{Spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Balanced Accuracy (Accuracy équilibrée)

Définition : L'accuracy équilibrée est une mesure de l'accuracy prenant en compte le déséquilibre entre les classes.
Formule :

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

où :
Sensitivity (Sensibilité) équivaut au Recall Specificity (Spécificité) mesure la capacité à détecter les vrais négatifs

Mean Absolute Error (Erreur absolue moyenne)

Définition : L'erreur absolue moyenne mesure la moyenne des valeurs absolues des écarts entre les prédictions et les valeurs réelles.
Formule :

$$\text{MAE} = \frac{\sum |y_i - \hat{y}_i|}{n}$$

où :
y_i : Valeur réelle
$\hat{y}_i$: Valeur prédite
n : Nombre d'observations

Erreur Quadratique Moyenne (MSE)

Définition : L'erreur quadratique moyenne (Mean Squared Error, MSE) est une mesure de la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles dans un ensemble de données. Elle est couramment

utilisée pour évaluer la performance des modèles de régression.
Formule :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Où : - n est le nombre d'observations dans l'ensemble de données. - y_i est la valeur réelle de la variable cible pour la i-ème observation. - $\hat{y}_i$ est la valeur prédite pour la i-ème observation.

R-squared Score (Coefficient de détermination)

Définition : Le coefficient de détermination mesure la proportion de la variance totale de la variable dépendante expliquée par le modèle.
Formule :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

où :
y_i : Valeur réelle
$\hat{y}_i$: Valeur prédite
$\bar{y}$: Moyenne des valeurs réelles

ROC AUC Score or AUROC (Aire sous la courbe ROC)

Définition : L'Aire sous la courbe ROC est une métrique qui mesure la capacité d'un modèle de classification à discriminer entre les classes.
Formule : Calculée à partir de la courbe ROC, qui représente le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) à différents seuils de classification.

Matrice de confusion (Confusion Matrix)

La matrice de confusion n'est pas une métrique, mais elle représente un tableau qui résume les résultats de classification en mettant en évidence les vrais positifs, les vrais négatifs, les faux positifs et les faux négatifs.

	Prédit Positif	Prédit Négatif
Réellement Positif	VP	FN
Réellement Négatif	FP	VN

Ces métriques sont couramment utilisées pour évaluer les performances des modèles de classification (accuracy, precision, recall, F1 score, balanced accuracy, ROC AUC score, confusion matrix) et des modèles de régression (mean absolute error, R-squared score).

2.6 Conclusion

La SM par MALDI-TOF a révolutionné l'identification des micro-organismes en microbiologie clinique grâce à sa simplicité, précision, rapidité et rentabilité. Cette technologie détecte rapidement les protéines et peptides, réduisant les coûts de diagnostic. Elle est prisée en microbiologie pour sa sensibilité dans la plage de masse de 2000 à 20 000 Da. Les outils informatiques jouent un rôle essentiel en permettant le pré-traitement des données MALDI-TOF pour éliminer le bruit et extraire des connaissances pertinentes. Des méthodes de pré-traitement inadéquates peuvent biaiser les données, entravant la formulation de conclusions significatives, soulignant ainsi l'importance du pré-traitement pour une analyse rigoureuse. La gestion et l'analyse des données MALDI-TOF suscitent un intérêt croissant, avec le développement continu d'outils dédiés. Dans cette optique, plusieurs algorithmes et outils ont été développés pour répondre à ces besoins spécifiques.

Ces dernières années, elle sollicite particulièrement l'usage des méthodes statistiques et de l'apprentissage automatique pour découvrir des biomarqueurs, classifier automatiquement des échantillons et regrouper des échantillons. Cependant, les méthodes d'apprentissage automatique supervisé, bien que simples, exigent souvent une ingénierie de caractéristiques avancée et une sélection préalable des caractéristiques, les rendant fortement dépendantes du pré-traitement.

Toutefois, l'apprentissage profond, plus particulièrement les réseaux de neurone, est une approche encore en développement qui se distingue par sa capacité à utiliser les spectres complets sans nécessiter de pré-traitement particulier tout en automatisant la réduction du bruit, la sélection des pics et l'application d'algorithmes de classification ou de régression. . Il convient de souligner que l'apprentissage profond dans l'identification des spectres offre un potentiel significatif pour la surveillance des épidémies hospitalières bactériennes et fongiques, ainsi que des maladies vectorielles comme le paludisme. Ses multiples applications, allant du diagnostic précoce à la prédiction du pronostic, en passant par la surveillance de l'évolution de la maladie et l'identification des patients répondant le mieux aux traitements spécifiques, répondent aux besoins des cliniciens et des chercheurs en laboratoire. Cette approche ouvre ainsi de nouvelles perspectives dans le domaine du diagnostic et de la surveillance épidémiologique des maladies infectieuses.

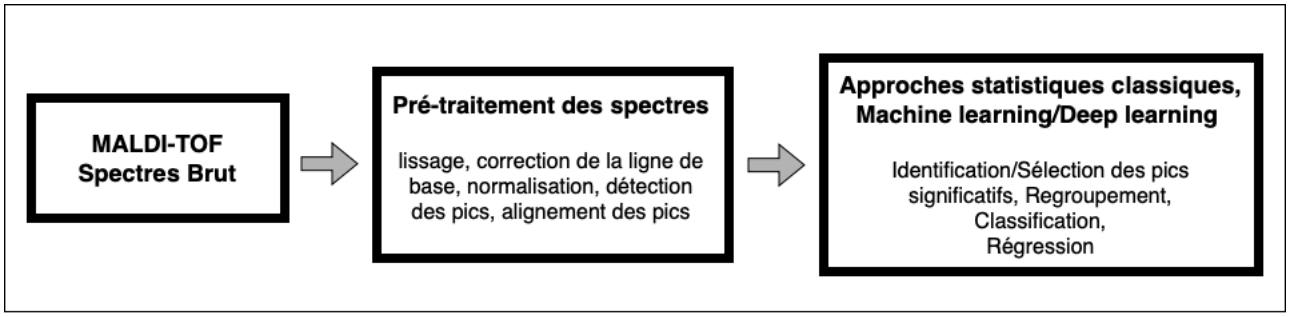


FIGURE 2.7 – **Récapitulatif des spécifications techniques abordées dans ce chapitre.** Source : Photothèque personnelle

Chapitre 3

Modèle d'apprentissage profond couplé à la protéomique pour détecter les clones d'épidémies de levure dans plusieurs hôpitaux

Ce chapitre est issu des recherches présentées dans l'article publié intitulé : "Improving the Detection of Epidemic Clones in *Candida parapsilosis* Outbreaks by Combining MALDI-TOF Mass Spectrometry and Deep Learning Approaches", **Noshine Mohammad**, Anne-Cécile Normand, Cécile Nabet, Alexandre Godmer, Jean-Yves Brossas, Marion Blaize, Christine Bonnal, Arnaud Fekkar, Sébastien Imbert, Xavier Tannier, Renaud Piarroux, 2023, Microorganisms.

Deux études sont associées à ce projet de recherche et ont été publiées ou sont en cours de révision pour une future publication (Voir la section 3.8 pour plus de détails).

3.1 Contexte

Candida parapsilosis est l'une des levures les plus fréquemment responsables d'infections humaines. Certaines études la placent en deuxième position, juste derrière le *Candida albicans*, parmi les espèces les plus fréquemment responsables de candidémies (Tadec et al., 2016). Cette levure a également été impliquée dans des infections nosocomiales survenant sur un mode épidémique (Weems et al., 1992). Plus récemment, des épidémies d'infections à *C. parapsilosis* dues à des isolats résistants au fluconazole et à d'autres azolés, qui constituent la première ligne de traitement, ont été décrites dans plusieurs pays (Choi et al., 2018 ; Govender et al., 2016 ; Pinhati et al., 2016 ; Thomaz et al., 2018) sans que l'on puisse expliquer cette émergence soudaine responsable de taux de mortalité élevés dans les unités de soins intensifs où sont pris en charge des patients sont immunodéprimés (Pfaller et al., 2008 ; Raghuram et al., 2012).
Des publications récentes rapportent jusqu'à 30 % d'isolats résistants au fluconazole avec la présence d'une mutation A395T (substitution Y132F) dans le gène *erg11* expliquant le phénotype observé. Cette mutation est probablement le principal mécanisme qui confère à ces isolats une résistance aux azolés. Les investigations en cours nous amènent à penser que la circulation d'un clone résistant s'étend au-delà de notre hôpital en région parisienne.
En effet, en 2021, notre équipe (Fekkar et al., 2021) a décrit une épidémie de *Candida parapsilosis* résistant au fluconazole à l'hôpital de la Pitié Salpêtrière (PSL) à Paris. Deux clones infectant principalement des patients en réanimation ont été identifiés ; l'un a été identifié entre 2012 et 2017, et l'autre a émergé en 2017 et est malheureusement toujours actif. La propagation inquiétante de ces clones épidémiques résistants rend nécessaire la construction d'outils de diagnostic appropriés pour détecter les isolats résistants clonaux parmi tous les *C. parapsilosis* non clonaux sensibles au fluconazole identifiés dans le flux de routine de nos services de microbiologie.

Cependant, pour l'instant, l'affectation d'un isolat donné à un clone épidémique nécessite l'utilisation de méthodes moléculaires telles que le typage des microsatellites ou le séquençage de l'ADN. Ces méthodes sont considérées comme l'étalon-or (Diab-Elschahawi et al., 2012 ; Sabino et al., 2010) pour déterminer si deux isolats appartiennent au même clone. Le principal inconvénient des méthodes moléculaires est qu'elles sont trop

coûteuses et prennent trop de temps pour être mises en œuvre en tant qu'activité de routine.
Nous avons donc cherché une méthode qui permettrait d'identifier les clones directement dans le flux des analyses de routine sans avoir à mettre en œuvre des tests biologiques supplémentaires basés sur la biologie moléculaire. La détection d'un cluster épidémiologique de micro-organismes résistants aux médicaments au moyen des méthodes d'analyse de routine offrirait aux microbiologistes la possibilité d'informer promptement les cliniciens. Cette réactivité permettrait une adaptation rapide du traitement administré au patient, contribuant ainsi à améliorer la prise en charge des infections. Actuellement, la spectrométrie de masse à temps de vol par désorption/ionisation laser assistée par matrice (SM MALDI-TOF) représente la principale approche de routine pour identifier les bactéries et les levures dans presque tous les laboratoires de microbiologie du monde.

Des études récentes ont ouvert la voie à de nouvelles applications des approches de typage MALDI-TOF grâce à l'utilisation d'algorithmes d'apprentissage automatique. Delavy et al. (2020) ont sélectionné un modèle d'apprentissage automatique parmi de nombreux autres modèles testés afin de détecter qualitativement la résistance au fluconazole chez l'espèce tolérante aux azoles *C. albicans* (Delavy et al., 2020). Plus récemment, nous avons développé un modèle simple d'apprentissage profond pour identifier une population clonale d'*Aspergillus flavus* par spectrométrie de masse MALDI-TOF avec des performances élevées (Normand et al., 2022).

Malheureusement, contrairement aux exemples cités précédemment, nous avons rapidement découvert que dans le cas de *Candida parapsilosis*, les profils protéiques obtenus par spectrométrie de masse MALDI-TOF étaient tellement similaires qu'il était impossible d'obtenir une bonne discrimination entre les isolats appartenant au clone résistant et les autres en utilisant le modèle précédemment mis au point pour identifier les clones d'*Aspergillus flavus*.

3.2 Objectif

La possibilité de détecter un groupe épidémiologique de micro-organismes résistants aux médicaments directement par l'analyse des spectres de spectromètre de masse MALDI-TOF permettrait aux microbiologistes d'alerter les cliniciens, qui seraient en mesure d'adapter rapidement le traitement administré au patient et, par conséquent, d'améliorer leur pronostic.

Dans cette étude, nous avons examiné les méthodes utilisées lors de la préparation des échantillons et lors de l'analyse informatique des spectres de masse pour améliorer la phase d'apprentissage et, par conséquent, le pouvoir de discrimination du réseau de neurone entraîné. Cette étude s'est particulièrement concentrée sur les étapes expérimentales qui peuvent influencer les performances d'identification de clones épidémiques en utilisant l'apprentissage profond appliqué aux spectres MALDI-TOF. Ce travail constitue le premier effort d'analyse des conditions requises pour l'utilisation optimale du MALDI-TOF et de l'apprentissage profond dans l'investigation des épidémies en mycologie médicale. Il peut être utile à d'autres équipes qui éprouvent des difficultés à distinguer avec succès des entités microbiennes ayant des profils MALDI-TOF très similaires.

Nous étudions la capacité d'un modèle de réseau de neurones à convolution à identifier des clones fongiques particuliers responsables de l'épidémie d'infections fongiques à *C. parapsilosis* survenue dans nos hôpitaux depuis 2012. Le but de cette étude est de faire une comparaison au sein d'une même espèce. Sa particularité est que les spectres des clones et des non-clones sont très similaires tant au niveau de la présence des pics que des variations d'intensités, surtout lorsque les spectres sont acquis par la même machine (Voir Figure 3.1 ci-dessous). Cette similitude rend le problème de la classification beaucoup plus complexe et nous a conduit à étudier les effets des principaux paramètres de la manipulation expérimentale.

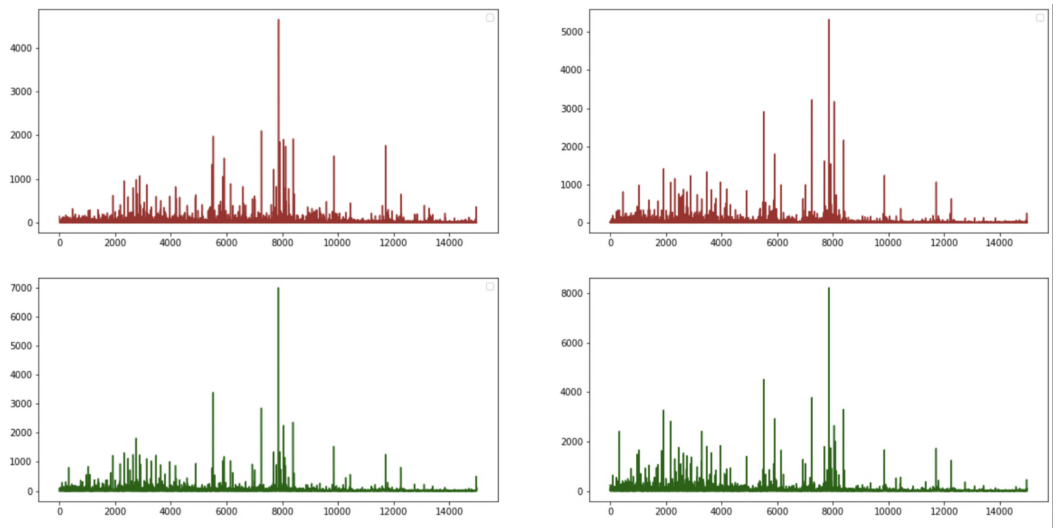


FIGURE 3.1 – **Visualisation de quatre spectres pré-traités de Clones et de Non-Clones de *Candida parapsilosis*.** Le rouge et le vert correspondent respectivement aux spectres des isolats appartenant ou non à l'ensemble clonal. Les isolats ont été cultivés dans les mêmes conditions et les spectres ont été acquis dans la même machine.

3.3 Acquisition et pré-traitement des données

3.3.1 Isolats

Quatre-vingt-seize isolats sensibles ou résistants au fluconazole ont été sélectionnés pour cette étude (Tableau supplémentaire Annexe A Table S2). Certains des isolats utilisés dans la présente étude ont été décrits précédemment (Fekkar et al., 2021 ; Presente et al., 2023). Parmi les isolats résistants, 39 appartiennent à l'ensemble clonal qui s'est récemment répandu dans différentes unités de soins intensifs de deux hôpitaux situés à Paris (l'hôpital de la Pitié-Salpêtrière (PSL) et l'hôpital Bichat Claude Bernard (BCH)). Les autres isolats ont été sélectionnés dans l'activité quotidienne de trois hôpitaux (PSL et BCH à Paris et l'hôpital Pellegrin à Bordeaux). Tous les isolats de *Candida parapsilosis* ont été cultivés en parallèle sur trois types de milieux de culture (Sabouraud Chloramphénicol Gentamycine (SAB-CG ; Oxoid, France), Chromagar (CHR ; BD, France) et Blood Agar (BLOOD ; BioMérieux, France)).

3.3.2 Diversité génétique et sensibilité au fluconazole

Le génotypage par microsatellite a été réalisé comme décrit par Diab-Elschahawi et al. (2012). Brièvement, un panel de 6 courtes répétitions en tandem a été utilisé, ce qui a permis d'obtenir un profil microsatellite de 12 marqueurs pour chaque isolat. Les profils microsatellites obtenus ont ensuite été exportés et soumis à une méthode de regroupement par paires non pondérées avec moyenne arithmétique (UPGMA) (Dendro-UPGMA [1] afin de générer un dendrogramme, en considérant les données comme des valeurs catégorielles. Les isolats présentant ≥ 11 génotypes identiques par typage microsatellite ont été regroupés et considérés comme appartenant au même ensemble clonal.

Les concentrations minimales inhibitrices de fluconazole ont été déterminées par une méthode de bande de concentration en gradient (Etest ; bioMérieux). Les isolats ont été classés comme sensibles, intermédiaires ou résistants selon les points de rupture cliniques EUCAST [2].

1. disponible à l'adresse http://genomes.urv.es/UPGMA/
2. disponible sur le site http://www.eucast.org/astoffungi/clinicalbreakpointsforantifungals/

3.3.3 Acquisition des spectres de masse MALDI-TOF

Toutes les cultures de *C. parapsilosis* ont été soumises au protocole d'extraction MALDI-TOF tel que décrit précédemment dans Normand et al. (2019) après 24 et/ou 48 heures de croissance. Brièvement, *C. parapsilosis* a été inactivé dans une solution d'EtOH à 70 %, et les protéines ont été extraites en utilisant de l'acide formique et de l'acétonitrile (v/v). Un microlitre d'extrait protéique a été déposé sur un point de deux cibles en acier poli (deux dépôts par isolat et par milieu de culture) et recouvert d'un microlitre de matrice HCCA. Dans chaque expérience, des échantillons de l'ensemble des clones et de l'autre catégorie ont été déposés alternativement afin d'éviter que tous les spectres clonaux se trouvent dans une moitié de la cible et tous les spectres non clonaux dans l'autre moitié. Les spectres ont été acquis à l'aide d'appareils Microflex situés dans quatre laboratoires parisiens différents : le service de mycologie (MYCO-PSL) et le service de bactériologie (BACT-PSL) de l'hôpital de la Pitié Salpêtrière, le service de bactériologie de l'hôpital Bichat Claude Bernard (BICHAT) et le service de bactériologie de l'hôpital Saint-Antoine (SAINT-ANTOINE). Pour les quatre spectromètres de masse, la méthode d'acquisition par défaut (MBT-AutoX) a été sélectionnée. Un paramètre d'acquisition des spectres par défaut du logiciel Flex Control a été modifié de la manière suivante : lorsqu'aucun des 800 tirs ne conduisait à un spectre répondant aux exigences du fabricant, la somme des spectres rejetés était sauvegardée au lieu de sélectionner l'option par défaut (i.e., ne pas sauvegarder), ce qui permettait une acquisition systématique d'un spectre, quelle que soit l'usure de la machine. L'identification de chaque dépôt a été vérifiée à l'aide de la base de données MSI-2 [3].

3.3.4 Analyse des données de spectrométrie de masse MALDI-TOF

Pour le pré-traitement, les données brutes MALDI-TOF ont été prétraitées à l'aide de l'environnement python 3.8 avec un lissage par la méthode de la moyenne mobile avec un facteur de 1/9, la méthode des moindres carrés asymétriques (Eilers et al., 2005) pour la correction de la ligne de base et la sélection des pics avec détection des changements de signe dans la dérivée des spectres (He et al., 2011) (Figure 3.2).

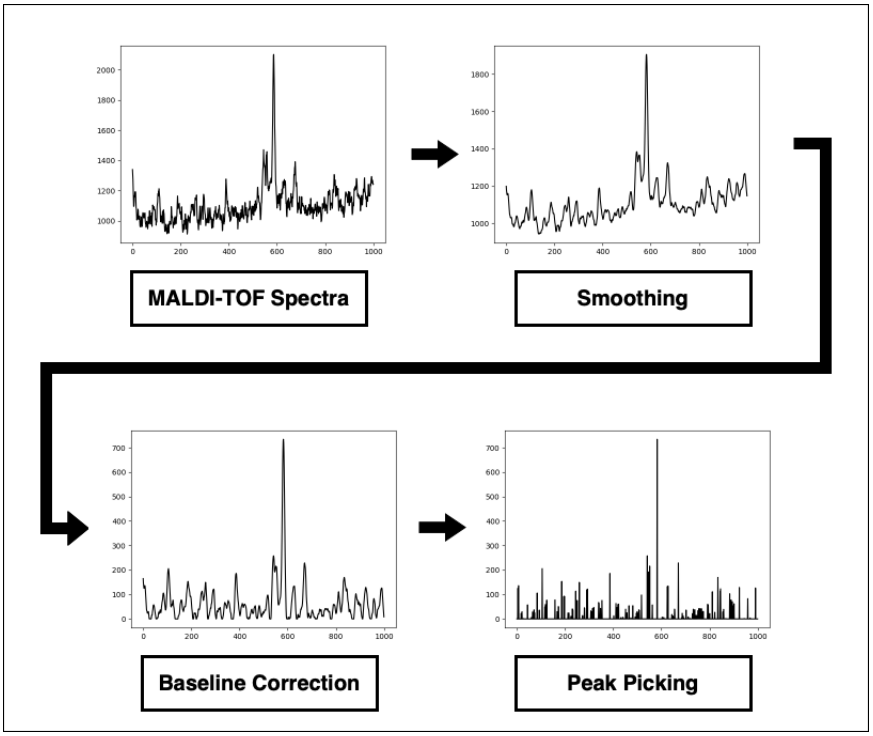


FIGURE 3.2 – **Prétraitement étape par étape des spectres, du spectre brut aux spectres traités, avant leur utilisation dans la phase d'apprentissage automatique.**

3. disponible à l'adresse https://msi.happy-dev.fr/

3.3.5 Alignement

L'alignement des spectres a été réalisé après l'étape de pré-traitement à l'aide de MSIWarp, une librairie Python fourni avec une implémentation C++. MSIWarp est un outil flexible compatible avec de nombreux types d'instruments pour réaliser l'alignement des spectres obtenus par spectrométrie de masse (Eriksson et al., 2020). L'approche d'alignement fonctionne sur les données à temps de vol (TOF) et réduit le décalage de la gamme de masse en appliquant une fonction de recalibrage sur les données de masse (m/z) et en maximisant un score de similarité qui prend en compte à la fois l'intensité et la position m/z des pics appariés entre deux spectres. Elle peut être appliquée à l'aide d'un spectre de référence. Ici, le spectre de référence choisi est celui qui présente le coefficient de corrélation le plus élevé avec tous les autres spectres.

3.4 Apprentissage automatique profond, évaluation et conception de l'étude

3.4.1 Apprentissage automatique

Pour mieux différencier les spectres du clone de ceux des autres souches, une méthode d'apprentissage profond impliquant un réseau de neurones à convolution (CNN : Convolutional Neural Network) a été mise en œuvre avec TensorFlow 2.7.0. Il est composé d'une partie convolutive et d'une partie entièrement connectée. Le classificateur (Figure 3.3) est un modèle CNN (Gu et al., 2017) très simple prenant un spectre de 18 000 valeurs en entrée. Le bloc convolutif est utilisé pour aider à la détection des motifs. Il est composé de plusieurs couches (3 filtres et une taille de noyau de 6) : une couche convolutive pour extraire les caractéristiques, une couche de mise en commun maximale (MaxPooling layer) pour réduire et transmettre l'information principale (Nirthika et al., 2022) (taille de la mise en commun = 100), et une couche d'aplatissement (Flatten layer) suivie de deux couches entièrement connectées (512 et 1024 unités). Une fonction d'unité linéaire rectifiée (ReLU : Rectified Linear Unit) (Agarap et al., 2019) est utilisée dans les couches convolutives et entièrement connectées comme fonction d'activation. La classification est ensuite effectuée avec une couche de normalisation (Ba et al., 2016) pour améliorer le score de classification et une couche dense finale de dimension 2, comportant une fonction softmax (Liu et al., 2017) pour produire la probabilité de prédiction sur les deux classes de sortie (clones et autres). Le taux d'apprentissage (Learning rate) est fixé par défaut à 0.001, et le nombre maximal d'époques est fixé à 50 avec un arrêt anticipé avec patience=20 (Early Stopping). Nous avons utilisé l'optimiseur Adam et la perte d'entropie croisée catégorielle (Categorical Cross-Entropy) (Zhang et al., 2018). La taille du lot (Batch size) est fixée à 60.

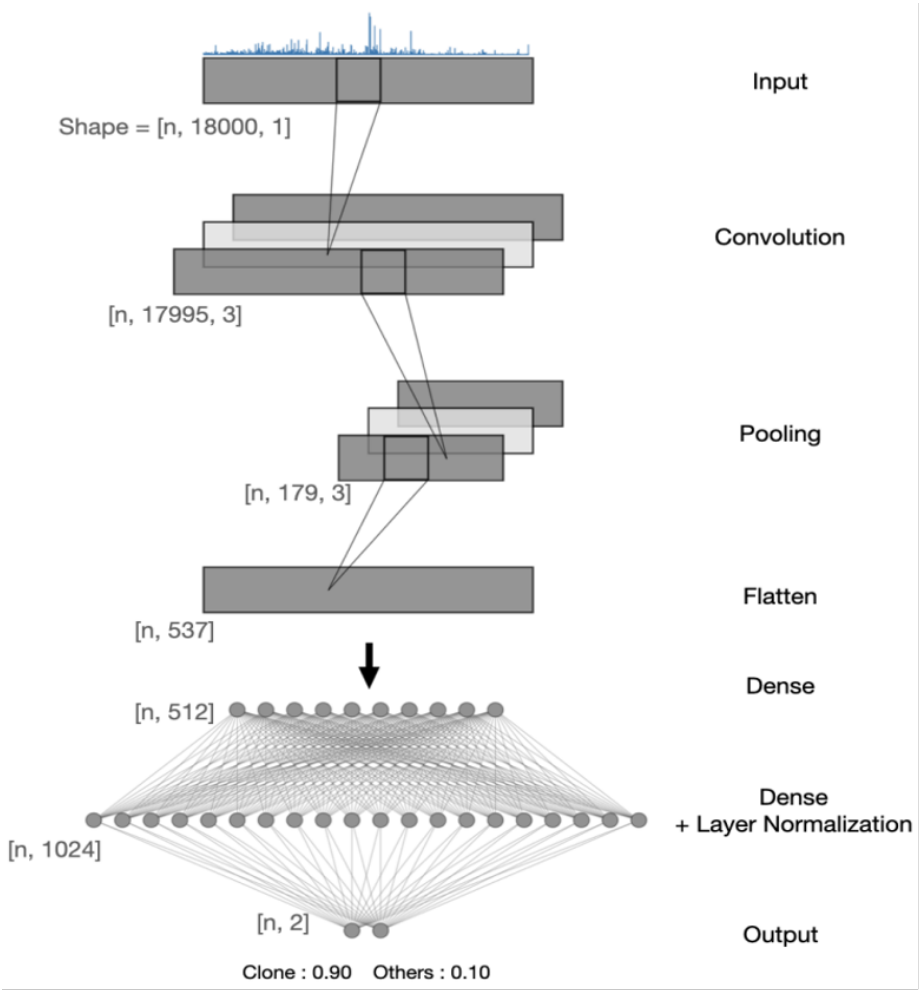


FIGURE 3.3 – **Architecture du modèle CNN créé et entraîné avec un ensemble de données.** La forme de la couche de sortie a été ajoutée avec une taille de lot n fixée à 1 pour simplifier l'illustration.

3.4.2 Méthode et mesures d'évaluation

Nous avons estimé les performances de notre système de classification en utilisant une technique de validation croisée imbriquée (nested CV : nested Cross Validation) stratifiée sur la classification clone/autre. Pour les différentes expérimentations les spectres ont été filtrés en fonction des différentes conditions pré-analytiques :

- le spectromètres de masse,
- le milieu de culture,
- et l'âge de la culture.

Pour chaque expérimentation, tous les spectres ont été divisés en 5 ensembles de taille égale selon les souches, en utilisant une sélection aléatoire préservant la distribution clone/autres. Chaque pli de CV était constitué :

(i) d'un ensemble d'entraînement composé de 80 % des isolats,

(ii) d'un ensemble de test composé des 20 % d'isolats restants.

Pour chaque pli, le système de classification clone/autre a été entraîné sur l'ensemble d'entraînement et validé sur l'ensemble de test.

Dans les expérimentations sollicitant les quatre instruments MALDI-TOF, au total, $5 \times 4 = 20$ plis (4 car quatre instruments MALDI-TOF) sont effectués avec une séparation stricte entre l'ensemble d'entraînement et l'ensemble de test, à la fois en termes d'isolats et de spectromètres. Il est à noter que chaque isolat est associé à plusieurs spectres.

Pour chaque évaluation d'impact, pour chaque entraînement par CV nous avons effectué une moyenne des performances obtenues et nous avons utilisé l'accuracy (pourcentage d'identifications correctes), le score F1, qui est un score de synthèse utilisé dans l'apprentissage automatique, le rappel (sensibilité) et la spécificité (Voir la section 2.5 pour les définitions des métriques). Les intervalles de confiance à 95 % ont été calculés à l'aide de la méthode bootstrap empirique (Dekking et al., 2005).

3.4.3 Conception de l'étude

L'étude a été conçue en quatre étapes (Figure 3.4). Tout d'abord, en utilisant tous les spectres acquis après 24 heures de croissance sur les trois milieux de culture, nous avons comparé l'effet des machines. Nous avons utilisé les spectres obtenus avec trois des quatre machines pour la phase d'apprentissage, et nous avons testé la classification "clone/autre" du CNN sur les spectres obtenus avec la quatrième machine. Deuxièmement, nous avons répété les mêmes tests, mais avec une différence : nous avons préalablement appliqué la méthode d'alignement MSIWarp avant les phases d'entraînement. Troisièmement, pour tester l'effet du milieu de culture, nous avons détaillé les résultats en fonction des milieux de culture utilisés pour la croissance des isolats. Enfin, en utilisant deux des quatre machines, nous avons nouvellement acquis les spectres des 96 isolats à 24 et 48 heures de croissance pour évaluer l'impact de l'âge de la culture.

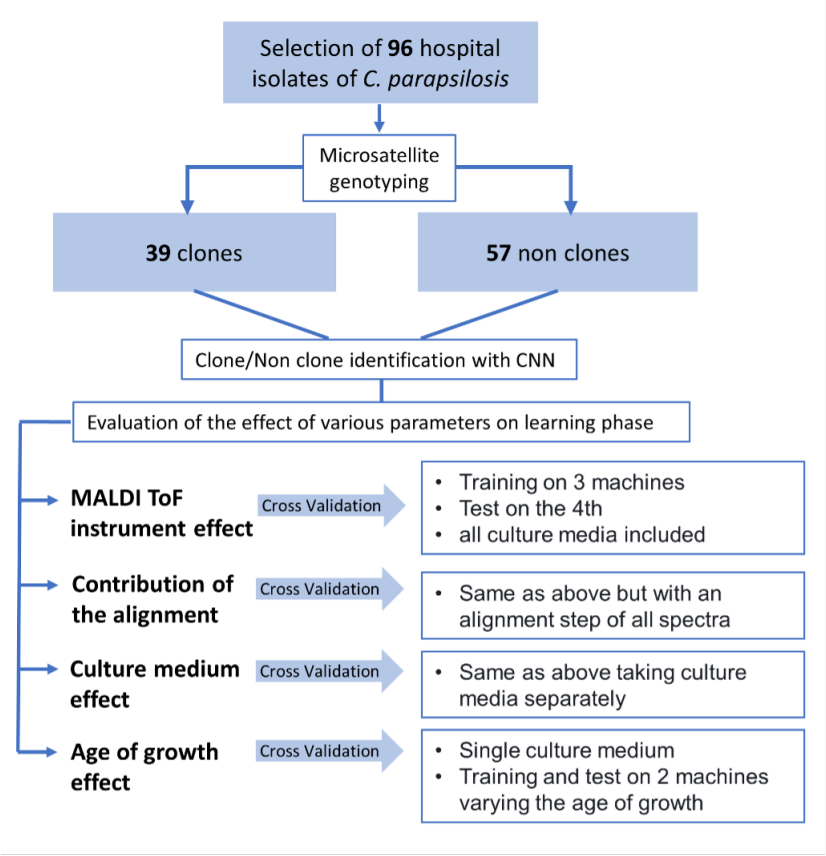


FIGURE 3.4 – **Organigramme de la conception de l'étude.**

Considérations éthiques : cette étude a été réalisée conformément à la Déclaration d'Helsinki. La présente étude n'a pas été considérée comme une étude impliquant des êtres humains selon la loi française n° 2012-300, car aucune donnée clinique ou d'identification n'a été utilisée. Toutes les souches ont été conservées de manière anonyme dans le laboratoire de mycologie de l'hôpital de la Pitié Salpêtrière.

3.5 Résultats

Au total, 2258 spectres ont été acquis et utilisés pour déterminer l'impact de la machine, de l'alignement et du milieu de culture, et 768 nouveaux spectres ont été acquis pour l'évaluation de l'impact de l'âge de la culture.

3.5.1 Diversité génétique

Parmi les 96 isolats sélectionnés, 39 étaient étroitement liés et appartenaient à notre ensemble de clones que nous avons appelé R2 (Figure 3.5 ; Tableau en Annexe A Table S2). Trente-sept des 39 isolats correspondaient à deux clones répandus qui ne différaient que par 7 répétitions sur l'un des allèles (26 isolats avec le profil R2a (3A : 28-28 ; 3B : 49-82 ; 3C : 48-51 ; 6A : 8-8 ; 6B : 7-7 ; 6C : 7-7) et 11 isolats avec le profil R2b (3A : 28-28 ;

3B : 49-75 ; 3C : 48-51 ; 6A : 8-8 ; 6B : 7-7 ; 6C : 7-7)). Deux autres isolats semblaient proches de ces deux clones et présentaient des variations du microsatellite 3B : l'isolat 307 (3A : 28-28 ; 3B : 49-76 ; 3C : 48-51 ; 6A : 8-8 ; 6B : 7-7 ; 6C : 7-7) et l'isolat 329 (3A : 28-28 ; 3B : 46-82 ; 3C : 48-51 ; 6A : 8-8 ; 6B : 7-7 ; 6C : 7-7).

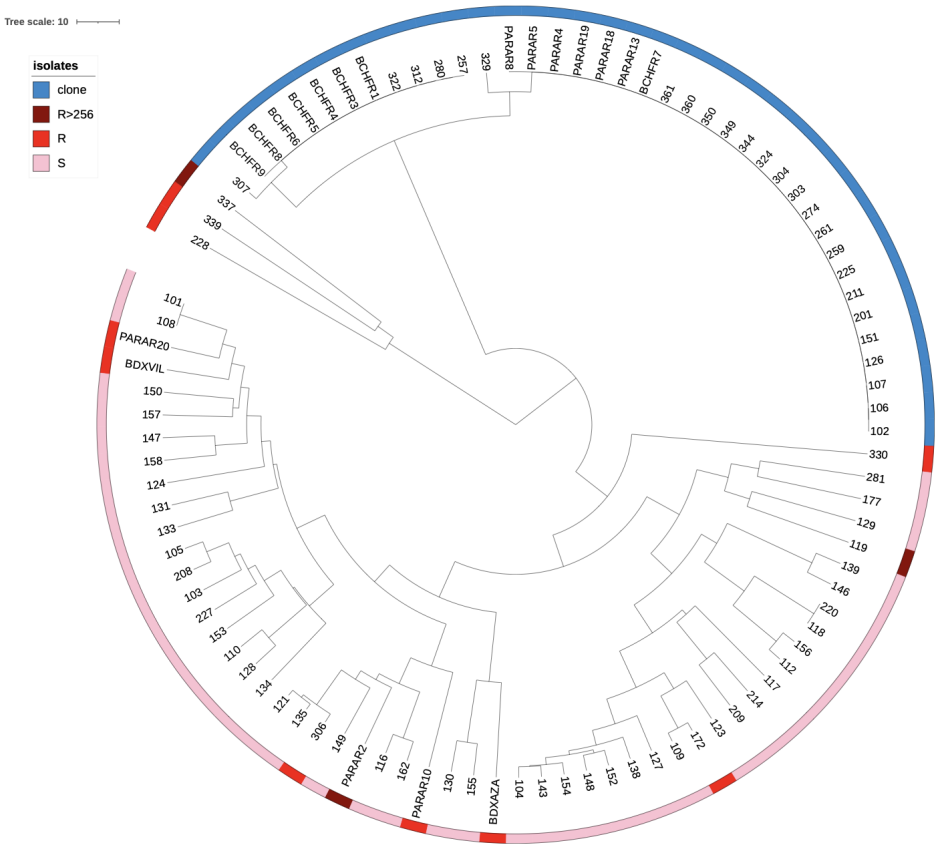


FIGURE 3.5 – **Dendrogramme des 96 isolats de *Candida parapsilosis* sélectionnés pour cette étude et typés par une approche microsatellite.**

3.5.2 Sensibilité au fluconazole

Parmi nos 96 isolats, 51 ont montré une résistance au fluconazole (concentration minimale inhibitrice (CMI) $\geq$ 4 mg/litre), et 45 isolats étaient sensibles ou intermédiaires au fluconazole (CMI $<$ 4 mg/litre) (Tableau supplémentaire Annexe A Table S2). Tous les isolats appartenant à l'ensemble de clones R2 étaient résistants au fluconazole, et 38/39 présentaient une résistance $\geq$ 256 mg/litre. Seul l'isolat R2 329 présentait une CMI de 16 mg/litre s. La plupart des isolats du profil R2a (25/26) provenaient de l'hôpital de La Pitié-Salpêtrière de novembre 2017 à octobre 2020, tandis qu'un isolat de ce clone a été identifié à l'hôpital Bichat-Claude Bernard en juin 2021. Des isolats du profil R2b ont été détectés à partir de février 2020 à La Pitié-Salpêtrière et de mai 2021 à Bichat-Claude Bernard. La propagation de *C. parapsilosis* des deux profils est toujours en cours dans ces deux hôpitaux.

3.5.3 Impact de la machine et de l'alignement avec MSIWarp

Les résultats de la CV réalisé pour chacune des quatre machines sont compilés dans le Tableau 3.1. Sans alignement, les performances du modèle CNN ont montré des résultats allant de 68 % à 89 % pour l'accuracy moyenne en fonction de la machine testée. Cette expérience montre que toutes les machines ne sont pas égales, même lorsque les mêmes paramètres d'acquisition ont été utilisés. L'alignement de MSIWarp a permis d'améliorer significativement les performances du modèle CNN, en particulier pour les deux machines qui présentaient des performances plus faibles sans alignement. En particulier, le rappel (sensibilité) des machines BACT-PSL et BICHAT est passé de 0,30 à 0,64 et de 0,29 à 0,75, respectivement, sans perte de spécificité dans les deux cas.

TABLEAU 3.1 – **Impact de la machine et de l'alignement avec MSIWarp.** Performance de l'identification des isolats appartenant à l'ensemble des clones par le modèle CNN (validation croisée sur 5 plis). Moyenne des ensembles d'entraînement de 1355 spectres obtenus sur trois machines et moyenne des ensembles de test de 113 spectres obtenus sur la quatrième machine.

	Accuracy	**Score F1**	**Rappel**	**Spécificité**
Machine Testée ; Performances Sans Alignement				
MYCO-PSL	0.89	0.87	0.90	0.86
	[0.87,0.92]	[0.84,0.90]	[0.86,0.94]	[0.83,0.90]
BACT-PSL	0.70	0.44	0.30	0.95
	[0.66,0.73]	[0.37,0.51]	[0.24,0.37]	[0.93,0.97]
SAINT-ANTOINE	0.83	0.74	0.65	0.90
	[0.80,0.86]	[0.71,0.80]	[0.59,0.71]	[0.87,0.93]
BICHAT	0.68	0.42	0.29	0.88
	[0.65,0.72]	[0.36,0.50]	[0.24,0.35]	[0.84,0.91]
Machine Testée ; Performances Avec Alignement				
MYCO-PSL	0.91	0.89	0.92	0.88
	[0.88,0.93]	[0.86,0.92]	[0.88,0.95]	[0.84,0.91]
BACT-PSL	0.81	0.74	0.64	0.92
	[0.78,0.85]	[0.68,0.78]	[0.57,0.70]	[0.89,0.95]
SAINT-ANTOINE	0.91	0.89	0.92	0.87
	[0.89,0.93]	[0.86,0.92]	[0.88,0.95]	[0.83,0.90]
BICHAT	0.84	0.79	0.75	0.87
	[0.81,0.87]	[0.75,0.83]	[0.68,0.80]	[0.84,0.91]

3.5.4 Impact du milieu de culture par machine

En conservant l'alignement avec MSIWarp, nous avons comparé les résultats obtenus sur les trois milieux de culture par machine (Tableau 3.2). En fonction du milieu de culture et de la machine testée, les performances du modèle CNN varient de 0,77 % à 0,96 % en accuracy moyenne. À l'exception de la machine MYCO-PSL, pour laquelle le rappel (sensibilité) était équivalente sur les trois milieux de culture, des performances supérieures ont été obtenues sur Sabouraud-GC pour une spécificité équivalente.

TABLEAU 3.2 – **Impact du milieu de culture par machine.** Performance de l'identification des isolats appartenant à l'ensemble des clones par le modèle CNN (validation croisée sur 5 plis). Moyenne des ensembles d'entraînement de 452 spectres obtenus sur trois machines et moyenne des ensembles de test de 38 spectres obtenus sur la quatrième machine.

	Accuracy	**Score F1**	**Rappel**	**Spécificité**
Machine Testée Avec Alignement ; Performances sur Chromagar				
MYCO-PSL	0.91	0.89	0.88	0.83
	[0.86,0.94]	[0.83,0.94]	[0.80,0.95]	[0.75,0.90]
BACT-PSL	0.77	0.62	0.47	0.96
	[0.71,0.83]	[0.55,0.75]	[0.37,0.59]	[0.92,0.99]
SAINT-ANTOINE	0.88	0.84	0.80	0.90
	[0.83,0.92]	[0.78,0.90]	[0.71,0.88]	[0.84,0.95]
BICHAT	0.81	0.72	0.61	0.96
	[0.75,0.86]	[0.64,0.82]	[0.50,0.71]	[0.91,0.99]
Machine Testée Avec Alignement ; Performances sur Sabouraud-CG				
MYCO-PSL	0.88	0.86	0.84	0.89
	[0.84,0.93]	[0.80,0.92]	[0.75,0.91]	[0.83,0.94]
BACT-PSL	0.93	0.92	0.95	0.88
	[0.89,0.96]	[0.87,0.96]	[0.89,0.99]	[0.82,0.94]
SAINT-ANTOINE	0.89	0.89	0.95	0.84
	[0.84,0.93]	[0.83,0.93]	[0.89,0.99]	[0.76,0.90]
BICHAT	0.89	0.88	0.89	0.92
	[0.85,0.94]	[0.82,0.92]	[0.81,0.95]	[0.87,0.97]
Machine Testée Avec Alignement ; Performances sur gélose Sang				
MYCO-PSL	0.92	0.89	0.87	0.95
	[0.88,0.96]	[0.86,0.96]	[0.79,0.94]	[0.90,0.98]
BACT-PSL	0.86	0.81	0.73	0.97
	[0.81,0.90]	[0.74,0.88]	[0.63,0.83]	[0.94,1.00]
SAINT-ANTOINE	0.93	0.92	0.96	0.87
	[0.89,0.96]	[0.87,0.96]	[0.91,1.00]	[0.81,0.93]
BICHAT	0.96	0.95	0.93	0.87
	[0.94,0.99]	[0.91,0.99]	[0.87,0.99]	[0.81,0.93]

3.5.5 Impact de l'âge de la culture sur milieu de Sabouraud

En conservant l'alignement avec MSIWarp, nous avons comparé les performances obtenues après 24 h et 48 h de croissance sur deux machines (MYCO-PSL et SAINT-ANTOINE). Les spectres des deux machines ont été regroupés et la CV n'a été réalisé que sur l'âge de la culture (Tableau 3.3). En considérant les mêmes âges de culture pour l'entraînement et le test, les performances se sont avérées égales, quelle que soit la métrique prise en compte (>90 %). Lorsque les âges de la culture sont croisés, notamment lorsque le CNN est entraîné avec des spectres issus de cultures cultivées pendant 48 h et testé avec des spectres issus de cultures cultivées pendant 24 h, les performances sont désastreuses, tous les spectres étant identifiés comme des non-clones.

TABLEAU 3.3 – **Impact de l'âge de la culture.** Performance de l'identification des isolats appartenant à l'ensemble des clones par le modèle CNN (validation croisée sur 5 plis). L=ensemble d'entraînement ; T=ensemble de test. Au total, 307 spectres ont été utilisés par âge de la culture pour l'ensemble d'entraînement, tandis que 77 spectres ont été utilisés par âge de la culture pour l'ensemble de test.

	Accuracy	**Score F1**	**Rappel**	**Spécificité**
Âge de la culture testé avec l'alignement ; performances par âge de la culture				
L24 h/T24 h	0.92	0.91	0.91	0.95
	[0.90,0.95]	[0.87,0.94]	[0.86,0.94]	[0.92,0.98]
L48 h/T48 h	0.94	0.92	0.93	0.95
	[0.91,0.96]	[0.89,0.95]	[0.89,0.97]	[0.92,0.97]
L(24 h+48 h)/T(24+48 h)	0.93	0.91	0.91	0.96
	[0.91,0.94]	[0.88,0.93]	[0.87,0.94]	[0.94,0.97]
Âge de la culture testé avec l'alignement ; performance des âges mixtes de la culture				
L24 h/T48 h	0.70	0.71	0.90	0.56
	[0.65,0.74]	[0.65,0.76]	[0.85,0.94]	[0.50,0.63]
L48 h/T24 h	0.59	0.00	0.00	1.00
	[0.54,0.64]	[0.00,0.00]	[0.00,0.00]	[1.00,1.00]
L(24 h+48 h)/T24 h	0.92	0.90	0.87	0.96
	[0.90,0.95]	[0.86,0.94]	[0.82,0.92]	[0.93,0.98]
L(24 h+48 h)/T48 h	0.91	0.89	0.89	0.93
	[0.89,0.94]	[0.86,0.93]	[0.84,0.94]	[0.90,0.96]

3.6 Discussion

3.6.1 Points forts et points faibles de l'étude

Jusqu'à présent, aucune étude ne s'est intéressée aux étapes pré-analytiques de la classification des spectres à l'aide d'un réseau de neurones. Nous montrons ici que ces étapes sont importantes en mettant en évidence le rôle des milieux de culture, du temps de croissance, de la machine utilisée pour acquérir les spectres et, enfin, du traitement mathématique appliqué au spectre, en particulier son alignement avec un spectre de référence avant sa classification par le réseau de neurones.

Le résultat peut être excellent, médiocre ou désastreux selon la maîtrise de ces paramètres. Ainsi, un apprentissage réalisé sur deux machines (MYCO-PSL et SAINT ANTOINE) à partir de colonies cultivées 48 heures sur gélose Sabouraud-GC a permis de classer correctement 94 % des spectres acquis dans les mêmes conditions, alors qu'une tentative de classification de spectres acquis après 24 heures de croissance à l'aide du même réseau de neurones entraîné a conduit à des résultats désastreux (tous les spectres ont été classés comme des non-clones). Nos résultats montrent également que cet écueil peut être contourné en incluant les deux temps de culture dans le processus d'apprentissage, ce qui permet d'obtenir une classification satisfaisante des isolats après 24 et 48 heures de culture. L'impact de l'âge de la culture sur la forme du spectre a déjà été observé dans des études visant à évaluer les performances d'identification en microbiologie médicale, notamment pour les dermatophytes (Jabet et al., 2022 ; Normand et al., 2022). Dans certains cas, cela a conduit à l'inclusion de spectres acquis à différents âges de la culture dans les bases de données de référence afin d'améliorer les performances d'identification. Dans le cas particulier de la recherche de clones au sein d'une espèce de levure, le degré de précision rend indispensable la maîtrise de ce paramètre.

Au-delà du temps de croissance de la colonie, notre étude a montré l'importance du milieu de culture sur lequel les colonies sont cultivées pour obtenir les résultats les plus fiables. Ce n'est pas une surprise pour nous, car ce paramètre a souvent été mis en avant dans les études, même si celles-ci concluaient que l'impact d'une telle variation sur la fiabilité de l'identification n'était pas un obstacle. Dans le cas de la recherche de clones, le niveau de précision est tel qu'il est nécessaire de prendre en compte ce paramètre. Notre étude montre que la classification des clones est possible soit en étendant l'apprentissage à plusieurs milieux de culture, soit en limitant l'utilisation du modèle aux spectres obtenus à partir d'isolats cultivés sur le même milieu que celui

utilisé pour l'apprentissage.

Les mêmes conclusions peuvent être tirées concernant les machines utilisées pour la phase d'apprentissage et pour les tests. Dans une étude précédente sur la détection clonale d'*Aspergillus flavus*, nous avions mis en évidence un effet machine pour les phases d'apprentissage et de test. Nous avons également souligné les difficultés à obtenir des résultats satisfaisants avec l'une des machines testées (BACT-PSL) qui était surutilisée (Normand et al., 2022). Néanmoins, nous montrons ici qu'en augmentant le nombre de spectres et de machines utilisés dans la phase d'apprentissage, il a été possible d'obtenir une classification satisfaisante des spectres pour une nouvelle machine non utilisée pour l'apprentissage, avec 81 à 91 % (d'accuracy) de spectres correctement classés, selon la machine utilisée pour tester le modèle. Cependant, pour obtenir ces résultats, la classification par réseau de neurone devrait être précédée d'une étape d'alignement des spectres afin de minimiser la variabilité des spectres d'une machine à l'autre. Heureusement, cette étape peut être réalisée automatiquement et ne prend qu'une fraction de seconde pour chaque nouveau spectre testé sur le modèle entraîné. De manière tout à fait inattendue, nous avons pu observer que notre CNN pouvait très facilement identifier la machine sur laquelle les spectres avaient été acquis et le milieu de culture sur lequel la colonie avait été cultivée.

Dans l'ensemble, ces résultats montrent qu'il est possible d'utiliser des réseaux de neurones pour réaliser des études épidémiologiques au niveau local ou même sur plusieurs centres, à condition de contrôler certains paramètres. Sur la base des recherches effectuées dans cette étude, nous recommandons à tout centre recherchant des clones spécifiques dans le contexte de la propagation locale d'une épidémie d'effectuer la phase d'apprentissage à l'aide de spectres acquis localement, puis de tester le modèle ultérieur en utilisant le même spectromètre de masse MALDI-TOF. En outre, les conditions, c'est-à-dire le milieu et la durée de culture, dans lesquelles les colonies ont été obtenues doivent être identiques entre la phase d'apprentissage et la phase de test. Dans le cas où les spectres à tester devraient correspondre à différentes conditions d'acquisition (par exemple, utilisation de plusieurs milieux de culture ou de plusieurs spectromètres de masse), nous recommandons de prendre en compte ces conditions dans la phase d'apprentissage. L'impact important de paramètres tels que le milieu de culture ou le temps de croissance a également été observé avec la spectrométrie infrarouge et le typage bactérien (Quintelas et al., 2018), pour lesquels il est recommandé d'effectuer tous les échantillons à typer dans la même expérience. Nous montrons ici qu'il est possible d'obtenir des résultats satisfaisants lorsque l'apprentissage et le test ne sont pas effectués en même temps ou sur la même machine. Il s'agit là d'une découverte intéressante qui mérite d'être soulignée. Un autre avantage notable est qu'une technologie couramment utilisée dans les laboratoires biomédicaux a servi de point de départ, ce qui n'était pas le cas de l'étude sur la spectrométrie infrarouge.

Notre étude présente toutefois des limites. Tout d'abord, le nombre d'isolats testés (96, dont 39 correspondants à une épidémie) est faible. Cela a certainement limité les capacités d'apprentissage de notre réseau de neurone, car il est bien connu que plus il y a d'éléments inclus dans la phase d'apprentissage, meilleurs sont les résultats. Cependant, les épidémies survenant en milieu hospitalier impliquent généralement un nombre limité de cas, en particulier celles impliquant des agents fongiques ; il est donc nécessaire de développer des approches adaptées pour faciliter les enquêtes épidémiologiques dès la découverte de l'épidémie et lorsque le nombre de cas est encore faible. Ainsi, une épidémie impliquant 39 cas dans deux hôpitaux différents est déjà un problème, d'où la nécessité de mettre en place de bons outils de détection.

3.6.2 Observation générale

En microbiologie, notamment pour la détection de la résistance aux antimicrobiens, l'utilisation des réseaux de neurones a fourni des informations intéressantes (Popa et al., 2022). Un certain nombre d'architectures différentes peuvent être utilisées, notamment les réseaux neuronaux convolutifs (CNN), qui sont connus pour être très puissants dans la reconnaissance d'images (Hung et al., 2020). Par exemple, ces algorithmes ont démontré leur utilité en microbiologie pour la lecture automatisée de la coloration de Gram (Smith et al., 2018).

Toutefois, contrairement à la reconnaissance d'images, les données expérimentales sur les spectres de masse MALDI-TOF restent rares. Bien que la spectrométrie de masse MALDI-TOF soit devenue la principale méthode utilisée pour l'identification de routine des bactéries, des levures et des champignons filamenteux, seules quelques études ont exploré les avantages des algorithmes d'apprentissage profond dans la classification des spectres MALDI-TOF. Cette observation peut être appliquée soit dans des études visant à distinguer des espèces étroitement apparentées, soit pour identifier une caractéristique particulière au sein d'une espèce microbienne, comme la résistance à certaines molécules antimicrobiennes ou l'appartenance à un clone épidémique.

3.6.3 Comparaison avec des études antérieures associant la technologie MALDI-TOF à l'apprentissage automatique

Dans la plupart des travaux de microbiologie combinant l'apprentissage automatique et le MALDI-TOF, la préparation des spectres est l'un des points forts de la méthodologie. Delavy et al. (2020), Yan et al. (2021) et Mortier et al. (2021) ont traité les données brutes avec le logiciel R Studio, ce qui leur a permis d'utiliser la librairie MALDIquant qui offre plusieurs techniques de traitement. Outre le traitement principal comme le lissage, la soustraction de la ligne de base et la technique de sélection douce des pics, la plupart des méthodes poussent le traitement des spectres jusqu'au recalibrage, à l'extraction des pics et à la normalisation. Nous nous sommes limités aux trois premières étapes (lissage, correction de la ligne de base et sélection des pics) afin de ne pas alourdir le traitement le spectre et de conserver autant d'informations que possible sur le spectre.
De plus, dans la plupart des travaux antérieurs, des algorithmes de contrôle qualité ont été utilisés pour améliorer les prédictions. Dans notre cas, le tri est limité à l'élimination des spectres vides ou ne présentant pas de pic significatif lors de l'acquisition sur le MALDI-TOF. Un pré-traitement simple et rapide a été choisi afin que tout le traitement des spectres soit effectué par le modèle. Compte tenu de la difficulté du problème, nous supposons que l'information relative au caractère clonal se trouve dans les pics de faible intensité et les petites variations d'intensité, ce qui justifie d'éviter la perte d'information par un pré-traitement drastique.

La représentation des spectres à l'entrée du modèle est également un point crucial de la méthodologie. L'identification est fortement dépendante de cette étape. Les articles dont la méthode nécessite l'utilisation de MALDIquant ont tendance à réaliser une étape d'extraction de pics dans le pré-traitement (Delavy et al., 2020), (Yan et al., 2021), pour l'extraction de caractéristiques car les modèles traditionnels d'apprentissage automatique sont les plus utilisés (Mortier et al., 2021).
Certaines recherches utilisent l'apprentissage profond pour filtrer les pics d'intérêt et effectuer l'étape d'extraction des caractéristiques juste avant la classification par un modèle d'apprentissage automatique (Li et al., 2022). Ling et al (2020) ont une approche différente tant au niveau de la préparation des spectres que de la représentation. Comme les spectres sont des données unidimensionnelles, les auteurs se sont penchés sur la question de la transformation du spectre en une image 2D qui est ensuite envoyée à un modèle CNN bi-dimensionnel. Dans notre cas, nous avons choisi de nous limiter à la représentation unidimensionnelle le plus simple possible après pré-traitement, sans extraction de pics, ni réduction de dimension, sous la forme d'un vecteur de taille 18000. De cette manière, nous sommes en mesure de réduire la complexité informatique tout en analysant les variations physiologiques, protéiques et protéomiques propres à chaque clone. Nous réservons la recherche d'une représentation plus complexe pour des travaux ultérieurs.

Contrairement à la plupart des études, nous avons travaillé sur une approche multicentrique.
Cependant, les performances d'identification des clones est comparable à celle de l'identification des résistances chez les bactéries (Weis et al., 2021) et chez les levures (Delavy et al., 2020). En effet, Weis et al. (2021) ont des performances allant jusqu'à 80 % en AUROC dans la détermination de la résistance aux antimicrobiens. De même, pour le *Candida albicans*, Delavy et al. (2020) ont des performances atteignant 86 % d'accuracy pour la détection de la résistance au fluconazole. Ces deux travaux ont utilisé des modèles classiques d'apprentissage automatique impliquant des étapes de sélection des caractéristiques dans le traitement des spectres (LDA, Régression logistique, Random Forest, Light GBM,...). Notre équipe, (Normand et al., 2022) a évalué un CNN simple (un bloc de convolution) avec une méthodologie d'étude incluant la phase d'entraînement sur un site et la phase de test sur un ensemble de données composé de spectres provenant de divers sites, à titre de preuve de concept. Comme dans cette étude, les performances varient d'un site à l'autre, mais les meilleurs résultats sont comparables aux nôtres en ce qui concerne les performances d'identification des clones sur un support. Néanmoins, nous n'avions pas encore employé la méthode d'alignement, démontrant ainsi son avantage en termes d'amélioration des performances d'identification.

3.6.4 Perspectives

Concrètement, cette étude ne visait pas à exploiter les possibilités offertes par diverses approches d'apprentissage automatique en dehors des réseaux neuronaux, notamment la machine à vecteurs de support (SVM), la régression logistique, les K plus proches voisins ou les arbres de décision. De même, d'autres architectures de réseaux neuronaux, comme les réseaux récurrents ou siamois, n'ont pas été envisagées. Il est important de souligner que les méthodes alternatives qui n'ont pas été examinées dans cette étude, représentent des domaines de recherche à considérer pour l'avenir. L'objectif principal de cette étude résidait davantage dans l'exploration des différentes étapes préalables à la phase d'apprentissage, un aspect souvent sous-estimé dans la littérature scientifique traitant de ce sujet.

3.7 Conclusion

L'optimisation de la préparation des spectres de masse MALDI-TOF avant la classification à l'aide de techniques d'apprentissage profond est un nouveau sujet émergent, et il reste encore beaucoup à explorer à ce sujet. Cependant, avec cette étude, nous démontrons qu'une telle optimisation peut améliorer les résultats de l'apprentissage profond et devrait éventuellement permettre de repousser les limites de la spectrométrie de masse MALDI-TOF. Cela pourrait ouvrir la voie à de nouvelles améliorations dans le diagnostic des épidémies fongiques et bactériennes en complément des méthodes moléculaires.

3.8 Études associées

3.8.1 Identification de clones d'*Aspergillus flavus* par MALDI-TOF et apprentissage profond

Nous avons réalisé une étude, preuve de concept, sur l'identification d'une population clonale d'*Aspergillus flavus* par spectrométrie de masse MALDI-TOF à l'aide d'un réseau de neurone à convolution simple que j'ai conçu.

En effet, la propagation des clones fongiques est difficile à détecter dans les routines quotidiennes des laboratoires cliniques, et il est nécessaire de disposer de nouveaux outils pouvant faciliter la détection des clones au sein d'un ensemble de souches. Nous avons réalisé une expérience dans laquelle 19 isolats clonaux d'*Aspergillus flavus* initialement collectés sur des masques chirurgicaux contaminés ont été inclus dans un ensemble de 55 isolats d'*A. flavus* d'origines diverses. Un simple réseau de neurone à convolution (CNN) a été entraîné pour détecter les isolats appartenant au clone. Dans cette expérience, les ensembles d'entraînement et de test étaient totalement indépendants et des appareils MALDI-TOF différents (Microflex) ont été utilisés pour les phases d'entraînement et de test. Le CNN a été utilisé pour trier correctement une grande partie des isolats, avec une excellente ($>$ 93 %) accuracy pour deux des trois appareils utilisés et avec une accuracy moindre pour le troisième appareil (69 %), qui était plus ancien et dont le laser devait être remplacé.

Article : "Identification of a clonal population of *Aspergillus flavus* by MALDI-TOF mass spectrometry using deep learning", Anne-Cécile Normand , Aurélien Chaline, **Noshine Mohammad**, Alexandre Godmer, Aniss Acherar, Antoine Huguenin, Stéphane Ranque, Xavier Tannier, Renaud Piarroux, 2022, Scientific Reports.

3.8.2 Transmission nosocomiale d'*Aspergillus flavus* en unité néonatale : persistance et détection via MALDI-TOF et CNN

Également, une étude sur la transmission nosocomiale d'*Aspergillus flavus* dans une unité de soins intensifs néonatale, a été faite, dans le but d'identifier des clones épidémiques dans une population d'isolats comportant des clones et des non-clones. J'ai utilisé le réseau de neurone construit dans l'étude précédente pour effectuer la classification.

L'aspergillose du nouveau-né reste une maladie rare mais grave. Nous rapportons quatre cas d'infections cutanées primaires à *A. flavus* chez des nouveau-nés prématurés liés à la contamination d'incubateurs par des souches clonales putatives. Notre objectif était d'évaluer la capacité du MALDI-TOF couplé à un CNN pour la reconnaissance de clones dans un contexte où seul un très petit nombre de souches est disponible pour l'apprentissage automatique. Des isolats cliniques et environnementaux d'*A. flavus* (n=64) ont été étudiés, dont 15 étaient épidémiologiquement liés aux quatre cas. Toutes les souches ont été typées à l'aide du polymorphisme de longueur des microsatellites. Nous avons trouvé un génotype commun pour 9/15 souches apparentées. Les isolats ont été sélectionnés pour obtenir un ensemble de données d'entraînement (6 isolats clonaux/25 non-clonaux) et un ensemble de données de test (3 isolats clonaux/31 non-clonaux), et les spectres ont été analysés à l'aide d'un modèle CNN simple.
Sur l'ensemble de données de test, les 31 isolats non clonaux ont été correctement classés, 2/3 isolats clonaux ont été correctement classés sans ambiguïté, tandis que la troisième souche est restée indéterminée. Les souches

clonales d'*A. flavus* persistent dans les unités de soins intensifs néonatals depuis plusieurs années. En effet, deux souches d'*A. flavus* isolées d'incubateurs en septembre 2007 sont identiques à la souche responsable du second cas survenu 3 ans plus tard.
Nous avons pu constater que le MALDI-TOF est un outil prometteur pour la détection d'isolats clonaux d'*A. flavus* à l'aide de CNN, même avec un ensemble d'entraînement limité, pour un coût et un temps de manipulation limités.

Article : "Nosocomial transmission of *Aspergillus flavus* in a neonatal intensive care unit : long term persistence in environment and interest of MALDI-ToF Mass-Spectrometry coupled with Convolutional Neural Network (CNN) for rapid clone recognition", **Noshine Mohammad***, Antoine Huguenin*, Annick Lefebvre, Laura Menvielle, Dominique Toubas, Stéphane Ranque, Isabelle Villena, Xavier Tannier, Anne-Cécile Normand, Renaud Piarroux, 2024, Medical Mycology - Oxford Academic.
* : Ces deux premiers auteurs ont contribué à parts égales à cet article.

Chapitre 4

Estimation précise de l'âge des moustiques anophèles par régression de réseau de neurones pour améliorer la surveillance épidémiologique de la transmission du paludisme

Ce chapitre est issu des recherches présentées dans l'article intitulé : "Accurate prediction of *Anopheles* mosquito age for epidemiological monitoring of malaria transmission", **Noshine Mohammad** et Pauline Naudion, Abdoulaye K. Dia, Pierre-Yves Boëlle, Abdoulaye Konaté, Lassana Konate, El-Hadji A. Niang, Renaud Piarroux, Xavier Tannier and Cécile Nabet, en cours de révision pour une future publication.

4.1 Contexte

Les maladies transmises par les moustiques sont en augmentation dans le monde entier [1]. Les parasites du paludisme, transmis par les moustiques anophèles, infectent encore plus de 200 millions de personnes et en tuent plu d'un demi-million chaque année [2]. Dans ce contexte, la lutte antivectorielle est une priorité stratégique pour prévenir la transmission de la maladie. Seuls les moustiques les plus âgés peuvent transmettre le paludisme au cours d'un repas sanguin infectieux (Macdonald, 1956), en raison de la période d'incubation nécessaire au développement de l'agent pathogène, qui est de 9 à 14 jours pour les parasites du paludisme (Gilles et al., 2002). À ce jour, nous avons peu d'informations sur la longévité des moustiques, mais seule une petite fraction d'entre eux peut survivre suffisamment longtemps (>10 jours) pour devenir infectieuse (Lambert et al., 2022). En ciblant la longévité des moustiques, les stratégies basées sur les insecticides ont été très efficaces (Johnson et al., 2020). Cependant, leur efficacité est menacée par la résistance aux insecticides (Ranson et al., 2016) et de nouvelles stratégies de réduction de la durée de vie, telles que les moustiques infectés par le virus *Wolbachia* ou génétiquement modifiés, ont été proposées (Shaw et al., 2018 ; Flores et al., 2018). Pour évaluer l'efficacité de ces nouvelles stratégies de lutte antivectorielle, il est essentiel d'examiner la structure d'âge des populations de moustiques ciblées et de mesurer l'impact sur la condition physique des moustiques (Johnson et al., 2020). Une détermination précise de l'âge des moustiques pourrait également permettre de mieux comprendre la biologie des vecteurs et l'impact de la longévité des moustiques sur la transmission du paludisme (Lambert et al., 2022). Il est donc nécessaire de disposer d'outils de terrain efficaces pour déterminer l'âge des moustiques.

Actuellement, seuls quelques outils entomologiques sont disponibles pour déterminer l'âge des moustiques, et la plupart d'entre eux demandent beaucoup de travail, sont approximatifs et pas toujours cohérents (Lambert et al., 2022 ; Johnson et al., 2020 ; Matthews et al., 2020). La méthode de marquage-relâchement-recapture (MRR) consiste à relâcher et à recapturer des moustiques marqués à différents moments (Guerra et al., 2014), mais elle nécessite un grand nombre de moustiques en raison des faibles taux de recapture, ce qui prend beaucoup

1. disponible à l'adresse https://apps.who.int/iris/bitstream/handle/10665/259205/9789241512978-eng.pdf
2. disponible à l'adresse https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022

de temps. La méthode morphologique consiste à observer au microscope les changements ovariens et à mesurer la proportion d'adultes physiologiquement plus âgés (qui ont déjà pondu), ce qui est approximatif (Detinova et al., 1962). Elle nécessite également des compétences techniques et du matériel de haute qualité, ce qui la rend difficile à utiliser sur le terrain. Comme alternative, de nouvelles techniques ont été développées, telles que le profilage transcriptionnel (Cook et al., 2007 ; Cook et al., 2006 ; Wang et al., 2013 ; Cook et al., 2010 ; Weeraratne et al., 2021), l'analyse des hydrocarbures cuticulaires (Caputo et al., 2005 ; Hugo et al., 2006) et la spectroscopie infrarouge (Siria et al., 2022 ; González Jiménez et al., 2019 ; Lambert et al., 2018 ; Sikulu et al., 2010 ; Mayagaya et al., 2009). Jusqu'à présent, ces méthodes n'ont pas supplanté les méthodes traditionnelles, car leur application sur le terrain s'avère compliquée (Johnson et al., 2020).

La détection des changements protéiques chez les moustiques est prometteuse pour la détermination précise de l'âge des moustiques. Des biomarqueurs protéiques du vieillissement, tels que la liaison au calcium, les chaperons moléculaires liés au stress et les protéines cuticulaires, ont été identifiés chez les moustiques *Anopheles* (Sikulu et al., 2015) et *Aedes* (Hugo et al., 2013) à l'aide de méthodes de SM haut de gamme. La SM à temps de vol par désorption et ionisation laser assistée par matrice (SM MALDI-TOF) a déjà révolutionné la microbiologie clinique pour l'identification rapide des espèces pathogènes (Evangelista et al., 2022 ; Elbehiry et al., 2022). Les spectres de masse des protéines de moustiques peuvent être générés en quelques minutes à partir d'un extrait protéique d'une partie anatomique disséquée. Notre équipe a récemment montré la faisabilité de cette méthode couplée à l'apprentissage profond pour prédire l'âge des moustiques dans une preuve de concept utilisant des souches d'anophèles de laboratoire (Nabet et al., 2020). Cependant, la robustesse de la méthode aux variations environnementales chez les moustiques sauvages est encore inconnue. En outre, l'estimation de l'âge a été limitée à de grandes catégories telles que les moustiques jeunes (0 à 3 jours), intermédiaires (4 à 10 jours) et âgés (11 à 28 jours).

4.2 Objectif

Notre équipe a mené la première étude pour explorer l'applicabilité sur le terrain de la SM par MALDI-TOF couplée à l'apprentissage profond (MALDI-TOF-DL) pour prédire l'âge des moustiques anophèles sauvages adultes, à partir de spectre de leur tête, de leur thorax ou de leurs pattes. Ici, nous souhaitons démontrer la robustesse de notre approche en utilisant des ensembles de données de moustiques indépendants obtenus à partir de deux sites écologiques au Sénégal, et en nous concentrant sur les moustiques *An. arabiensis*, une espèce majeure de vecteur du paludisme du complexe *Gambiae* en Afrique (Sinka et al., 2012). Également, nous tenterons d'augmenter le niveau de granularité de la prédiction de l'âge des moustiques en utilisant des modèles d'apprentissage profond optimisés, atteignant une précision suffisante pour surveiller la structure de l'âge des populations de moustiques *anophèles* sauvages. Cette méthode innovante pourrait fournir un outil d'estimation de l'âge des moustiques simple et précis pour surveiller la transmission des maladies et évaluer l'efficacité des interventions ciblées de lutte antivectorielle, qui peut être adapté à d'autres espèces de moustiques vecteurs pour améliorer la surveillance des maladies transmises par les moustiques.

4.3 Collecte et acquisition des données

4.3.1 Collecte sur le terrain et élevage de moustiques

Pour constituer deux ensembles de données indépendants présentant des variations écologiques, des moustiques anophèles ont été collecté au stade larvaire dans deux endroits différents du Sénégal. L'ensemble de données **Sénégal 1** (n=183) a été collecté dans la zone urbaine de Dakar (N14°46'11.532" / W17°18'25.091", district sanitaire de Keur Massar) en décembre 2021, et l'ensemble de données **Sénégal 2** (n=68) a été collecté dans la zone rurale de Keur Socé (N 14°00'01.3" / W 16°03'07.5", district sanitaire de N'Dofan) en novembre 2019. Pour mieux s'adapter au milieu naturel, les larves collectées sur le terrain ont été élevées dans un insectarium à température et humidité ambiantes. Les moustiques adultes ont eu accès ad libitum à une solution de saccharose à 10 %. Les moustiques femelles de la génération F0 ont été sacrifiés à un moment fixe par congélation (30 minutes à -20°C) à différents âges post-émergence et jusqu'à 28 jours d'âge, le jour 0 étant défini comme le jour d'émergence. Les spécimens ont ensuite été stockés à +4°C (délai maximum de 3 semaines) et transportés plus tard au laboratoire de parasitologie-mycologie de l'hôpital de la Pitié-Salpêtrière, Paris (France), avec l'approbation nécessaire de l'IRD (Institut de Recherche pour le Développement) sénégalais et en suivant l'accord de transfert de matériel (Material Transfer Agreement, MTA). Dès leur arrivée, les échantillons ont été congelés et conservés à -20°C jusqu'à leur analyse. L'identification morphologique a confirmé leur classification comme membres du complexe *Gambiae*, et le marqueur IGS a été utilisé pour les identifier comme *An. arabiensis*,

suivant un protocole précédemment publié (Wilkins et al., 2006).

4.3.2 Préparation des échantillons pour la SM MALDI-TOF et acquisition des spectres de masse

Après la dissection du moustique, l'extraction des protéines a été réalisée à partir de la tête, du thorax avec les ailes et des pattes selon un protocole précédemment décrit (Nabet et al., 2020). Ensuite, les extraits de protéines ont été déposés sur une plaque d'acier et recouverts d'une matrice HCCA. Pour garantir la précision et la fiabilité, chaque spécimen et partie anatomique a été soumis à quatre répétitions techniques.

Les spectres de masse ont été acquis à l'aide d'un instrument Microflex LT (Bruker France SAS) avec les paramètres d'acquisition par défaut, comme cela a été fait dans le protocole précédent (Nabet et al., 2020). Les données acquises ont ensuite été exportées vers l'application Maldi Biotyper v4.1 pour évaluer la reproductibilité des réplicats de spectre et la qualité des spectres de masse. Pour créer la base de données, un profil de spectre de référence (Main Spectrum Profile : MSP) a été généré pour chaque spécimen et chaque partie anatomique en utilisant le logiciel Bruker MBT Compass. Ces profils sont établis en calculant la moyenne des spectres issus de quatre réplicas techniques réalisés à partir du même échantillon. Les MSP tiennent compte de l'intensité, de la position et de la fréquence d'apparition des pics.
Avant d'intégrer les spectres dans la base de données, un contrôle de qualité a été effectué pour chaque MSP afin d'éliminer tout spectre de qualité insuffisante. Deux méthodes d'analyse de la reproductibilité ont été appliquées à cet effet. La première méthode consiste à évaluer la reproductibilité des résultats entre les quatre dépôts d'un même spécimen en comparant le MSP de référence à chacun des dépôts dont il est issu, en utilisant un algorithme de MALDI Biotyper v4.1 de Bruker France SAS qui a produit des valeurs logarithmiques (Log Score Value : LSV) comprises entre 0 et 3. Un LSV plus élevé indique une meilleure reproductibilité des spectres. Une valeur seuil de 2 a été fixée, en-dessous de laquelle le dépôt était considéré comme insuffisamment reproductible, entraînant le retrait des spectres correspondants de la base de données. La deuxième méthode consiste à générer un coefficient de corrélation (CC) par un autre algorithme du même logiciel, qui varie de 0 à 1 en fonction du degré de corrélation entre les spectres, en prenant en compte la distribution, l'intensité et la fréquence des pics. Le CC est représenté dans une matrice de chaleur (heatmap) avec des couleurs variant du bleu au rouge pour indiquer le niveau de corrélation, se rapprochant de 1 en cas de forte corrélation. Tout MSP ayant un CC <0.8 par rapport à lui-même a été retiré De la même manière, si un MSP avait un index de corrélation $< 0{,}8$ vis-à-vis des MSPs des autres spécimens de la même espèce d'anophèle, il était aussi exclu (défaut de reproductibilité inter-spécimen). Pour minimiser les biais d'acquisition potentiels, les spectres de toutes les classes d'âge ont été acquis simultanément et placés sur la même plaque.

Pour minimiser les biais d'acquisition potentiels, les spectres de toutes les classes d'âge ont été acquis simultanément et placés sur la même plaque. Pour une analyse robuste des variations spectrales, les ensembles de données du **Sénégal 1** et du **Sénégal 2** ont été acquis indépendamment par différents opérateurs utilisant le même instrument. Les moustiques et la composition des spectres des ensembles de données **Sénégal 1** et **Sénégal 2** sont fournis dans les données supplémentaires Annexe B Data S1 et Data S2.

4.3.3 Profilage protéique et préparation des données pour l'analyse de l'apprentissage profond

Pour observer la possibilité de retrouver des variations protéiques en lien avec l'âge des moustiques, nous avons comparé les profils des différentes catégories d'âges d'anophèles en important les spectres de masse dans le logiciel ClinProTools 3.0, en suivant le protocole précédent (Nabet et al., 2020). Cette analyse nous a permis de générer une liste de pics présentant des différences d'intensité significatives. Compte tenu de la gamme de masse, le pouvoir de résolution de masse a été estimé à 5-10 Da. Nous avons analysé les spectres en utilisant l'outil Peak Statistic avec le mode de tri t-test/ANOVA. Nous avons sélectionné les 25 pics les plus discriminants (valeur $p < 0{,}05$) pour une analyse et une interprétation plus approfondies (Données supplémentaires Annexe B Data S3 et Data S4).

Plusieurs techniques ont été utilisées pour le pré-traitement des spectres de masse : lissage à l'aide de la méthode de la moyenne mobile, correction de la ligne de base à l'aide de la méthode des moindres carrés asymétriques (Eilers et al., 2005) et sélection des pics à l'aide de la méthode de la dérivée (He et al., 2011). Ces étapes avaient pour but de minimiser le bruit tout en préservant autant que possible les signaux de faible intensité. L'aligne-

ment spectral a été réalisé à l'aide de l'algorithme MSIWarp (Eriksson et al., 2020). L'ensemble des données de spectres de moustiques contenait 2 763 spectres (tête, thorax et pattes) et a été divisé en un ensemble d'entraînement (80 %) et deux ensembles de test indépendants (20 %) pour le **Sénégal 1** et le **Sénégal 2**, qui n'ont pas été vus auparavant par le modèle. Selon l'expérience, chaque partie anatomique a été utilisée indépendamment ou ensemble pour l'entraînement et le test. Chaque spectre a été traité comme une entrée unique, mais pour éviter l'ajustement excessif, nous avons stratifié l'échantillonnage par spécimen afin d'affecter les réplicats d'un même spécimen de moustique soit à l'ensemble d'entraînement soit à l'ensemble de test.

Une autre question importante était de développer des modèles robustes aux variations génétiques et techniques afin de valider notre approche de prédiction de l'âge pour une utilisation sur le terrain. Pour y répondre, la diversité génétique de notre échantillonnage a été augmenté en collectant des larves d'*An. arabiensis* dans deux sites écologiques du Sénégal (Figure 4.2 A), situés à 205 km l'un de l'autre. Le premier site, **Sénégal 1** (n=183), était une zone urbaine avec des habitats larvaires temporaires à semi-permanents, et les larves ont été collectées au début de la saison sèche (décembre 2021). Le deuxième site, **Sénégal 2** (n=68), était une zone rurale avec des habitats larvaires temporaires, et les larves ont été collectées à la fin de la saison des pluies (novembre 2019). Les larves ont ensuite été élevé jusqu'au stade adulte d'âge connu dans un insectarium sous température et humidité non contrôlées.

Nous avons également augmenté la variabilité technique en traitant indépendamment les moustiques de chaque site d'étude avec le même instrument, mais par des opérateurs différents et à des moments différents, avec un intervalle de 2 ans, créant ainsi deux ensembles de données indépendants pour le Sénégal 1 et le Sénégal 2. L'analyse non supervisée (analyse en composantes principales) des spectres de masse MALDI-TOF de la tête, du thorax et des pattes n'a révélé aucun regroupement susceptible de séparer les ensembles de données du **Sénégal 1** et du **Sénégal 2**. Cela indique un certain degré d'homogénéité entre les spectres de masse des deux sites, malgré la variabilité potentielle (Figure 4.2 B). Par conséquent, un ensemble de données variable a été créé, de 2 763 spectres provenant de **Sénégal 1** et de **Sénégal 2**. Nous en avons utilisé 80 % pour l'entraînement du modèle et 20 % pour les tests sur le **Sénégal 1** ou le **Sénégal 2**.

4.4 Méthodes d'apprentissage profond et évaluation

Notre étude suit les recommandations de la liste de contrôle "Minimum Information about the Clinical Artificial Intelligence Modeling" (MI-CLAIM) (Norgeot et al., 2020) pour la modélisation par apprentissage automatique profond (Voir le Tableau supplémentaire Annexe B Table S3).

4.4.1 Modèles d'apprentissage profond

L'un des principaux défis de cette étude était de développer des modèles d'apprentissage profond (Figure 4.1) pour détecter les variations temporelles dans les profils protéiques des spectres de masse des moustiques, afin d'améliorer la granularité de la prédiction de l'âge des moustiques. Pour la validation initiale de notre approche MALDI-TOF-DL sur des moustiques collectés sur le terrain, nous avons d'abord utilisé un réseau de neurones à convolution (CNN) et une technique conventionnelle de prédiction de la classification, comme cela avait été fait précédemment pour classer les moustiques par catégories d'âge (Siria et al., 2022, Nabet et al., 2020). Nous avons ensuite exploré de nouvelles techniques de prédiction par apprentissage profond : nous avons testé la méthode de classification cohérente avec les rangs (ranking) car cette approche s'est avérée plus précise que la classification conventionnelle pour prédire l'âge des individus en utilisant la reconnaissance des images faciales (Shi et al., 2023), car elle utilise l'ordre chronologique de l'âge pour classer les données et les assigner à des catégories d'âge. Ensuite, nous avons testé la méthode de régression car elle pouvait fournir une estimation de l'âge du moustique en jours et améliorer la granularité. Pour améliorer encore nos prédictions d'âge, nous avons évalué un réseau de neurones à convolution temporel (TCN) qui a été développé à l'origine pour traiter les signaux audio (Lemaire et al., 2019). Le TCN et le CNN sont tous deux efficaces pour analyser les données séquentielles, mais le TCN peut être particulièrement performant pour capturer les dépendances temporelles dans les données de séries temporelles (Lea et al., 2016).

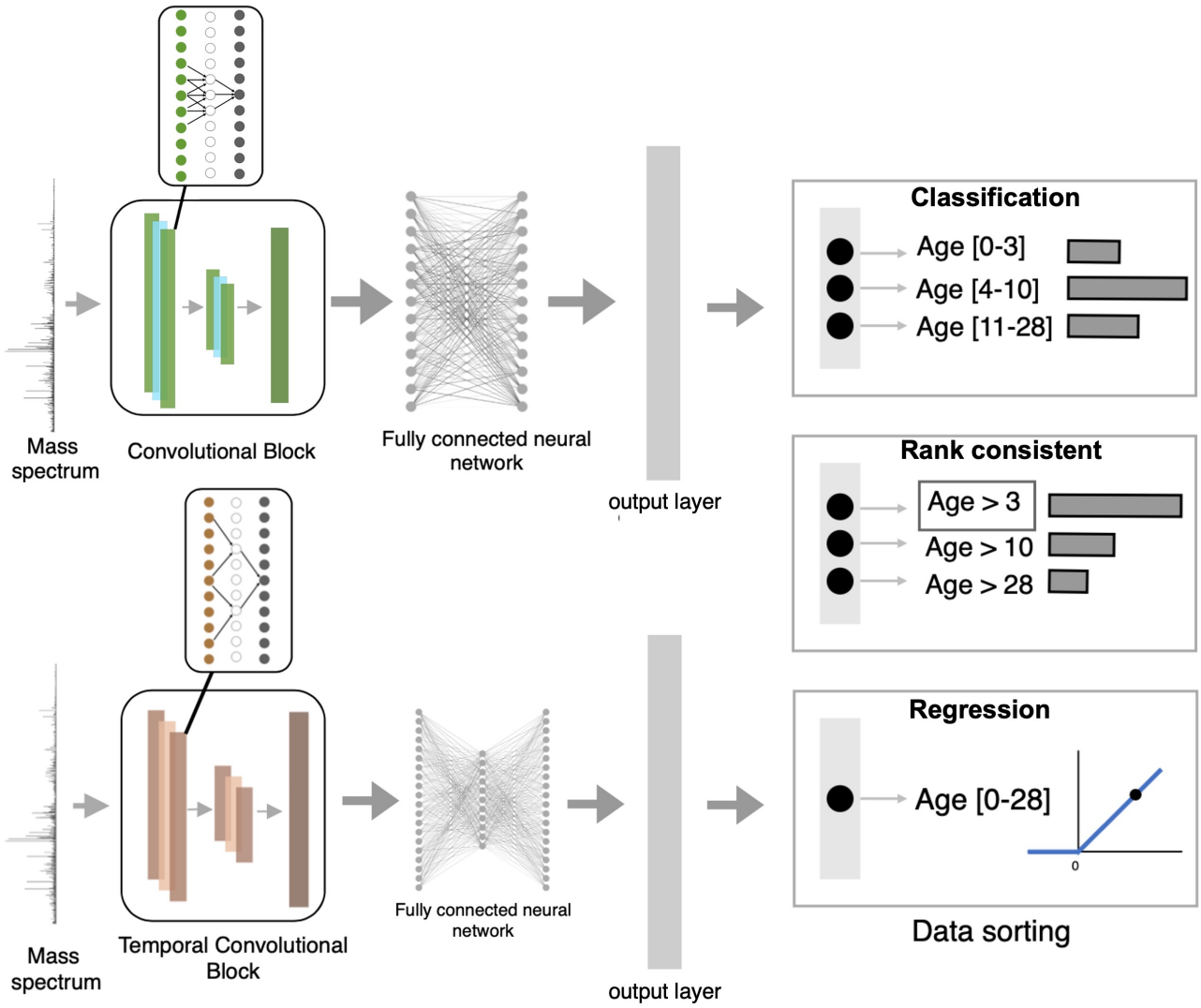


FIGURE 4.1 – **Architectures des modèles d'apprentissage profond pour prédire l'âge des moustiques anophèles à partir des spectres de masse MALDI-TOF.** Les architectures du CNN (bleu) et du TCN (marron) diffèrent en termes de filtres convolutifs. Le CNN utilise des filtres qui glissent linéairement sur les données d'entrée, en se concentrant sur les éléments proches, tandis que le TCN utilise des filtres à pas variable, permettant la détection de dépendances entre des éléments non adjacents. Le TCN est particulièrement adapté à la capture de dépendances temporelles complexes.

Les deux réseaux de neurones artificiels ont été mis en œuvre à l'aide de TensorFlow 2.7.0. L'entrée des réseaux consistait en un tableau unidimensionnel de 18 000 valeurs.

Réseau de neurones à convolution (CNN) : Le modèle comprend un seul bloc convolutif. Il se compose d'une couche convolutive (Conv1D) avec 3 filtres et une taille de noyau de 6, suivie d'une couche de mise en commun maximale (MaxPooling) avec une taille de mise en commun de 100. La sortie de la couche de regroupement maximal est ensuite aplatie (Flatten) et connectée à un réseau de neurones composé de deux couches entièrement connectées (Dense) de 512 unités chacune. Cette partie du réseau est connectée à une couche de normalisation (LayerNormalization) et à une couche dense finale. La fonction Rectified Linear Unit (ReLU) (Agarap et al., 2018) sert de fonction d'activation pour les couches convolutives et entièrement connectées. Le taux d'apprentissage (Learning rate) est fixé, par défaut, à 0,001 et le nombre maximal d'époques est de 50. L'optimisation est réalisée à l'aide de l'optimiseur Adam. Pour éviter le sur-apprentissage, nous utilisons une taille de lot (Batch size) de 60 et un arrêt précoce (EarlyStopping) avec une patience de 20.

Réseau de neurones à convolution temporelle (TCN : Temporal convolutional Network) : Le modèle s'inspire d'une étude publiée (Lea et al., 2016). Il comprend des couches de régularisation initiales, comprenant une couche de normalisation des lots (Batch Normalization layer) et une couche d'abandon spatial 1D (1D Spatial Dropout layer) avec un taux d'abandon de 0,3. Ces couches sont connectées à un bloc convolutif temporel,

qui consiste en une couche TCN non causale avec 3 filtres, une taille de noyau de 10, un empilage (stack), et une dilatation de 10. Un taux d'abandon (dropout rate) de 0,3 est appliqué à la couche, des connexions de saut (skip connections) sont établies, et un retour à la séquence (return sequences) est configuré. Ce bloc est suivi d'une couche de mise en commun maximale (MaxPooling) avec une taille de mise en commun de 50 et se termine par une couche aplatie (Flatten).
La sortie du TCN non causal est acheminée à travers trois couches entièrement connectées, avec 512, 128, et 512 unités respectivement, pour former un réseau de neurones entièrement connecté. La fonction PReLU (Parametric Rectified Linear Unit) (Xu et al., 2015) est utilisée comme fonction d'activation dans les couches entièrement connectées, et la matrice de poids du noyau est initialisée à l'aide de l'initialisateur He normal. Une pénalité de régularisation L2 de 10^{-4} est également employée comme régularisateur. La partie entièrement connectée est connectée à une couche de normalisation (LayerNormalization) et à une couche dense finale pour la prédiction, similaire au modèle CNN. Les hyperparamètres d'apprentissage sont définis de manière identique à ceux du CNN, à l'exception de la taille du lot (Batch size), qui est fixée à 32.

4.4.2 Méthodes d'apprentissage profond pour la prédiction de l'âge

Trois approches prédictives ont été évaluées dans le but de déterminer la méthode optimale pour prédire l'âge.

Classification conventionnelle : Pour cette méthode, la couche de sortie a été configurée pour générer trois sorties à l'aide d'une fonction softmax. Nous avons utilisé la perte d'entropie croisée catégorielle (Categorical Cross Entropy) pour générer la probabilité de prédiction dans les trois différentes classes d'âge de sortie (0-3 jours, 4-10 jours, 11-28 jours).
Nous avons établi les trois catégories d'âge selon une étude récente de la spectroscopie infrarouge moyenne (MIRS) (Siria et al., 2022) pour l'estimation de l'âge des moustiques : 0-3 jours ; 4-10 jours ; et $\geq$ 11 jours, représentant respectivement les moustiques improbablement infectés , les moustiques possiblement infectés, les moustiques improbablement infectieux et les moustiques possiblement infectieux.

Classification cohérente en termes de rangs (Rank-consistent classification or ranking) : Contrairement aux méthodes de classification conventionnelles, la classification cohérente avec les rangs est une approche de classification ordinale spécifiquement conçue pour les réseaux de neurones profonds qui prennent en compte l'ordre de la variable cible (chronologie de l'âge). Nous avons utilisé un cadre de classification ordinale cohérente avec les rangs appelé CORN (Conditional Ordinal Regression for Neural Networks) (Shi et al., 2023). Dans ce cadre, les groupes d'âge (0-3 jours, 4-10 jours, 11-28 jours) ont été traités comme des catégories ordonnées, et le modèle a été entraîné à prédire la probabilité qu'un individu tombe dans chaque catégorie. La sortie de chaque probabilité de groupe d'âge représente la probabilité que l'entrée appartienne à ce groupe d'âge ou à un groupe d'âge supérieur dans l'ordre.

Régression : Pour estimer directement l'âge du moustique en jours, nous avons configuré la couche de sortie pour qu'elle fournisse une sortie numérique unique représentant l'âge prédit. Nous avons utilisé la fonction ReLU (Rectified Linear Unit) et la fonction de perte de Huber (Huber loss) pour l'apprentissage.

4.4.3 Performances et mesures de prédiction

Nous avons d'abord procédé à une validation croisée (cross validation) sur l'ensemble d'apprentissage pour construire nos modèles. Après avoir construit les modèles, nous avons évalué les performances de prédiction sur chaque ensemble de données de test indépendant. Pour les classifications conventionnelles et cohérente en termes de rangs, nous avons utilisé les mesures d'accuracy, de précision (valeur prédictive positive VPP) et de rappel (sensibilité).

Nous avons mesuré l'erreur absolue moyenne (MAE : Mean Absolute Error) et le R carré pour la régression.
Pour comparer les performances de toutes les méthodes de prédiction, nous avons utilisé l'aire sous la courbe caractéristique de l'opérateur récepteur (AUROC). Nous avons légèrement adapté cette mesure pour évaluer l'approche de régression. L'AUROC reflète la capacité de diagnostic d'une méthode en comparant le taux de vrais positifs (sensibilité ou rappel) au taux de faux positifs (1-spécificité, c'est-à-dire (1-VN/(VN+FP)). D'autre part, la mesure de la précision reflète le taux de prédiction correcte. Il est important de noter que ces mesures ne sont pas directement comparables et que l'AUROC est plus approprié pour comparer les performances (Voir la section 2.5 pour les définitions des métriques).

Nous avons calculé des intervalles de confiance à 95 % à l'aide de la méthode bootstrap empirique (Dekking et al., 2005).

Pour comparer les résultats des trois techniques de prédiction, nous avons utilisé un test ANOVA à sens unique, en considérant chaque partie anatomique indépendamment. Pour les résultats de la régression, nous avons arrondi les prédictions pour qu'elles correspondent à trois classes catégorielles : 0-3 jours, 4-10 jours et 11-28 jours. Cela nous a permis d'effectuer le test statistique. Nous avons comparé les prédictions de régression obtenues par CNN et TCN à l'aide d'un test de Student bilatéral.

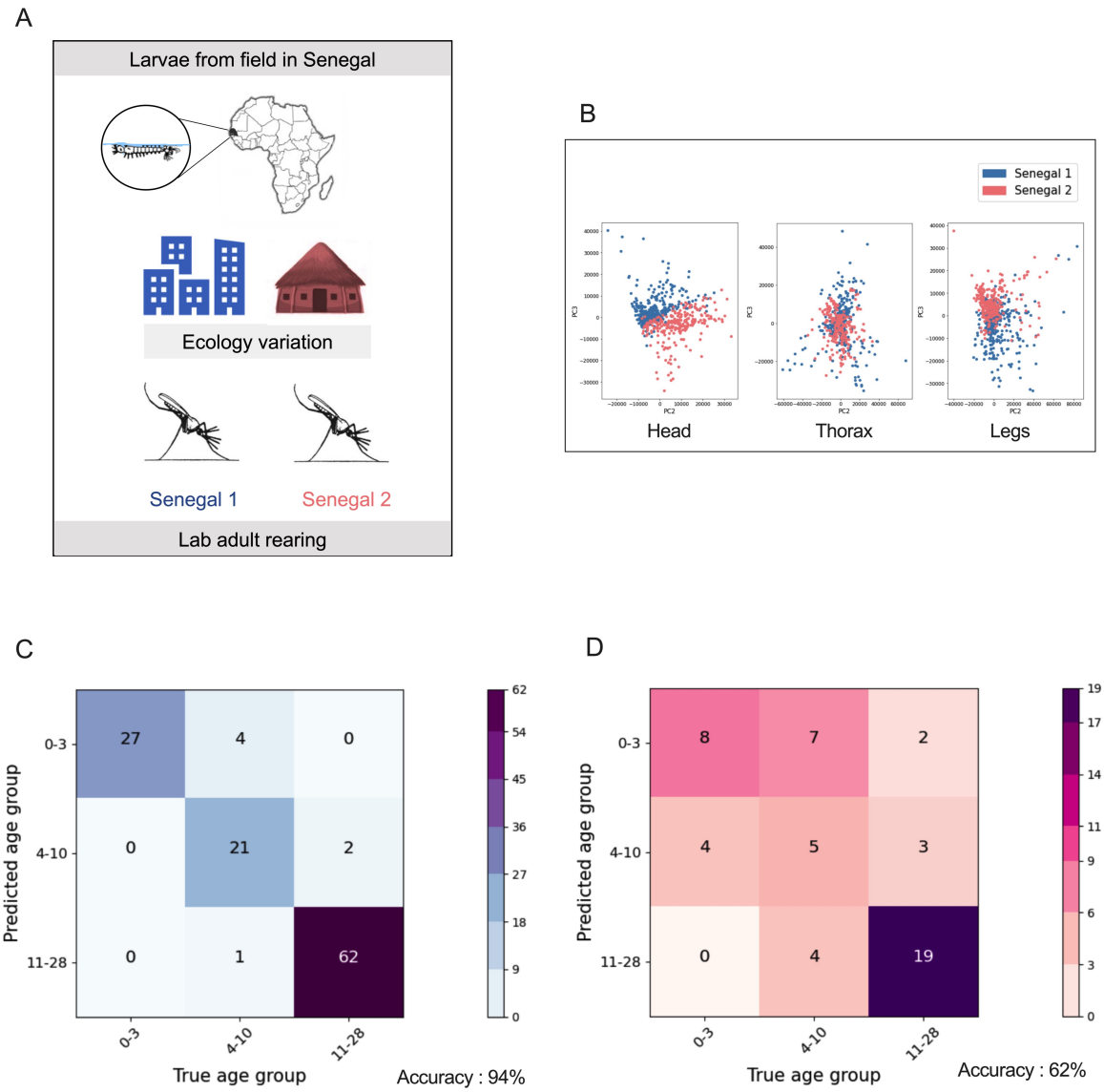
4.4.4 Modélisation de la structure d'âge des populations de moustiques

Pour évaluer l'efficacité de notre approche dans la prédiction de la distribution des âges des populations d'*An. arabiensis* dans différentes conditions, nous avons utilisé des techniques de modélisation. Nous avons créé deux scénarios théoriques : l'un représentant la mortalité naturelle et l'autre reflétant une mortalité accrue due à une intervention de lutte antivectorielle utilisant des moustiquaires imprégnées d'insecticide. Dans le scénario de la moustiquaire imprégnée d'insecticide, nous avons supposé que le taux de mortalité des moustiques femelles adultes serait quatre fois plus élevé que dans des conditions naturelles, à partir du premier repas sanguin, qui se produit généralement autour de trois jours de vie adulte, comme dans une étude de spectroscopie infrarouge i.e. MIRS (Mid-InfRared spectroscopy) précédente (Siria et al., 2022). Pour générer un ensemble de données de test avec une distribution d'âge compatible avec chaque scénario, nous avons utilisé une fonction de survie de Gompertz (Iacovidou et al., 2022). En supposant une population hypothétique de 10 000 moustiques femelles, nous avons utilisé une méthode de bootstrapping pour échantillonner au hasard 100 (on a aussi testé avec 50, 200 et 300) moustiques pour l'ensemble de données de test. L'échantillonnage était basé sur la distribution théorique des âges, traitée comme la probabilité de tirer chaque moustique. Nous avons répété ce processus 100 fois pour tenir compte de la variabilité des résultats et générer des intervalles de confiance. Pour évaluer la précision de nos prédictions d'apprentissage profond à l'aide de ces nouveaux échantillonnages de test, nous avons comparé les distributions d'âge prédites avec les distributions d'âge réelles à l'aide du test de Kolmogorov-Smirnov. L'apprentissage a été effectuée sur les ensembles de données combinés du **Sénégal 1** et du **Sénégal 2** (80 %). Le test a été réalisé sur les (20 %) restant du panel **Sénégal 1** à l'aide du modèle TCN car il contenait le plus grand nombre d'échantillons, minimisant ainsi les effets négatifs de la faible disponibilité de données pour les différentes tranches d'âge dans le modèle (contrairement au panel test Sénégal 2). Les trois parties anatomiques des moustiques ont été incluses à la fois dans l'apprentissage et dans les tests.

4.5 Résultats

4.5.1 Validation de l'approche MALDI-TOF-DL pour les moustiques anophèles collectés sur le terrain

En utilisant un modèle CNN et une classification conventionnelle en trois catégories d'âge, nous avons validé notre méthode MALDI-TOF-DL pour prédire l'âge des moustiques anophèles collectés sur le terrain, élevés en laboratoire et expédiés à température ambiante. Parmi les parties anatomiques testées, les spectres de masse du thorax ont montré la meilleure performance et l'accuracy moyenne la plus élevée de 0,94, pour le **Sénégal 1** (Figure 4.2 C, Tableau 4.1). Pour le **Sénégal 2**, qui avait la moitié du nombre de spécimens, l'accuracy moyenne n'était que de 0,62 pour le thorax (Figure 4.2 D, Tableau 4.1). Nous avons également observé une diminution de l'accuracy pour le groupe d'âge intermédiaire de 4 à 10 jours, qui était plus prononcée au **Sénégal 2** (Tableau 4.1).


A
Larvae from field in Senegal
Ecology variation
Senegal 1
Senegal 2
Lab adult rearing
B
Senegal 1
Senegal 2
Head
Thorax
Legs
C
Predicted age group
True age group
0-3
4-10
11-28
27
4
0
0
21
2
0
1
62
Accuracy : 94%
D
Predicted age group
True age group
8
7
2
4
5
3
0
4
19
Accuracy : 62%

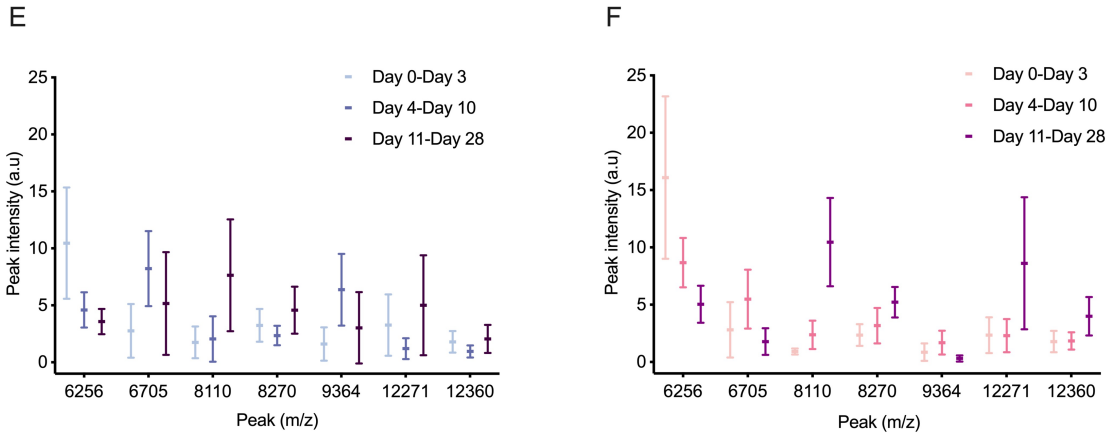


FIGURE 4.2 – **Validation de la SM MALDI-TOF couplée à l'apprentissage profond pour prédire l'âge des moustiques anophèles de terrain** (A) Pour établir un ensemble de données écologiquement variables, des larves de terrain d'*Anopheles arabiensis* ont été collectées dans deux sites écologiquement distincts au Sénégal : urbain (**Sénégal 1**, bleu, n=184 spécimens) et rural (**Sénégal 2**, rose, n=68 spécimens). Les moustiques ont été élevés en laboratoire dans des conditions de température et d'humidité non contrôlées, tout en tenant compte de l'âge des moustiques. (B) Regroupement non supervisé par analyse en composantes principales des spectres de masse MALDI-TOF des parties anatomiques des moustiques (tête, thorax et pattes), avec les ensembles de données complets. Les spectres sont représentés dans un espace bidimensionnel, les couleurs indiquant l'origine géographique. (C, D) Matrices de confusion montrant la classification précise (diagonale) des spectres de masse du thorax en trois catégories d'âge (0-3 jours, 4-10 jours et 11-28 jours) à l'aide d'un réseau de neurones à convolution et d'une classification conventionnelle pour le **Sénégal 1** (C) et le **Sénégal 2** (D). (E, F) Diagrammes en boîte et à moustaches montrant la variation d'intensité des sept pics les plus discriminants (valeurs m/z en Daltons) provenant du profilage des protéines des spectres de masse du thorax pour chacune des trois catégories d'âge (0-3 jours, 4-10 jours et 11-28 jours) dans les ensembles de données du **Sénégal 1** (E) et du **Sénégal 2** (F). La ligne représente l'intensité moyenne, tandis que les moustaches indiquent l'écart-type.

TABLEAU 4.1 – **Performance de la prédiction de l'âge par MALDI-TOF MS couplée au CNN et à une classification conventionnelle ou cohérente avec les rangs, à partir des spectres de masse du thorax.** IC : intervalle de confiance à 95 %. Le nombre indiqué de spectres comprend l'entraînement et le test.

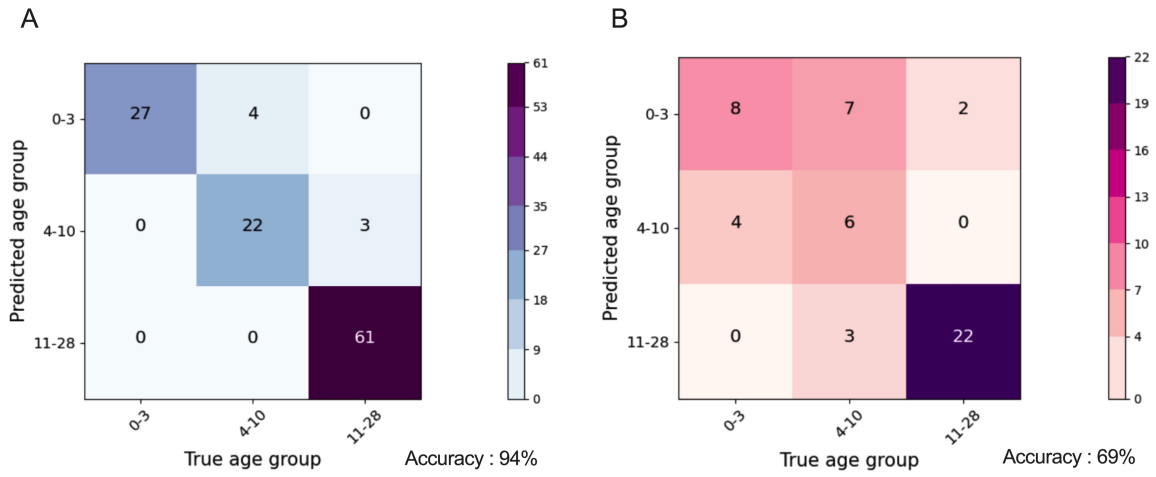
	Senegal 1 (Thorax, n= 806 spectres)				Senegal 2 (Thorax, n=741 spectres)			
Techniques de prédiction	Accuracy (IC)	Accuracy par catégorie	Précision (IC)	Rappel (IC)	Accuracy (IC)	Accuracy par catégorie	Précision (IC)	Rappel (IC)
Classification conventionnelle	0.94 [0.90, 0.97]	[1.0, 0.81, 0.97]	0.92 [0.84, 0.98]	0.93 [0.84, 0.98]	0.62 [0.48, 0.75]	[0.67, 0.31, 0.79]	0.57 [0.36, 0.74]	0.59 [0.37, 0.75]
Rank-consistent	0.94 [0.90, 0.97]	[1.0, 0.85, 0.95]	0.92 [0.85, 0.98]	0.93 [0.85, 0.98]	0.69 [0.56, 0.81]	[0.67, 0.38, 0.92]	0.65 [0.43, 0.81]	0.65 [0.43, 0.8]

Notamment, la performance des différentes parties anatomiques diffère de manière significative entre les ensembles de données du **Sénégal 1** et du **Sénégal 2** (Tableaux supplémentaires Annexe B Table S1 et Table S2). Au **Sénégal 2**, la tête a montré une accuracy moyenne significativement plus élevée de 0,79. Par conséquent, pour améliorer encore les performances de notre méthode, nous avons inclus toutes les parties anatomiques dans l'ensemble de données d'entraînement. Nous avons ensuite fait varier le nombre de répétitions du spectre de 1 à 4 dans l'ensemble de données d'entraînement. Cette combinaison a permis d'améliorer l'accuracy de la prédiction, avec un impact plus important sur le **Sénégal 2** que sur le **Sénégal 1** (Figure supplémentaire Annexe B Fig. S1). L'accuracy moyenne est passée de 0,66 à 0,75 pour le **Sénégal 2** et de 0,85 à 0,90 pour le **Sénégal 1**, de 1 à 4 répétitions.

Il est intéressant de noter que nos prédictions d'apprentissage profond étaient cohérentes avec les résultats du profilage des protéines, ce qui constitue une preuve supplémentaire de la pertinence de la méthode. Nous avons observé des variations dans l'intensité des pics parmi les trois groupes d'âge (0-3 jours, 4-10 jours et 11-28 jours), ce qui a probablement contribué à la classification des spectres de masse du thorax d'*An. arabiensis* par le modèle d'apprentissage profond (Figure 4.2 E, F). Parmi les 25 pics les plus discriminants, sept pics ont été partagés entre les deux ensembles de données et ont montré des comportements similaires de variation d'intensité, malgré la distance géographique de 205 km. Pour la tête et les pattes, certains des pics les plus discriminants ont également été partagés entre les deux ensembles de données, mais moins que pour le thorax (Données supplémentaires Annexe B Data S3). Ces résultats confirment la robustesse de la méthode aux variations génétiques et techniques, comme le montre la Figure 4.2 B. Entre le thorax et/ou la tête et/ou les pattes, nous avons également observé des pics discriminants communs, certains étant partagés entre les ensembles de données (Données supplémentaires Annexe B Data S4) et d'autres semblant être spécifiques au Sénégal 1 ou au Sénégal 2 (Données supplémentaires Annexe B Data S4). Les pics communs aux trois parties du corps étaient présents chez les moustiques du Sénégal 1 (m/z 4137, m/z 4433) et du Sénégal 2 (m/z 5192), mais n'étaient pas partagés entre les ensembles de données. Cela suggère la présence de biomarqueurs spécifiques à la population étudiée, ce qui est cohérent avec la diversité génétique attendue entre les deux ensembles de données.

4.5.2 Optimiser la précision à l'aide de nouvelles méthodes de prédiction.

Nous avons d'abord utilisé le CNN sur les spectres du thorax pour évaluer les nouvelles techniques de prédiction par apprentissage profond. La classification cohérente par rangs (ranking) a donné des résultats similaires à la classification conventionnelle pour le **Sénégal 1** (Figure 4.3 A, Tableau 4.1). Pour le **Sénégal 2**, la classification cohérente avec les rangs a légèrement amélioré les performances, ce qui s'est traduit par une accuracy moyenne plus élevée de 0,69 et un intervalle de confiance réduit (Figure 4.3 B, Tableau 4.1). Cependant, les différences d'accuracy entre les groupes d'âge ont persisté.



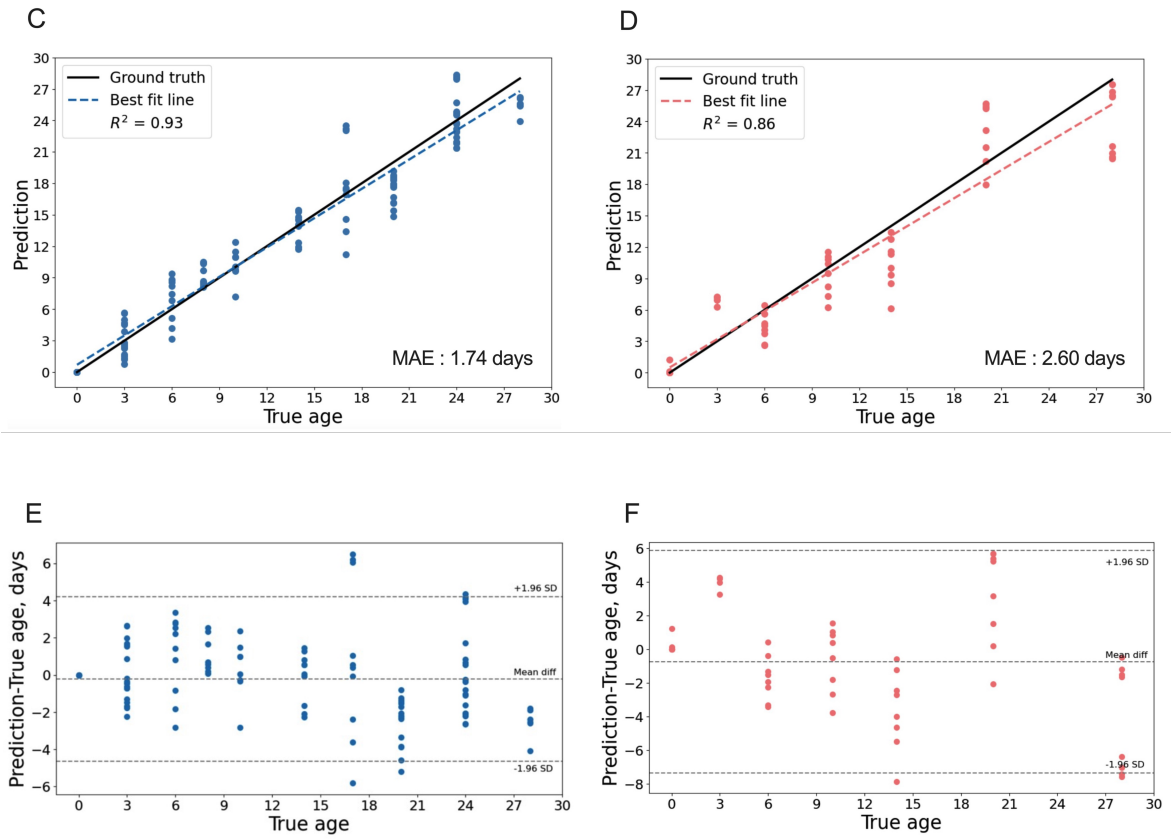


FIGURE 4.3 – **Application de nouvelles techniques de prédiction par apprentissage profond pour améliorer la précision de la prédiction de l'âge des moustiques anophèles.** Les performances des techniques de prédiction ont été évalué en utilisant un réseau de neurones à convolution et des spectres de masse du thorax. (A, B) Matrices de confusion montrant la classification précise (diagonale) des spectres de masse du thorax en trois catégories d'âge (0-3 jours, 4-10 jours, 11-28 jours) en utilisant une classification cohérente avec les rangs pour le **Sénégal 1** (A) et le **Sénégal 2** (B). (C, D) Analyse de la régression montrant la droite de meilleur ajustement (bleu pointillé) et le coefficient de corrélation (Rsquared, R^2) entre l'âge prédit et l'âge réel du moustique en utilisant une prédiction de régression pour le **Sénégal 1** (C) et le **Sénégal 2** (D). La droite de régression idéale, où la prédiction est égale à l'âge réel, est représentée par la droite continue. (E, F) Graphiques de Bland-Altman modifiés illustrant la concordance entre l'âge réel et l'âge prédit à l'aide de la régression. La différence entre l'âge prédit et les valeurs de l'âge réel est tracée en fonction des valeurs de l'âge réel pour le **Sénégal 1** (E) et le **Sénégal 2** (F). La dispersion des valeurs résiduelles des âges prédits est illustrée par les droites pointillées, qui représentent la moyenne ± 1,96 écart-type. MAE : erreur absolue moyenne (Mean Absolute Error).

En revanche, la technique de régression (régression MALDI-TOF-DL) a montré une forte corrélation entre l'âge réel et l'âge prédit pendant toute la durée de vie du moustique. Les valeurs de R carré étaient élevées pour les deux ensembles de données, atteignant 0,93 d'AUROC pour le **Sénégal 1** et 0,86 pour le **Sénégal 2** (Figure 4.3 C, D, Tableau 4.2). Les droites de régression obtenues à partir des âges prédits présentaient un faible écart par rapport à la droite théorique (vérité de terrain), ce qui indique une bonne calibration. L'erreur absolue moyenne (MAE) entre l'âge réel et l'âge prédit était de 1,74 jour pour le **Sénégal 1** et de 2,6 jours pour le **Sénégal 2**. Pour valider davantage nos prédictions, nous avons utilisé un graphique de comparaison inspiré du graphique de Bland-Altman (Figure 4.3 E, F). Ce graphique a montré une concordance entre l'âge réel et l'âge prédit. La plupart des prédictions se situent dans une fourchette de +/- 1,96 écart-type, ce qui indique qu'il y a peu de valeurs aberrantes. La prédiction était suffisamment précise d'un point de vue entomologique, avec des intervalles de prédiction de +/-4 jours pour le **Sénégal 1** et de +/-6 jours pour le **Sénégal 2** sur l'ensemble du spectre d'âge.

Ensuite, nous avons comparé les performances des trois techniques de prédiction pour chaque partie anatomique et ensemble de données en utilisant l'aire sous la caractéristique de l'opérateur récepteur (AUROC) comme mesure commune (Figure 4.4 A, B). Une valeur AUROC de 0,5 représente une performance aléatoire, tandis qu'une valeur de 1 représente une précision parfaite. Nous avons observé une augmentation significative

des valeurs AUROC pour toutes les techniques de prédiction pour la tête du **Sénégal 1** (p<0,001, ANOVA à sens unique) et le thorax du **Sénégal 2** (p=0,02, ANOVA à sens unique). La régression était la méthode la plus précise et la moins variable entre les ensembles de données et les parties anatomiques, avec les valeurs AUROC moyennes les plus élevées et les intervalles de confiance à 95 % les plus faibles.
En utilisant la régression, la valeur AUROC la plus faible était de 0,89 (IC 95 % : 0,83-0,94) pour les pattes du **Sénégal 2**, et la plus élevée était de 0,96 (IC 95 % : 0,94-0,98) pour le thorax du **Sénégal 1**. Enfin, lorsque nous avons combiné la prédiction par régression avec un modèle TCN (Figure 4.4 C, D, et Tableau 4.3), nous avons observé des performances comparables à celles du modèle CNN, sauf pour le thorax du Sénégal 2 et la tête du Sénégal 1, où les valeurs AUROC se sont améliorées de manière significative (p=0,01 et p=0,003, respectivement, test t bilatéral).

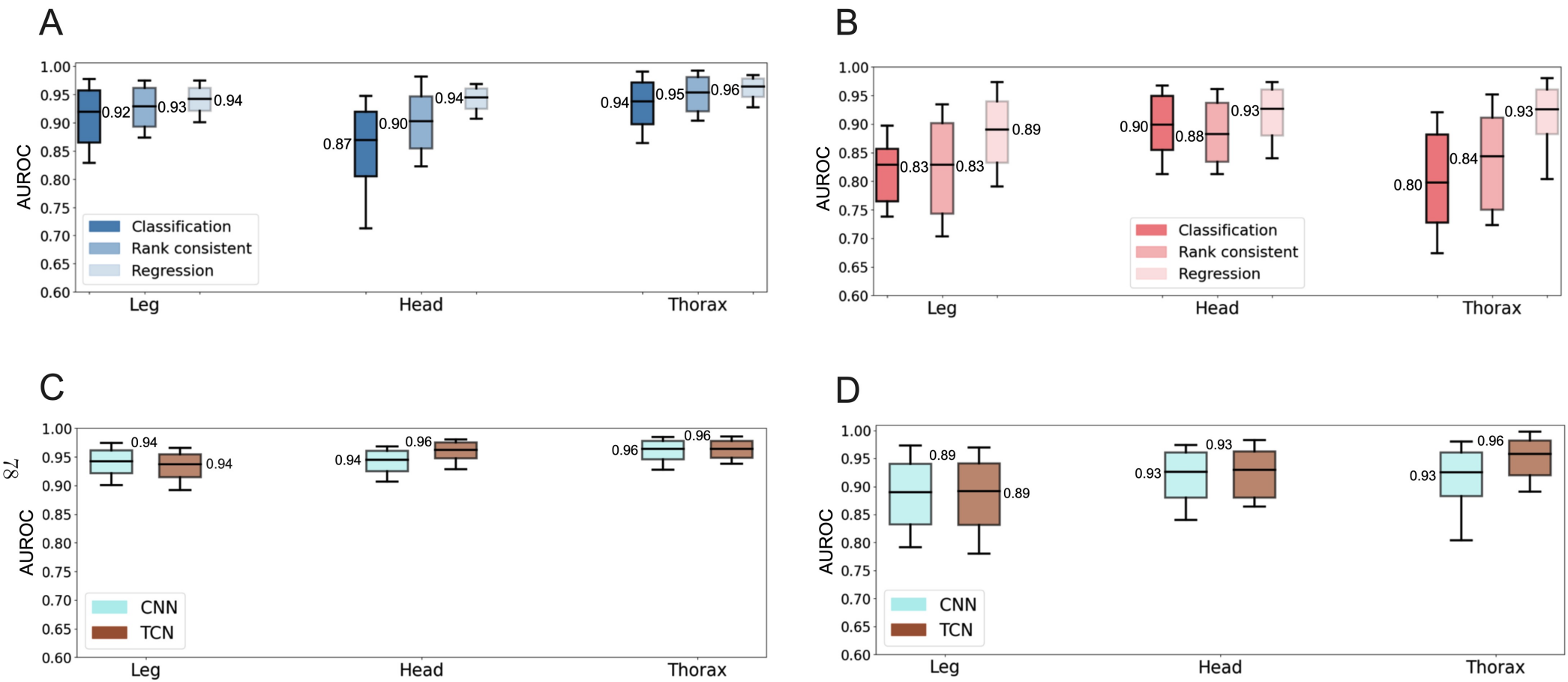


FIGURE 4.4 – **Comparaison des performances des techniques de prédiction et des modèles de réseaux de neurones pour la prédiction de l'âge des moustiques anophèles.** Les performances ont été comparé en utilisant l'aire sous la caractéristique de l'opérateur récepteur (AUROC) comme mesure commune. (A, B) Diagrammes en boîte des valeurs AUROC pour comparer les performances de la classification conventionnelle, de la classification cohérente avec les rangs et de la régression pour le **Sénégal 1** (A) et le **Sénégal 2** (B) à l'aide d'un réseau de neurones à convolution. (C, D) Diagrammes en boîte des valeurs AUROC pour comparer les performances du réseau de neurones à convolution et du réseau de neurones à convolution temporel pour le **Sénégal 1** (C) et le **Sénégal 2** (D) à l'aide de la technique de prédiction par régression. AUROC : aire sous la caractéristique de l'opérateur récepteur.

4.5.3 Généraliser notre modèle d'apprentissage profond à de nouvelles populations cibles

Dans les étapes précédentes, nous avons entraîné nos modèles à l'aide d'un ensemble de données contenant des spectres provenant des deux populations (**Sénégal 1** et **Sénégal 2**) et nous avons testé les modèles sur des sous-ensembles indépendants de chaque population. Pour évaluer la capacité des modèles à se généraliser à une population cible indépendante, sans formation spécifique préalable, nous avons entraîné le modèle CNN sur le **Sénégal 1** et l'avons testé sur le **Sénégal 2**, et vice versa, en utilisant la régression MALDI-TOF-DL (Tableau 4.2).

TABLEAU 4.2 – **Performances de la prédiction de l'âge et capacité de généralisation en utilisant la SM par MALDI- TOF couplé à un CNN et à une régression, à partir des spectres de masse des pattes, de la tête et du thorax.** S1 : **Sénégal 1** ; S2 : **Sénégal 2** ; MAE : erreur absolue moyenne (jours) ; AUROC : aire sous la caractéristique de l'opérateur récepteur ; R^2 : R-carré. *Les spectres des jours 8, 17 et 24 ont été exclus.

	Pattes			Tête			Thorax		
Apprentissage/test (nb spectres Pattes, Tête, Thorax)	MAE	R2	AUROC	MAE	R2	AUROC	MAE	R2	AUROC
S1+S2 (806, 727, 689)/S1 (144, 124, 117)	1.98	0.88	0.94	2.19	0.90	0.94	1.74	0.93	0.96
S1+S2 (806, 727, 689)/S2 (52, 52, 52)	2.92	0.77	0.89	2.77	0.79	0.93	2.60	0.86	0.93
S1 (586, 511, 469)/S2 (52, 52, 52)	4.48	0.56	0.85	3.94	0.57	0.85	3.95	0.66	0.85
S2 (220, 216, 220)/S1 (144, 124, 117)	5.28	0.39	0.80	4.39	0.59	0.84	4.48	0.57	0.90
S1 (418, 352, 312)/S1 (76, 70, 79)*	2.25	0.85	0.91	1.94	0.93	0.97	2.01	0.96	0.90
S2 (220, 216, 220) /S2 (52, 52,79)*	1.96	0.92	0.95	4.2	0.70	0.86	2.31	0.87	0.94

L'accuracy a baissé mais est restée élevée, avec des valeurs AUROC comprises entre 0,80 et 0,90, même avec une formation et des tests totalement indépendants. Notamment, l'ensemble de données du Sénégal 2 comportait des classes d'âge manquantes par rapport au Sénégal 1 (Figure 4.3 C, D), mais les performances sont restées élevées. La MAE était comprise entre 3,94 et 5,28 jours, en fonction de la partie anatomique et de la taille de l'ensemble de données d'entraînement (de 216 à 586 spectres). Dans l'ensemble, la performance la plus faible a été observée lorsque le **Sénégal 2** a été utilisé pour l'entraînement, car il contenait moins de spectres et de classes d'âge que le **Sénégal 1**. La performance la plus élevée a été obtenue lorsque l'entraînement comprenait les deux populations. Des résultats similaires ont été obtenus en utilisant un modèle TCN (Voir le Tableau 4.3 ci-dessous).

Pour vérifier si la baisse de précision était due à la difficulté du modèle à se généraliser ou à des variations inhérentes à certaines populations de moustiques nécessitant un étalonnage local, nous avons entraîné et testé le modèle en utilisant un seul site d'étude (Sénégal 1 ou Sénégal 2), en excluant les jours 8, 17 et 24 de l'ensemble de données du Sénégal 1 pour tenir compte des différences dans la composition des âges. Par rapport à la combinaison des deux sites pour l'entraînement, nous avons observé que l'utilisation d'un seul site d'étude n'améliorait pas la performance, sauf pour les pattes du Sénégal 2 où la performance a augmenté. Pour les autres parties anatomiques, les performances étaient similaires ou augmentées, ce qui montre que les variations locales des spectres ne sont pas une limitation et que le modèle se généralise mieux lorsque l'ensemble des données d'entraînement est plus variable, ce qui est attendu de l'apprentissage automatique. Par rapport à une formation et à des tests totalement indépendants, l'inclusion des souches locales augmente les performances.

TABLEAU 4.3 – **Performances de la prédiction de l'âge et capacité de généralisation en utilisant la SM par MALDI- TOF couplé au TCN et à une régression, à partir des spectres de masse des pattes, de la tête et du thorax.** S1 : **Sénégal 1** ; S2 : **Sénégal 2** ; MAE : erreur absolue moyenne (jours) ; AUROC : aire sous la caractéristique de l'opérateur récepteur ; R^2 : R-carré. *Les spectres des jours 8, 17 et 24 ont été exclus.

	Legs			Head			Thorax		
Apprentissage/test (nb spectres Pattes, Tête, Thorax)	MAE	R2	AUROC	MAE	R2	AUROC	MAE	R2	AUROC
S1+S2 (806, 727, 689)/S1 (144, 124, 117)	2.20	0.69	0.93	1.81	0.80	0.96	2.05	0.87	0.96
S1+S2 (806, 727, 689)/S2 (52, 52, 52)	3.30	0.76	0.89	3.34	0.77	0.93	2.45	0.85	0.96
S1 (586, 511, 469)/S2 (52, 52, 52)	5.55	0.41	0.80	3.85	0.58	0.86	4.06	0.59	0.85
S2 (220, 216, 220)/S1 (144, 124, 117)	4.80	0.49	0.81	4.88	0.51	0.83	5.39	0.37	0.87
S1 (418, 352, 312)/S1 (76, 70, 79)*	2.02	0.87	0.95	1.83	0.92	0.98	2.32	0.85	0.93
S2 (220, 216, 220) /S2 (52, 52,79)*	2.04	0.95	0.91	3.90	0.72	0.86	2.18	0.87	0.94

4.5.4 Simulation de populations de moustiques anophèles sauvages

Pour confirmer l'utilité de notre méthode de régression MALDI-TOF-DL pour les enquêtes de terrain sur les populations d'anophèles, nous avons simulé la structure d'âge d'*An. arabiensis* avant et après une intervention traditionnelle de lutte antivectorielle représentée par une moustiquaire imprégnée d'insecticide (Figure 4.5). Les populations simulées ont montré des différences significatives dans la structure d'âge, avec un changement marqué vers des âges plus jeunes et une réduction substantielle de la survie globale des moustiques dans le scénario post-intervention (Figure 4.5 B) par rapport au scénario de la population naturelle (Figure 4.5 A). Mais l'augmentation simulée du taux de mortalité après exposition à une moustiquaire imprégnée d'insecticide était compatible avec une population de moustiques présentant un certain niveau de résistance, étant donné qu'une population sensible mourrait dans les 24 heures suivant l'exposition, ce qui entraînerait un changement plus spectaculaire du taux de mortalité [3]. Pour les deux scénarios, les distributions d'âge des populations théoriques et prédites étaient statistiquement similaires (p = 0,99, test de Kolmogorov-Smirnov). Cela démontre que l'âge de l'échantillon ne fausse pas la précision de la structure d'âge prédite. En effet, la précision globale est restée élevée bien que l'intervalle de confiance ait été plus large pour les classes d'âge intermédiaires. Cela était attendu en raison du niveau élevé de corrélation entre l'âge réel et l'âge prédit à l'aide de la régression MALDI-TOF-DL (Figure 4.3 C, D). Cela démontre la précision de notre approche d'apprentissage profond dans la reconstruction de la structure d'âge des populations d'anophèles.

La simulation a été réalisée avec un échantillon de taille n=100 moustiques pour l'entraînement, grâce à des itérations aléatoires successives pour modéliser la structure d'âge d'une population de moustiques. Nous avons évalué l'impact de cette taille d'échantillon sur les prédictions d'âge pour les deux scénarios, de n=50 à n=300 moustiques. Les intervalles de confiance étaient larges pour n=50 moustiques, mais ils se sont considérablement réduits lorsque l'échantillonnage a augmenté jusqu'à n=100 moustiques et semblent être stables pour une taille d'échantillon supérieure à n=200 moustiques. Notamment, même avec un échantillon modeste de seulement 100 individus, nous avons pu modéliser avec précision les prédictions de la structure d'âge d'une population de moustiques et notre approche d'apprentissage profond a réussi à capturer les structures d'âge distinctes et à détecter le changement dans la distribution des âges résultant de l'intervention. Ces résultats soulignent l'utilité pratique de notre approche pour évaluer l'efficacité des stratégies d'intervention qui ont un impact sur la longévité des moustiques, à condition que l'effet de l'intervention sur la survie soit suffisamment important pour être détecté.

3. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022

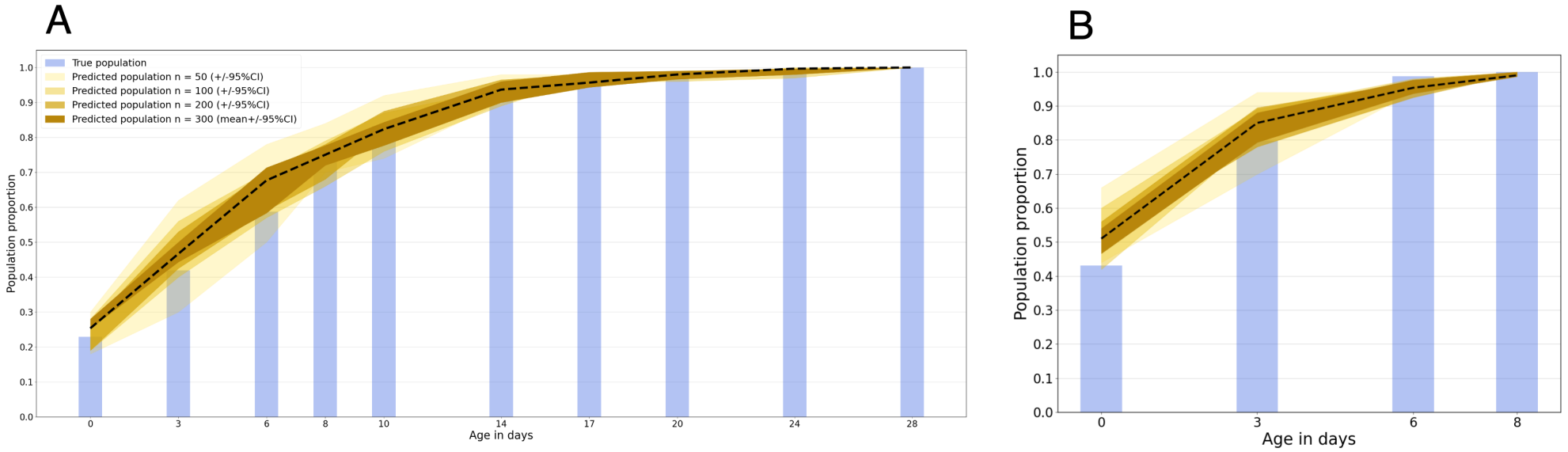


FIGURE 4.5 – **Prédiction des distributions d'âge de populations simulées de moustiques anophèles sauvages avant et après l'intervention d'une moustiquaire imprégnée d'insecticide.** La structure d'âge des populations d'*Anopheles arabiensis* a été reconstruite en utilisant la modélisation (fonction de survie de Gompertz) selon deux scénarios théoriques : l'un avec une mortalité naturelle (A) et l'autre avec une mortalité accrue résultant d'une intervention efficace au moyen de moustiquaires imprégnées d'insecticide (B). Pour évaluer la précision de notre approche dans la prédiction de l'âge de ces nouveaux ensembles de données de test, nous avons utilisé la technique de prédiction par régression et un réseau de neurones à convolution temporel. Les véritables proportions d'âge cumulées de la population d'*An. arabiensis* sont représentées par les barres bleues, tandis que les proportions d'âge cumulées prédites sont représentées par la courbe noire en pointillés. CI : intervalle de confiance.

4.6 Discussion

4.6.1 Points forts et perspectives de cette étude

À notre connaissance, il s'agit de la première étude visant à évaluer l'applicabilité sur le terrain du MALDI-TOF-DL en tant qu'outil de recherche pour surveiller l'âge des moustiques adultes sauvages. Nous avons utilisé 2 763 spectres de la tête, du thorax ou des pattes d'*An. arabiensis* collectés sur le terrain au stade larvaire au Sénégal. Notre méthode démontre sa robustesse face aux variations génétiques et techniques, avec des résultats cohérents sur deux sites écologiques. Grâce à une nouvelle approche de régression par apprentissage profond (régression MALDI-TOF-DL) sans entraînement spécifique préalable, nous avons obtenu la meilleure MAE de seulement 1,74 jour, lorsqu'elle a été entraînée et testée sur des souches de moustiques locales. Lorsqu'elle est appliquée à des populations indépendantes sans étalonnage local, la meilleure MAE était de 3,94 jours. L'estimation de l'âge des moustiques dans une fenêtre étroite de 2 à 4 jours est particulièrement pertinente pour évaluer l'efficacité de nouvelles stratégies ciblant la durée de vie des moustiques au cours des essais de lutte antivectorielle. Cette fenêtre est nettement plus courte que la période d'incubation extrinsèque (PIE) de 9 à 14 jours du parasite Plasmodium chez les moustiques anophèles, à laquelle s'ajoute la période estimée à 2 à 3 jours pour trouver un repas sanguin infectieux (Gilles et al., 2002). Les études de modélisation ont confirmé l'efficacité de notre approche pour surveiller la structure d'âge des populations de moustiques anophèles sauvages avec un effort d'échantillonnage minimal, ce qui est essentiel pour les futures applications de lutte antivectorielle.
Nous démontrons ici l'importance d'un cadre d'apprentissage profond optimisé. Dans notre étude précédente utilisant la SM MALDI-TOF couplée à un CNN, nous avons obtenu une précision moyenne de 73 % dans la prédiction de trois groupes d'âge (0-10 jours ; 11-20 jours ; 21-28 jours) des moustiques *An. stephensi* de laboratoire (Nabet et al., 2020). En utilisant la technique de prédiction par régression, appliquée pour la première fois aux spectres de masse MALDI-TOF, nous avons maintenant amélioré de manière significative la précision et la granularité de la prédiction de l'âge en utilisant des *An. arabiensis* collectés sur le terrain. La prédiction par régression a montré une précision moyenne AUROC constamment élevée (>90 %) pour tous les ensembles de données et toutes les parties anatomiques (tête, thorax, pattes) et nous avons pu prédire de manière fiable l'âge des moustiques jusqu'à 28 jours, y compris pour les groupes d'âge intermédiaires. En utilisant les méthodes de classification, la précision était plus faible pour les groupes d'âge intermédiaires, comme dans notre étude précédente (Nabet et al., 2020) et lors des études MIRS (Siria et al., 2022) et REIMS (Wagner et al., 2023), probablement en raison des changements physiologiques intermédiaires liés au vieillissement, qui rendent plus difficile la classification des moustiques. La méthode de classement n'a pas obtenu de meilleures performances que la classification conventionnelle, cela peut s'expliquer par l'utilisation d'un faible nombre de groupes d'âge. En outre, nous avons ouvert la voie à de nouveaux modèles tels que le TCN, développé pour les signaux audio (Lemaire et al., 2019) même si ses performances ne sont pas significativement meilleures que celles du CNN, comme on aurait pu s'y attendre pour un modèle plus sensible aux dépendances temporelles. Les études précédentes sur les signaux spectraux des moustiques se sont principalement concentrées sur le 1DCNN (Siria et al., 2022 ; Nabet et al., 2020, Merchan et al., 2023), mais d'autres nouvelles architectures de modèles devraient être testées pour améliorer encore les performances, comme le réseau de neurones à retardement (Time Delay Neural Network, Peddinti et al., 2015), l'autoencodeur de débruitage (Li et al., 2022) et les modèles optimisés pour la prédiction de la régression (He et al., 2022). En outre, nous avons confirmé l'importance de l'alignement spectral pour la généralisation à de nouveaux ensembles de données (Mohammad et al., 2023) et nous avons montré que l'inclusion de la diversité des réplicats techniques, des parties anatomiques, de la classe d'âge et de l'origine géographique peut améliorer les résultats de manière significative.

Nous avons montré que nos modèles d'apprentissage profond peuvent se généraliser à de nouveaux ensembles de données, malgré les variations dans les spectres de masse qui pourraient résulter de la diversité génétique, de la dérive des instruments, de la variation des sources ou des différences entre les opérateurs. Ceci est significatif pour l'applicabilité sur le terrain de notre méthode, qui semble peu dépendante de la période et du lieu de collecte des spécimens lorsqu'un même instrument est utilisé. Les résultats étaient également cohérents entre les deux modèles, les trois techniques de prédiction et les trois parties anatomiques du moustique testées, ce qui démontre la robustesse de l'approche d'estimation de l'âge des moustiques. Cette robustesse peut s'expliquer par la présence de biomarqueurs de vieillissement dans les spectres de masse des protéines des moustiques, comme l'a révélé le profilage des protéines. Il est intéressant de noter que certains de ces biomarqueurs étaient conservés entre les deux sites écologiques testés, ce qui suggère une applicabilité potentielle parmi des populations génétiquement distinctes. D'autres biomarqueurs semblent être spécifiques à chaque site d'étude, illustrant la diversité génétique. La résistance aux insecticides pourrait modifier les spectres des moustiques, comme l'a récemment mis en évidence une étude MALDI-TOF, qui a montré des signatures protéiques de spectres de masse distincts entre des colonies *Aedes aegypti* résistantes et sensibles aux pyréthrinoïdes (Almeras et al., 2023). Les populations testées présentent des niveaux variables de résistance aux insecticides. À Keur Massar (Sénégal 1) situé dans la zone urbaine de Dakar, les populations ne sont sensibles qu'aux familles d'organophosphorés (Dia et al, 2018). Par contre, à N'dofan (Sénégal 2), les populations sont sensibles aux organophosphorés et aux

carbamates (Rapport national 2019). Elles sont toutes deux résistantes aux pyréthrinoïdes mais elles détiennent des fréquences alléliques distinctes dans la mutation kdr conférant une résistance croisée aux pyréthrinoïdes et au DDT (Dia et al, 2018 ; Rapport national 2019). Cela renforce le potentiel du modèle à se généraliser à des populations génétiquement distinctes, y compris des populations avec différents niveaux de sensibilité aux insecticides.
Par conséquent, nos résultats appuient une prédiction d'apprentissage profond robuste basée sur les changements protéiques liés à l'âge chez les moustiques anophèles et l'absence de sur-apprentissage, ce qui est prometteur pour le développement sur le terrain.

4.6.2 Comparaison avec les autres approches méthodologiques

Nos résultats sont cohérents avec l'approche du profilage transcriptionnel qui mesure l'expression génique des protéines liées à l'âge (Cook et al., 2007 ; 2006). Des études antérieures de profilage transcriptionnel utilisant des moustiques de laboratoire ont rapporté une valeur résiduelle moyenne de 4,3 jours pour *An. gambiae* (Cook et al., 2010) et de 3 jours pour *Aedes* (Weeraratne et al., 2021). Cette méthode est considérée comme l'un des outils les plus précis pour estimer l'âge des moustiques, offrant une précision suffisante pour des comparaisons au niveau de la population. Cependant, la sensibilité de l'expression des gènes des moustiques aux différentes conditions de développement nécessite un étalonnage local (Wang et al., 2013), ce qui pose un problème pour l'application sur le terrain, en plus du coût de traitement élevé et de la complexité technique. Notre approche de régression MALDI-TOF-DL présente une meilleure applicabilité sur le terrain et montre une précision comparable en utilisant des anophèles collectés sur le terrain, même sans apprentissage spécifique préalable utilisant des moustiques de la population cible. Cependant, une comparaison directe des deux méthodes sur les mêmes échantillons de moustiques devrait être effectuée pour une meilleure évaluation. En outre, il serait intéressant d'évaluer si les biomarqueurs protéiques décrits dans cette étude correspondent aux transcrits dépendant de l'âge identifiés dans les études transcriptionnelles, en utilisant des spectromètres de masse haut de gamme comme le LC-MS/MS.

Comparée à d'autres méthodes basées sur la spectrométrie pour estimer l'âge des moustiques (Siria et al., 2022 ; Wagner et al., 2023), notre méthode est rentable sur le plan informatique tout en étant robuste aux variations spectrales, ce qui constitue un avantage significatif pour une utilisation sur le terrain. Nos modèles sont rentables sur le plan informatique en raison de leur faible niveau de complexité architecturale et de l'optimisation de leurs algorithmes. Ils nécessitent peu de puissance de calcul et de temps de calcul, et une très faible consommation d'énergie, reflétée par une empreinte carbone dérisoire estimée par un traqueur de carbone (Antony et al., 2020). De plus, comme nos modèles sont entraînés avec une taille de lot fixe, le traitement d'ensembles de données plus importants n'augmenterait pas les coûts de calcul. La MIRS peut prédire l'âge des moustiques en détectant les changements dans la composition de la cuticule, mais elle a nécessité des modèles complexes d'apprentissage profond avec apprentissage par transfert (Transfert Learning) et étalonnage local pour traiter les variations spectrales associées aux différentes origines des moustiques (Siria et al., 2022). Plus récemment, la SM à évaporation rapide (REIMS) a permis de prédire l'âge des moustiques sur la base des variations de la teneur en lipides, mais une analyse préalable de regroupement a été nécessaire pour ajuster les catégories d'âge des moustiques, ce qui a soulevé des inquiétudes quant à la précision (Wagner et al., 2023). Les deux méthodes n'ont fourni des estimations de l'âge des moustiques que dans de grandes catégories (moustiques jeunes, intermédiaires et âgés) à l'intérieur de fourchettes d'âge spécifiques. En comparaison, la régression MALDI-TOF-DL est plus précise et peut prédire de manière fiable l'âge des moustiques jusqu'à 28 jours. Les cadres d'apprentissage profond développés dans cette étude pourraient améliorer la précision d'autres méthodes basées sur la spectrométrie ciblant l'âge des moustiques.

4.6.3 Limites de l'étude

Notre étude présente plusieurs limites. Nos résultats ne représentent pas entièrement les conditions naturelles car nous avons utilisé des moustiques collectés sur le terrain au stade larvaire et élevés jusqu'à l'âge adulte dans un insectarium. De plus, nous nous sommes concentrés uniquement sur l'espèce *An. arabiensis*. Si nous avions effectué l'élevage dans des systèmes de semi-campagne, nous aurions pu introduire une plus grande variabilité écologique et nous aurions pu mieux évaluer l'impact des variations environnementales. En effet, une étude MIRS récente s'est avérée précise dans des conditions de laboratoire, mais a échoué lorsque des sites d'élevage plus variés ont été utilisés dans un système de semi-campagne, ce qui suggère une sensibilité aux variations environnementales, telles que la température et l'humidité. Mais l'impact sera probablement moindre pour le MALDI-TOF, car des études antérieures ont montré de bonnes performances sur des spécimens de moustiques

adultes collectés sur le terrain avec des échantillons de moustiques stockés à température ambiante sur du gel silicique (Rakoto). Ceci a été confirmé dans cette étude, car les moustiques adultes ont été à la fois élevés et expédiés sur du gel silical du Sénégal vers la France à température ambiante. En outre, les moustiques qui se développent dans la nature peuvent présenter des schémas de vieillissement plus variables d'une espèce à l'autre, à la fois dans le temps et dans l'espace, ce qui peut nuire à la capacité de généralisation (Lambert et al., 2022 ; Johnson et al., 2020). Par conséquent, les études futures devraient inclure une plus grande hétérogénéité de moustiques adultes entièrement sauvages. Elles devraient inclure différentes espèces d'anophèles et tenir compte du régime alimentaire, de l'état physiologique, de la saison et de l'environnement. De telles études sont compliquées à mettre en oeuvre, car elles nécessitent des compétences techniques pour les comparer aux méthodes actuelles d'estimation de l'âge des moustiques adultes, telles que les méthodes morphologiques ou la recapture par marquage (Johnson et al., 2020). Un travail considérable est nécessaire pour enrichir les ensembles d'entraînement avec une plus grande diversité de spectres de masse d'origines plus variées, pour améliorer la précision du modèle et déterminer si un étalonnage local est nécessaire. Une autre limite est que notre modèle était restreint à l'identification de l'âge des moustiques. L'intérêt du MALDI-TOF pour l'identification des espèces de moustiques vecteurs a déjà été démontré (Yssouf et al., 2016 ; Sevestre et al., 2021), et il serait utile pour l'utilisation sur le terrain d'explorer des modèles qui identifient à la fois l'espèce et l'âge, comme dans les études MIRS (Cook et al., 2010 ; Weeraratne et al., 2021) et REIMS (Wagner et al., 2023).

4.7 Conclusion

Notre approche innovante MALDI-TOF-DL constitue une avancée significative par rapport aux méthodes existantes de suivi de l'âge des moustiques anophèles adultes. L'un des principaux inconvénients du MALDI-TOF est l'achat de l'instrument, qui représente un investissement important (200 000 euros pour un système complet) en plus des coûts annuels de maintenance du spectromètre de masse, et de l'accessibilité de ces services dans les pays où le paludisme est endémique. La méthode nécessite peu de compétences techniques, de consommables et de réactifs, et le coût de traitement est estimé à 1-2 $ par spécimen. Nous pensons donc qu'elle pourrait être facilement exportée vers des pays aux ressources limitées. Les spectres de masse pourraient être acquis localement dans des centres centralisés et analysés à distance à l'aide d'une application d'apprentissage profond en ligne, pour une surveillance en temps réel des populations de moustiques. L'estimation de l'âge des moustiques dans une fenêtre étroite de 2 à 4 jours est particulièrement pertinente pour évaluer l'efficacité de nouvelles stratégies ciblant la durée de vie des moustiques pendant les essais de lutte antivectorielle. Cette fenêtre est nettement plus courte que la période d'incubation extrinsèque (PIE) de 9 à 14 jours du parasite *Plasmodium* chez les moustiques anophèles, à laquelle s'ajoute la période estimée à 2 à 3 jours pour trouver un repas sanguin infectant (Gilles et al., 2002). Ce nouvel outil pourrait également faciliter les études de terrain visant à comprendre la contribution de l'âge des moustiques à la dynamique de la transmission du paludisme. À plus grande échelle, il pourrait fournir des informations inédites sur les facteurs écologiques à l'origine de la longévité des anophèles vecteurs sauvages (Lambert et al., 2022 ; Matthews et al., 2020). Par exemple, déterminer si la survie des anophèles dépend ou non de l'âge permettrait d'améliorer considérablement la lutte contre le paludisme grâce à des modèles épidémiologiques plus précis (Iacovidou et al., 2022). Étendue à d'autres vecteurs de maladies, tels que les *Aedes* vecteurs des virus du *zika*, du *chikungunya* et de la *dengue*, cette approche a le potentiel de révolutionner le domaine de la lutte antivectorielle.

Les codes nécessaires pour évaluer les conclusions de cette étude peuvent être consultés à l'adresse suivante : https://github.com/NoshineMo/Age_Anopheles_DL_MALDI_TOF_Mohammad_Nabet_et_al.

Chapitre 5

Exploration novatrice des modèles d'apprentissage statistique profonds pour l'analyse des spectres MALDI-TOF en épidémiologie des maladies infectieuses

Ce chapitre est issu d'une étude méthodologique réalisée en étroite collaboration avec Cécile Nabet, Alexandre Godmer et Xavier Tannier, dans le cadre de la rédaction en cours d'un article destiné à une revue d'informatique médicale (publication courant 2024).

5.1 Contexte

L'objectif crucial de la médecine moderne est d'obtenir des diagnostics rapides et précis. La technologie de désorption/ionisation laser assistée par matrice et de spectrométrie de masse à temps de vol (MALDI-TOF MS) a révolutionné le domaine de la microbiologie en facilitant l'identification précise et rapide des espèces (Hou et al., 2019). C'est une technique très développée dans les laboratoires de microbiologie clinique, beaucoup de laboratoires en sont équipés.
Une des limites de l'utilisation de la SM MALDI-TOF en routine clinique réside dans son incapacité à distinguer de manière précise certaines espèces bactériennes et fongiques étroitement apparentées déjà décrites chez les champignons (Imbert et al., 2019), leishmanies (Lachaud et al., 2017) ou encore les phlébotomes par les systèmes commerciaux (Chavy et al. 2019). L'application des algorithmes commerciaux utilisés en routine, qui, bien qu'efficaces pour l'identification de la majorité des espèces, montrent leurs limites lorsqu'on aborde des applications plus avancées telles que l'identification d'espèces proches, la caractérisation des souches, la détection de la résistance ou la caractérisation physiologique. Ces algorithmes ne peuvent pas détecter avec précision les variations de signaux, d'intensités et de nuances fines, ce qui représente un défi majeur.

Récemment, l'application des techniques d'apprentissage automatique a permis d'améliorer l'exploitation des données de la SM MALDI-TOF. Cependant, la plupart des études se concentrent principalement sur l'identification des espèces, et les tentatives d'utiliser l'apprentissage automatique avec les spectres MALDI-TOF pour déterminer la résistance aux antimicrobiens sont encore rares (Weis et al., 2020, Han et al., 2021, Popa et al., 2022). De plus, dans le domaine de la microbiologie, les études préfèrent généralement les modèles d'apprentissage automatique traditionnels en raison de leurs avantages, tels que leur capacité à fonctionner avec moins de données, leur facilité d'interprétation et d'utilisation, leur faible besoin de puissance de calcul, et leur capacité à fournir des résultats satisfaisants dans de nombreuses situations. Cependant, à mesure que les données s'accumulent et que l'accès aux ressources informatiques s'améliore, l'utilisation de modèles d'apprentissage profond pourrait devenir plus courante à l'avenir.

Actuellement, les études qui associent l'apprentissage profond aux spectres de masse MALDI-TOF, utilisent souvent des modèles simples tels que les perceptrons multicouches disponibles dans les logiciels de traitement des spectres MALDI-TOF pour l'identification à l'aide d'algorithmes de apprentissage automatique simples (Weis et al., 2022 ; Zhu et Girault, 2023, Rashidi et al., 2022), ainsi que des réseaux de neurones entièrement connectés (Papagiannopoulou et al., 2020 ; Guajardo et al., 2022 ; Deulofeu et al., 2023). Des études plus récentes

explorent l'utilisation de réseaux de neurones à convolution unidimensionnelle (Mortier et al., 2021, Wang et al., 2022, Mohammad et al., 2023). Par ailleurs, certaines recherches novatrices se penchent sur des architectures plus complexes pour améliorer les performances de l'apprentissage automatique profond dans l'identification des espèces (Orellana et al., 2022), en utilisant notamment des Autoencodeurs (Li et al., 2022 ; Zhou et al., 2020), des réseaux de neurones Siamois (Merchan et al., 2023), ou en transformant la représentation du spectre en une matrice bidimensionnelle pour l'utilisation de réseaux de neurones à convolution à deux dimensions (Ling et al., 2020).

Ces récentes études suscitent des interrogations sur la représentation optimale des spectres de masse MALDI-TOF et sur les diverses architectures de réseaux de neurones qui peuvent être exploitées à cette fin. Compte tenu de la diversité des architectures de réseaux neuronaux, incluant celles qui n'ont pas encore été exploitées dans le domaine des spectres MALDI-TOF, une étude comparative visant à développer et évaluer ces modèles inexplorés par rapport à ceux déjà documentés dans la littérature pourrait fournir des informations essentielles sur leur pertinence pour la spectrométrie de masse MALDI-TOF.

L'intérêt de cette démarche serait renforcé par une évaluation des modèles sur des ensembles de données variés en termes de complexité. Actuellement, une vue d'ensemble complète des performances de ces modèles sur des cohortes diverses fait défaut. L'exploration de diverses architectures de réseaux de neurones pourrait fournir des informations essentielles sur les composants et les structures de réseau qui améliorent l'analyse des spectres et les prédictions. En outre, cela pourrait ouvrir la voie à de nombreuses autres possibilités architecturales et stimuler de nouvelles perspectives de recherche dans ce domaine.

5.2 Objectif

Nous réalisons une étude systématique pour explorer divers modèles de réseaux neuronaux dans l'analyse des données de SM MALDI-TOF, incluant des modèles encore inexploités dans ce domaine. Notre démarche consiste à explorer différentes représentations possibles des spectres de masse et évaluer les modèles de réseaux neuronaux qui leur sont associés. Fournissant des résultats directement comparables sur diverses tâches, notre objectif est d'inspirer et d'informer les chercheurs désireux d'appliquer ces techniques à leurs problématiques spécifiques, ainsi que de guider les futures recherches dans ce domaine.

Pour cette étude, trois cohortes distinctes ont été étudiées, chacune présentant des caractéristiques spécifiques. Ces cohortes ont été sélectionnées pour représenter divers défis pertinents, notamment l'identification des souches, des espèces proches voire des sous-espèces, ainsi que la résistance et l'état physiologique. Nous avons également tenu compte des éventuelles contraintes liées aux ressources informatiques. Les cohortes utilisées sont les suivantes :

- Une cohorte déséquilibrée de souches de *Mycobacterium abscessus*. La tâche est de classer les trois sous-espèces *M. abscessus*, *M. bolletii* et *M. massiliense*, ainsi qu'à identifier les souches résistantes parmi les sensibles, en utilisant un jeu de test indépendant.
- Une cohorte de moustiques anophèles *An. arabiensis* provenant du Sénégal, incluant des moustiques de milieu rural et urbain. La tâche est d'estimer l'âge des moustiques en jours (par régression) à l'aide d'un jeu de test indépendant.
- Une cohorte de moustiques provenant de diverses régions géographiques, destinée à l'identification de quatre espèces de moustiques anophèles : *An. arabiensis*, *An. coluzzii*, *An. funestus* et *An. gambiae*.

Ces cohortes ont été conçues pour exposer nos réseaux de neurones à des défis représentatifs de la pratique de terrain.

Nous avons réalisé une évaluation approfondie en utilisant différents modèles de réseaux de neurones, chacun adapté à des problématiques spécifiques. Notre étude englobe diverses architectures, notamment le réseau de neurones à convolution à une dimension (1DCNN) et le réseau de neurones à convolution temporelle (TCN), qui ont été conçus pour traiter des données temporelles. Nous avons également exploré des modèles récurrents, comme le bidirectionnel GRU (RNN-BiGRU) et l'Echo State Network (ESN), traditionnellement utilisés dans le domaine temporel.
De plus, nous avons évalué l'applicabilité de modèles conçus pour des données bidimensionnelles, comme les réseaux de neurones à convolution à deux dimensions (2DCNN) généralement utilisés pour le traitement d'images, pour l'analyse des spectres MALDI-TOF représentées par des images bidimensionnelles (spectrogramme et scalogramme). En outre, nous avons examiné les avantages d'un modèle hybride combinant des composantes de convolution et de récurrence.

Enfin, nous avons exploré le potentiel de deux types d'Autoencodeurs : l'Autoencodeur à couches entièrement connectées (DAE) et l'Autoencodeur à couches de convolution (DCAE). Ces Autoencodeurs sont conçus pour réduire les spectres MALDI-TOF, filtrer leurs caractéristiques et les fournir en entrée à des modèles d'apprentissage automatique classiques.

Les critères de sélection des meilleurs modèles ne se basent pas seulement sur leurs performances, mais aussi sur leur temps d'exécution et leur consommation d'énergie pendant l'entraînement, mesurés simultanément.

5.3 Données

5.3.1 Constitution des cohortes et définition des tâches

Pour évaluer les modèles de manière efficace dans cette étude, nous avons utilisé plusieurs cohortes comportant des caractéristiques spécifiques. Chacune de ces cohortes est composée d'un ensemble d'apprentissage et d'un ensemble de test indépendant, permettant ainsi d'évaluer les performances des modèles sur des données inconnues.

Identification de sous espèces et détection de la résistance chez *Mycobacterium abscessus*

Le complexe *Mycobacterium abscessus* (MABSc) regroupe des bactéries opportunistes pathogènes (Forbes et al., 2018 ; McGrath et al., 2010 ; Van der Werf et al., 2014 ; Mougari et al., 2016). Les trois sous-espèces génétiquement proches (*M. abscessus*, *M. bolletii* et *M. massiliense*) de ce complexe possèdent des sensibilités différentes à la clarithromycine qui est un antibiotique dans le traitement de ces infections. Environ 70 à 80 % des souches de *M. abscessus* et les souches de *M. bolletii* expriment naturellement une résistance aux macrolides induite par le gène *erm-(41)*, tandis que *M. massiliense* est généralement naturellement sensible à la clarithromycine en raison de mutations génétiques spécifiques (Bastian et al., 2011 ; Brown-Elliott et al., 2015 ; Koh et al., 2017). Ces variations génétiques expliquent les différences de réponse au traitement entre ces sous-espèces (Huang et al., 2010 ; Harada et al., 2012). Par conséquent, il est essentiel de distinguer ces trois sous-espèces du MABSc (pour guider les décisions thérapeutiques.

Nous avons constitué une cohorte de spectres de masse MALDI-TOF provenant de souches caractérisées par méthode moléculaire dans le but d'évaluer la capacité des modèles d'apprentissage profond à accomplir deux tâches de classification : (i) identifier les trois sous-espèces (*M. abscessus*, *M. bolletii*, *M. massiliense*) et (ii) distinguer les souches résistantes (R : résistant) des sensibles (S : sensible) aux antibiotiques.

Cette cohorte présente des particularités notables. Tout d'abord, elle comporte un faible nombre de souches et présente un déséquilibre marqué entre les sous-espèces (Voir le Tableau 5.1). Ce déséquilibre est pertinent pour l'apprentissage automatique profond, car il pose le défi de la gestion des classes déséquilibrées. De plus, les spectres de MALDI-TOF sont complexes, car les trois sous-espèces partagent une proximité génétique qui se traduit par des similitudes marquées sur plusieurs parties du spectre, entraînant des chevauchements fréquents de pics aux mêmes positions en termes de m/z. L'ensemble d'apprentissage représente 80 % de la cohorte, tandis que les 20 % restants constituent le jeu de test indépendant.

Prédiction de l'âge des moustiques anophèles sur le terrain

La cohorte est détaillée dans le paragraphe 4.3 du chapitre précédent.

Comme indiqué dans le chapitre précédent, la survie des moustiques *Anopheles* est cruciale pour la surveillance épidémiologique du paludisme et l'évaluation des stratégies de lutte antivectorielle axées sur la durée de vie des moustiques, car le risque de transmission du parasite augmente avec l'âge des moustiques. Malheureusement, les méthodes actuelles sur le terrain pour estimer l'âge des moustiques sont souvent imprécises et chronophages. Nous cherchons à explorer différents modèles d'apprentissage profond pour prédire, par régression, l'âge en jours des moustiques *Anopheles arabiensis* sur le terrain, tout en garantissant leur capacité à généraliser à des populations génétiquement diverses et à des spectres présentant une variabilité technique importante. La particularité de cette étude réside dans l'identification de modifications protéiques liées à l'âge, sous la forme de biomarqueurs

protéiques de vieillissement (Sikulu et al., 2015, expliqué dans le chapitre précédent).

Pour créer un ensemble de données présentant une variabilité génétique, nous avons collecté des spécimens de larves d'*Anopheles arabiensis* sur un site urbain au Sénégal (Sénégal 1) et spécimens sur un site rural (Sénégal 2). Ces moustiques ont ensuite été élevés en laboratoire de manière à connaître leur âge de façon précise. L'ensemble d'apprentissage est constitué de 80 % des deux bases de données de Sénégal 1 et de Sénégal 2. Sur chaque partie anatomique (pattes, tête et thorax) traitée séparément, notre objectif est de prédire l'âge en jours des 20 % restants de la base Sénégal 1, qui constitue notre jeu de test indépendant (pour éviter la surcharge de présentation des résultats, nous ne présentons pas les performances des 20 % restant de la base Sénégal 2). Les cohortes Sénégal 1 et Sénégal 2 ont été analysées à deux ans d'intervalle par des opérateurs différents sur le même automate, ce qui introduit de la variabilité technique.

Identification des espèces de moustiques anophèles sur le terrain

Environ 500 espèces d'anophèles existent, mais seulement une douzaine d'entre elles sont des vecteurs efficaces en Afrique (Harbache et al., 2004). Les espèces les plus courantes sont *An. arabiensis*, *An. coluzzii*, *An. funestus* et *An. gambiae*, et sont responsables de 90 % de la transmission du paludisme sur le continent africain (Carnevale et al., 2009). Du fait de capacités de transmission du paludisme différentes entre les espèces, l'identification précise des espèces d'anophèles est essentielle pour le contrôle des vecteurs, mais les méthodes actuelles présentent des limitations (Manguin et al., 2013). Les méthodes moléculaires sont coûteuses et chronophages, la morphotaxonomie ne peut pas identifier les espèces cryptiques (morphologiquement identiques mais distinguables sur le plan génétique), et même les méthodes basées sur des critères morphologiques nécessitent un observateur expérimenté (Erlank et al., 2018).

Notre objectif est d'évaluer les performances des modèles de réseaux de neurones pour identifier les quatre espèces : *An. arabiensis*, *An. coluzzii*, *An. gambiae* et *An. funestus* . Il convient de noter que les espèces *arabiensis*, *coluzzii* et *gambiae* sont étroitement apparentées au sein du complexe *Gambiae* (Carnevale et al., 2009). En conséquence, les profils spectraux de ces espèces sont très similaires. *An. gambiae* et *An. coluzzii* sont encore plus génétiquement proches bien qu'elles aient été classées comme des sous-espèces auparavant, elles sont maintenant considérées comme des espèces distinctes (Coetzee et al., 2013). Cependant, malgré leur ressemblance morphologique, des biomarqueurs protéiques potentiels ont été mis en évidence par spectrométrie MALDI-TOF pour différencier les espèces proches d'anophèles (Muller et al., 2013).

D'autre part, *An. funestus* appartient à un groupe taxonomique plus éloigné, ce qui en fait une espèce utilisée comme contrôle interne de référence.

Cette cohorte présente des particularités importantes : les moustiques anophèles proviennent de diverses régions géographiques (Mali, Guinée, Sénégal, Afrique du Sud, République Démocratique du Congo et Kenya) et sont issus de colonies de laboratoire et du terrain, conférant ainsi une grande diversité génétique à l'échantillonnage. Parmi les moustiques de terrain, certains sont issus de la collecte de larve qui ont été élevées en adulte en insectarium alors que d'autres sont issus de la collecte directe d'adultes, dont certains d'entre eux sont gorgés ou semi-gorgés de sang, ce qui implique de la diversité physiologique (Voir le tableau Table S1 en Annexe C pour plus de détails). Ces particularités complexifient le défi, car elles contribuent à la variabilité des spectres. De plus, le déséquilibre dans le nombre de moustiques par espèce représente un défi supplémentaire pour l'identification des espèces. (Voir le Tableau 5.1 ci-dessous). Malgré ces variations, en traitant les parties anatomiques (pattes, tête et thorax) indépendamment, nous cherchons à développer des modèles d'apprentissage profond capables de distinguer ces quatre espèces.

L'ensemble d'apprentissage est constitué de 80 % des spécimens de la base de données qui comporte un mélange d'*An. arabiensis* (Sénégal, Afrique du Sud) , d' *An. coluzzii* (Mali, Guinée), d'*An. gambiae* (Mali, Guinée, Congo et Kenya) et d'*An. funestus* (Mali). Nous cherchons à l'aide de modèles d'apprentissage profond à identifier ces quatre espèces sur les 20 % des spécimens restant (jeu de test indépendant).

TABLEAU 5.1 – **Informations générales sur les cohortes.**

Cohorte	Tâche	Distribution des classes (ou plage de valeur)	Quantité de données (apprentissage/test)	Objectif
MABSc	Classification à 3 classes	*M. abscessus* : 63% *M. bolletii* : 26% *M. massiliense* : 11%	33/8 souches 814/187 spectres	Identification des trois sous-espèces : *M. abscessus*, *M. bolletii* et *M. massiliense*
MABSc	Classification binaire	R : 68% S : 32%	33/8 souches 814/187 spectres	Détection de la résistance : Résistant et Sensible
Anophèles âge	Régression	0, 3, 6, 8, 10, 14, 17, 20, 24, 28	Pattes : 202/36 spécimens 806/144 spectres Tête : 183/32 spécimens 727/124 spectres Thorax : 179/31 spécimens 689/117 spectres	Prédire l'âge des moustiques de 0 à 28 jours
Anophèles espèces	Classification à 4 classes	*An. arabiensis* : 47% *An. coluzzii* : 29% *An. gambiae* : 21% *An. funestus* : 3%	Pattes : 211/40 spécimens 1047/188 spectres Tête : 214/46 spécimens 1089/214 spectres Thorax : 203/54 spécimens 999/279 spectres	Classification des quatre espèces : *An. arabiensis*, *An. coluzzii*, *An. gambiae* et *An. funestus*

5.3.2 Acquisition des spectres de masse pour les différentes cohortes

Complexe *Mycobacterium abscessus*

Les 41 souches ont été cultivées à 37°C en atmosphère aérobie sur gélose au sang (COH, bioMérieux®) pendant environ 7 jours. Ensuite, une colonie a été extraite selon le protocole MycoEx (Bruker®). Chacun de ces extraits a été analysé par MALDI-TOF MS (Bruker®) en effectuant 8 répétitions techniques, avec l'application d'une matrice MALDI (α-HCCA) sur les spots séchés [1].

Les spectres de masse ont été obtenus à l'aide d'un instrument Microflex® LT (de Bruker® Daltonics), en respectant les paramètres par défaut de la méthode normalisée recommandée par Bruker pour les diagnostics in vitro (CE-IVD : Conformité Européenne pour Dispositif Médical de Diagnostic In Vitro). Cet appareil était équipé d'un laser N2 (λ =377 nm), avec les paramètres suivants : plage de masse de 200 à 20000 Da, tension de la source d'ions 1 : 20 kV, tension de la source d'ions 2 : 18,5 kV, tension Iens : 8,45 kV, extraction d'ions pulsés : 330 ns, fréquence du laser : 20,0 Hz. Les spectres ont été obtenus après 500 tirs. Chaque dépôt été réalisé 6 fois et analysé par le laser 3 fois. Pour l'étalonnage, un standard de calibration externe (Bacterial Test Standard de Bruker Daltonics) a été employé, et l'acquisition des données a été réalisée à l'aide de FlexControl (version 3.0 de Bruker Daltonics).

Les spectres de mauvaise qualité ont ensuite été éliminés après analyse visuelle. Conformément aux recommandations de Bruker pour la création d'une base de données spectrale, les critères suivants ont été utilisés : (i) présence de pics anormaux par rapport aux autres spectres, (ii) existence de spectres avec des pics plats par rapport aux autres spectres, et (iii) spectres présentant un décalage de masse supérieur à 500 ppm par rapport aux autres spectres. Si moins de 5 spectres satisfaisants étaient obtenus parmi les 48 spectres issus de souches cultivées pendant 4 à 7 jours, l'extraction des protéines était répétée.

1. Voir le lien suivant pour plus de détail https://zenodo.org/record/5793313

Âge des moustiques anophèles et identification des espèces proches de moustiques sur le terrain

Après dissection du moustique, l'extraction des protéines a été réalisée à partir de la tête, du thorax avec les ailes et des pattes selon un protocole antérieur (Voir la section 4.3 du chapitre précédent). Ensuite, les extraits de protéines ont été appliqués sur une plaque d'acier et recouverts d'une matrice HCCA. Pour garantir la reproductibilité des résultats, chaque spécimen et parties anatomiques ont été soumis à quatre réplicas techniques (dépôts d'un même extrait protéique) pour les différentes cohortes de moustiques. Jusqu'à dix réplicas ont été réalisés pour la cohorte sur l'identification des espèces proches de moustique anophèles en vue de la construction d'une banque d'identification.

Pour les deux cohortes, les spectres de masse ont été acquis avec un spectromètre de masse Microflex LT (Bruker France SAS) avec un laser émettant dans l'UV à une fréquence de 60 Hz qui tire 240 fois sur 6 régions au niveau de chaque dépôt. Les données ont été acquises en utilisant AutoXecute sur le logiciel FlexControl v3.4 softwares (Bruker France SAS) puis exportées sur le logiciel MALDI Biotyper v4.1 software (Bruker France SAS) avec les paramètres d'acquisition par défaut, comme cela a été fait dans le protocole précédent (Nabet et al., 2020). Les données acquises ont ensuite été exportées vers Maldi Biotyper v4.1 pour évaluer la reproductibilité des réplicats de spectre et la qualité des spectres de masse. Pour garantir la qualité des données, nous avons effectué une étape de contrôle de la qualité des spectres comme décrit dans Nabet et al., 2021 (Voir le Chapitre précédent section 4.3).

Pour une analyse robuste des variations spectrales, les deux ensembles de données ont été acquis indépendamment par différents opérateurs utilisant le même instrument.

5.4 Modèles avec différents contextes d'entrée

Les différents réseaux de neurones présentés dans cette section ont été répartis en fonction de leur type et de la nature des entrées qu'ils peuvent traiter. Pour cette raison, chaque modèle est accompagné de méthodes de pré-traitement spécifiques. Tous les modèles de cette section, associés à leurs méthodes de pré-traitement respectives, ont été appliqués aux spectres MALDI-TOF pour effectuer des classifications binaires, multi-classes et des régressions dans le cadre de diverses études d'identification et d'estimation de cohortes.

5.4.1 Les réseaux de neurones à convolution 1D

Cette section reprend les modèles utilisés dans les chapitres précédent, voir les sections 3.4 et 4.4.

Nous proposons deux réseaux de neurones à convolution appliqués sur des données unidimensionnelles : le réseau de neurones à convolution unidimensionnel (1DCNN) et le réseau de neurones à convolution temporel (TCN).

Le pré-traitement des spectres

Le traitement des spectres suit la même procédure que celle exposée dans les chapitres antérieurs. Il débute par un processus de lissage à l'aide de la méthode de la moyenne mobile avec un facteur de 1/9. Ensuite, la soustraction de la ligne de base est effectuée en utilisant la technique des moindres carrés asymétriques (Eilers et al., 2005). Enfin, la sélection des pics est réalisée au moyen de la méthode de la dérivée des spectres (He et al., 2011) dans le but de minimiser le bruit tout en préservant les signaux de faible intensité (Voir la figure 5.1 ci-dessous).

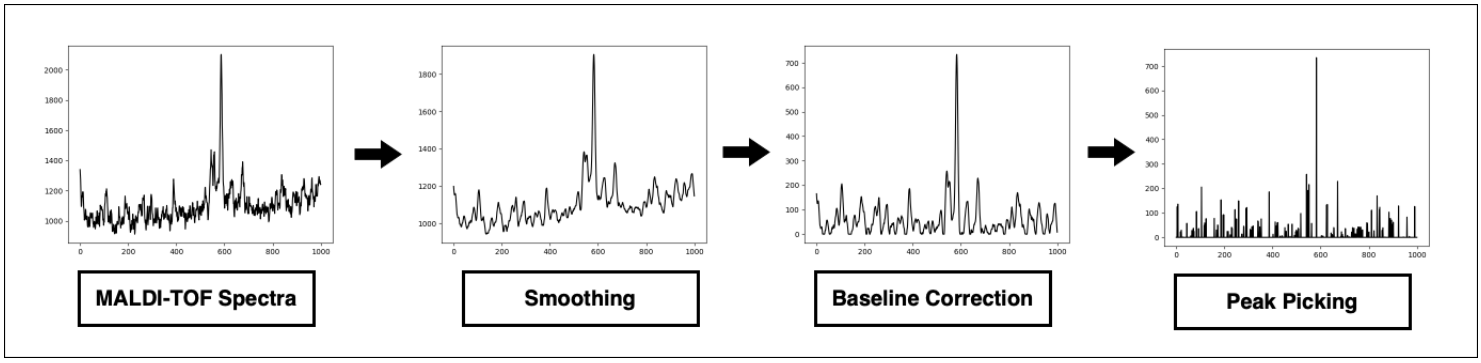


FIGURE 5.1 – **Pré-traitement des spectres sur CNN et TCN.**

Les modèles

Il s'agit des modèles présentés dans le chapitre précédent.

Les réseaux de neurones à convolution 1D (1DCNN) et ceux à convolution temporelle (TCN) utilisent des couches de convolution pour détecter, localiser, et quantifier les pics d'intérêt dans les spectres, puis les traiter dans les couches suivantes pour effectuer des prédictions (Voir la figure 5.2).
Contrairement au CNN, le TCN adopte une approche hiérarchique de convolutions pour capturer des modèles temporels étendus et résiste bien aux retards temporels. L'introduction de ces caractéristiques filtrées dans un classificateur permet d'exploiter des informations temporelles de haut niveau. Pour ce faire, les TCN sont équipés de couches de convolution dotées de filtres à pas variable, leur permettant ainsi de détecter des dépendances entre des éléments qui ne sont pas nécessairement adjacents (Lea et al., 2016). En revanche, les CNN utilisent des filtres qui se déplacent linéairement à travers les données d'entrée, en se concentrant sur les éléments proches (Kiranyaz et al., 2021).

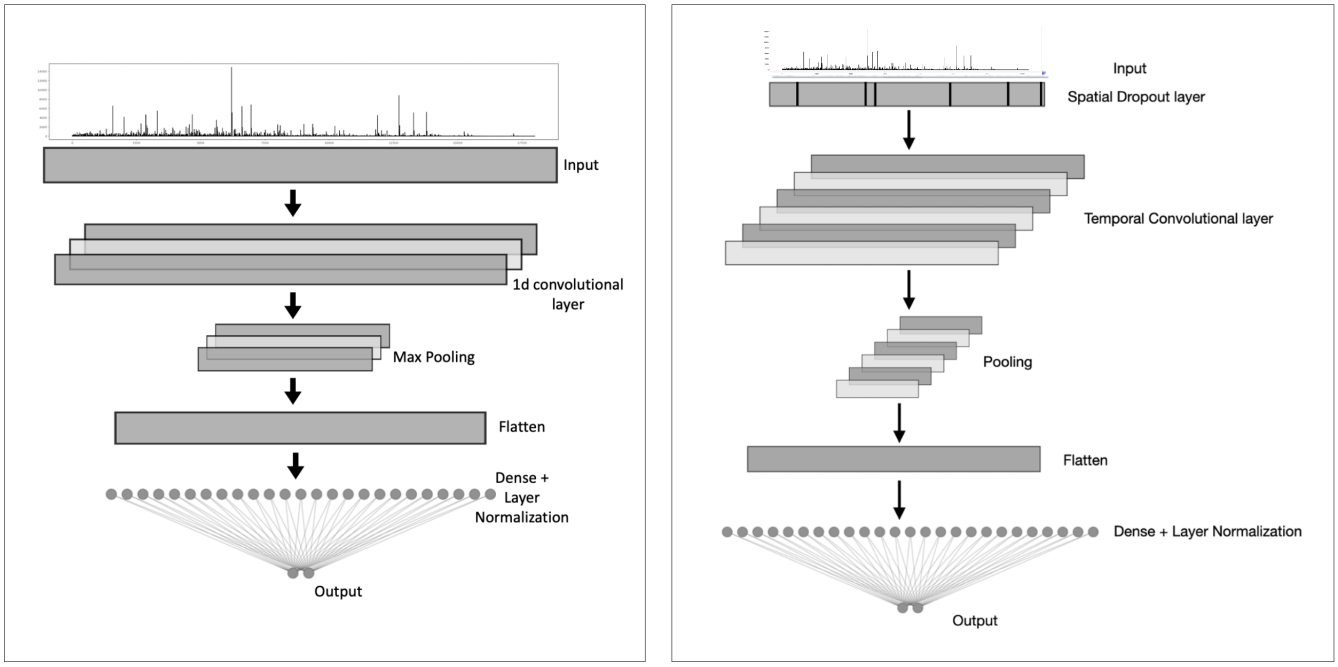


FIGURE 5.2 – **Architectures du CNN (gauche) et du TCN (droite)**

5.4.2 Les réseaux de neurones récurrents

Dans cette section, nous avons employé des réseaux de neurones récurrent (RNN : Recurrent Neural Network) appliqué sur des spectres unidimensionnelles. Nous proposons deux modèles de réseaux de neurones récurrent : le RNN-BiGRU et l'ESN.

Le pré-traitement des spectres

Les spectres ont été prétraités pour être analysés comme des signaux sonores. Tout d'abord, nous avons appliqué un lissage en utilisant la méthode de la moyenne mobile, suivi de la soustraction de la ligne de base grâce à la méthode des moindres carrés asymétriques, directement sur le spectre brut. Ensuite, nous avons choisi le spectre le plus court en fonction de la plage de temps de vol, puis nous avons interpolé (He et al., 2011) tous les spectres sur l'axe des temps de vol du spectre le plus court sélectionné.
Pour l'entraînement, nous avons utilisé les spectres de l'ensemble d'apprentissage, tandis que ceux de l'ensemble de test ont été ajustés pour correspondre aux temps de vol du spectre le plus court de l'ensemble d'apprentissage. Cette interpolation, avec un espacement de 4, a réduit la taille initiale du spectre de 20 000 à 5 000 points, sans altérer le contenu du spectre, car généralement un pic s'étend sur 30 points. Ensuite, nous avons appliqué une dérivation première aux spectres résultants de 5 000 points pour éliminer le bruit et permettre aux informations du spectre de s'étendre sur des valeurs positives et négatives (Voir la figure 5.3 ci-dessous).

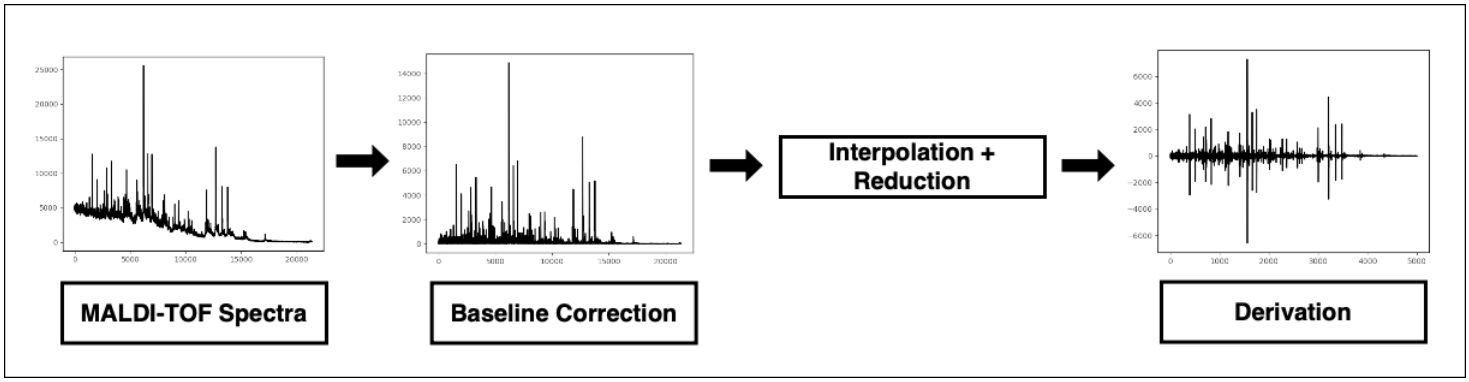


FIGURE 5.3 – **Pré-traitement des spectres sur RNN-BiGRU et TCN.**

Les modèles

Les réseaux de neurones récurrents, notamment les RNN-GRU bidirectionnels (Lynn et al., 2019), sont largement utilisés pour traiter des données séquentielles et extraire des motifs complexes et irréguliers, améliorant ainsi la représentation hiérarchique des dépendances séquentielles.
Un RNN-GRU bidirectionnels (RNN-BiGRU) est utilisé ici pour détecter des variations inhabituelles dans les spectres de masse qui ne suivent pas les schémas normaux, en tenant compte des caractéristiques spécifiques du spectre et de son historique temporel. Il s'inspire de la complexité des spectres MALDI-TOF, qui contiennent des relations entre des éléments à différentes distances, permettant de capturer des caractéristiques liées aux protéines. Pour ce faire, cette méthode divise les données en sous-séquences, transformant chaque point temporel en une caractéristique du réseau (Liu et al., 2019).

L'Echo State Network (ESN) (Sun et al., 2020) est une variante de réseau récurrent qui utilise des couches cachées initialisées de manière aléatoire pour former un réservoir. Ce réservoir, en tant que composante centrale de l'ESN, fonctionne comme un réseau de neurones récurrent (RNN) aléatoire faiblement connecté. Chaque neurone du réservoir produit sa propre réponse non linéaire en fonction de l'entrée. Les poids des connexions internes du réservoir et les poids d'entrée restent constants, seules les pondérations des sorties sont ajustées grâce à un algorithme d'apprentissage spécifique. Nous évaluons ce modèle en utilisant des spectres de masse MALDI-TOF pour déterminer sa capacité à identifier et prédire de manière précise, malgré la faible connectivité du réservoir, qui stocke un volume limité d'informations dans les séquences des spectres.

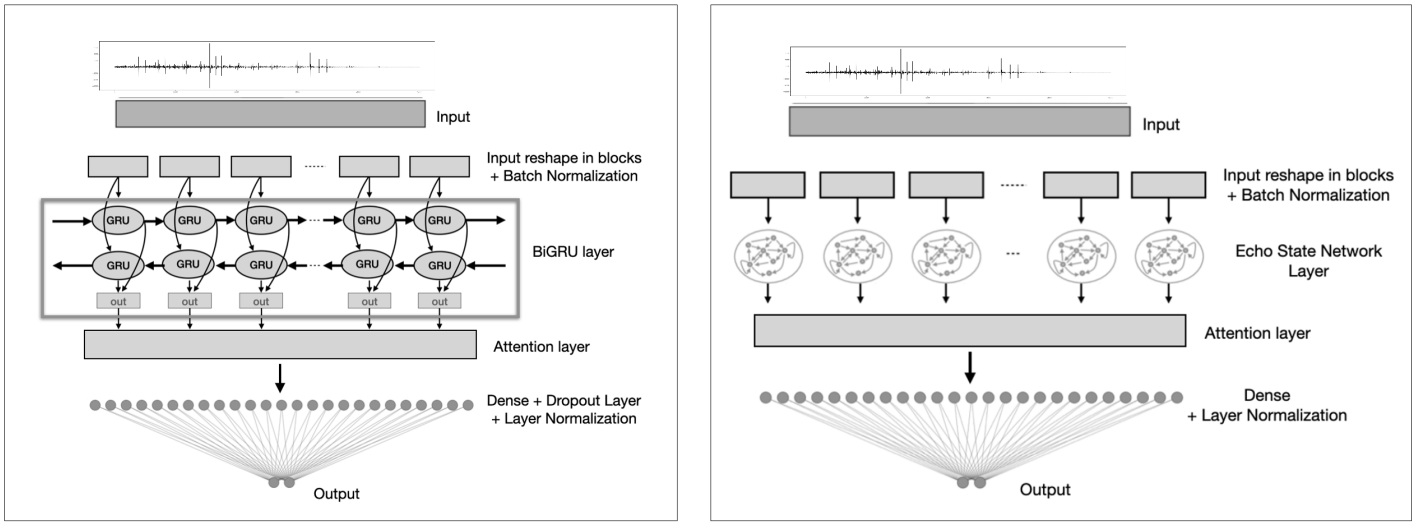


FIGURE 5.4 – **Architectures du RNN-BiGRU (gauche) et du ESN (droite)**

5.4.3 Les réseaux de neurones à convolution 2D pour les spectrogrammes

Les réseaux de neurones présentés dans cette section sont utilisés pour traiter des images, nécessitant donc une représentation bidimensionnelle des spectres MALDI-TOF. Les modèles de réseaux de neurones implémentés sont le réseau de neurones à convolution à deux dimensions pour les spectrogrammes (2DCNN spectrogram) et une version hybride de ce modèle avec une composante récurrente (2DCNN BiGRU - Hybrid spectrogram).

Le pré-traitement des spectres

Le prétraitement débute par la soustraction de la ligne de base en utilisant la méthode des moindres carrés asymétriques, suivie d'une interpolation à un espacement de 4, réduisant ainsi la taille du spectre initial de 20 000 à 5 000 points (comme mentionné précédemment avec le spectre de référence). Ensuite, le spectre subit une dérivation puis est traité par la méthode de la transformée de Fourier à court terme (STFT - Short Time Fourier Transform) (Zhang et al., 2007). Par la suite, il est transformé en un spectrogramme à l'aide du package Librosa en Python, puis normalisé et converti en une image en niveaux de gris avec le package cv2. Le spectre, initialement un signal unidimensionnel de 20 000 points, est ainsi transformé en une image bidimensionnelle en niveaux de gris de taille 128x128 (Voir la figure 5.5 ci-dessous).

La transformée de Fourier joue un rôle fondamental dans l'analyse des spectres MALDI-TOF. Elle est utilisée pour extraire des informations essentielles, telles que les masses moléculaires, la composition chimique, et d'autres caractéristiques des ions présents dans les échantillons (Rockwood, 2003 ; Haegler et al., 2009). De plus, elle est reconnue pour sa capacité à réduire le bruit, la distinguant ainsi de la méthode de lissage (Zhang et al., 2007). Grâce à la conversion du signal temporel en un signal fréquentiel, la méthode STFT (Short-Time Fourier Transform) combine les informations temporelles pour générer des spectrogrammes (Wyse et al., 2017). Ces spectrogrammes offrent une représentation visuelle des données spectrales, où les variations d'intensité sont symbolisées par des couleurs allant du clair au foncé (sur l'axe vertical). Les transitions de couleur reflètent les changements d'intensité des pics, tandis que l'axe horizontal indique la distribution temporelle des ions ainsi que les informations liées aux masses moléculaires.

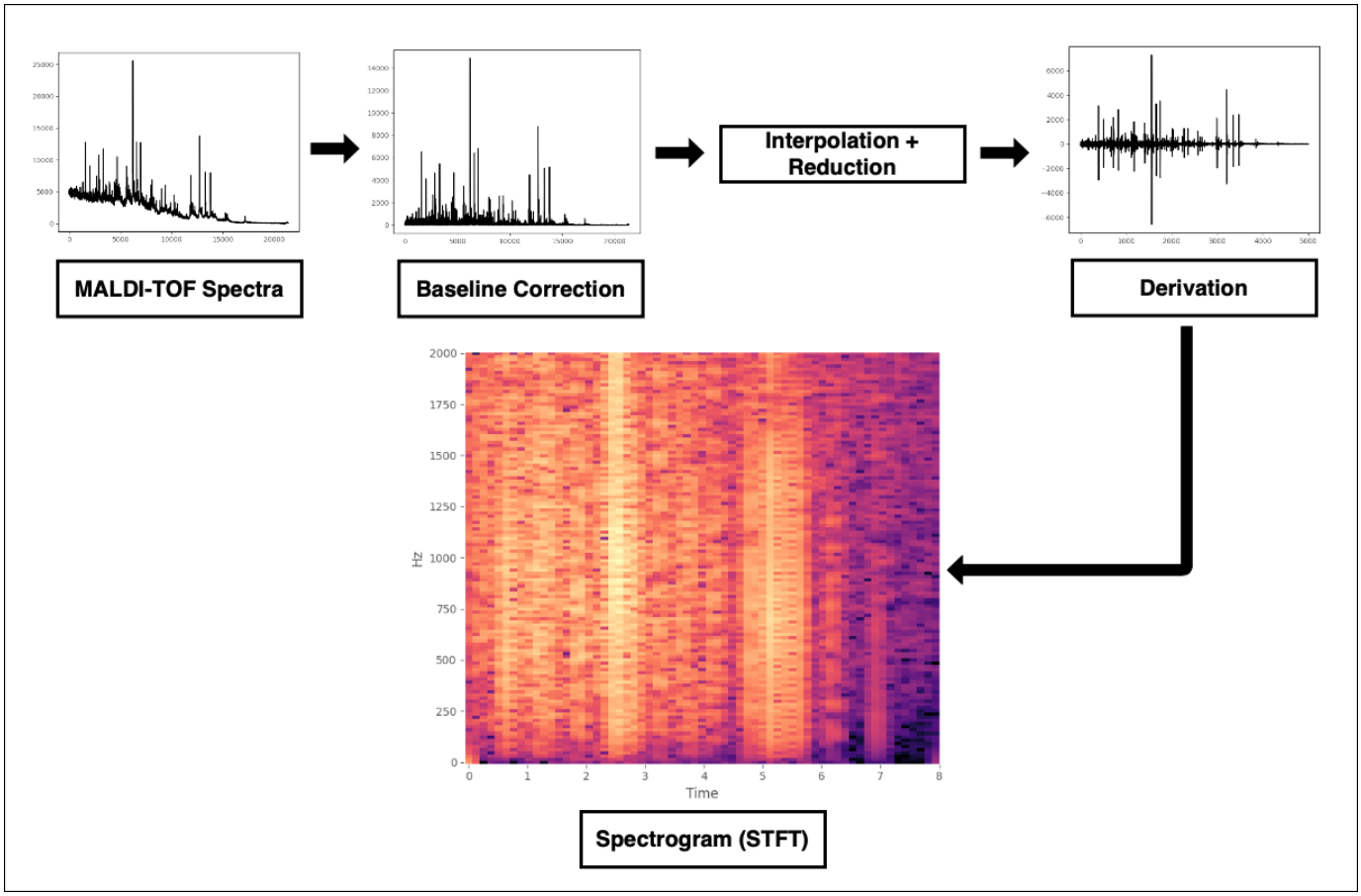
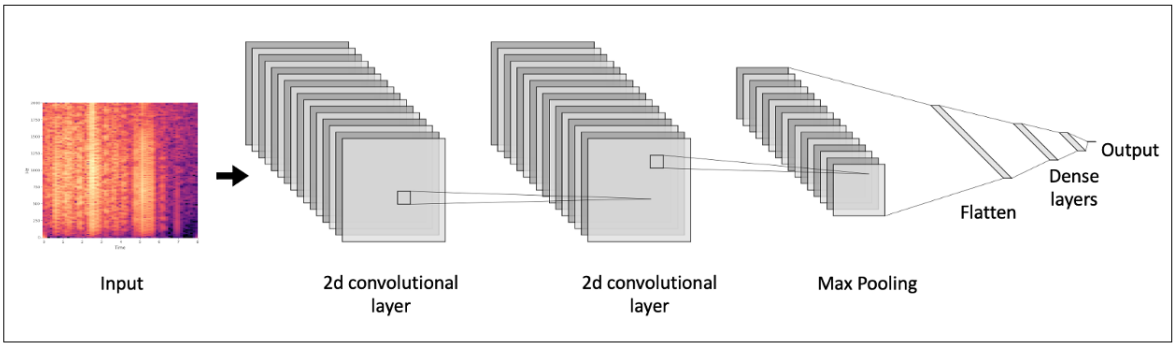


FIGURE 5.5 – **Pré-traitement des spectres sur 2DCNN spectrogram et 2DCNN BiGRU - Hybrid spectrogram.**

Les modèles

Pour le traitement efficace des spectrogrammes, un réseau de neurones convolutionnels en 2D (2DCNN), conçu initialement pour les images, est utilisé. Dans notre modèle, les variations de couleur, correspondant aux changements d'intensité des pics, ainsi que les informations locales dans les spectrogrammes, sont repérées et capturées par le 2DCNN. Ces données extraites permettent la reconnaissance de similitudes entre les spectrogrammes, facilitant ainsi la classification ou la régression ultérieure. En résumé, les motifs spécifiques à différentes positions du spectrogramme sont identifiés par les filtres de convolution d'un 2DCNN, créant ainsi des cartes de caractéristiques qui aident le réseau à localiser et extraire des informations cruciales des données spectrales MALDI-TOF. Cela a pour résultat l'amélioration de diverses tâches de traitement et d'analyse (Huang et al., 2019).

En utilisant l'architecture précédente comme base, un modèle hybride a été créé, intégrant des couches de convolution ainsi que des couches récurrentes GRU bidirectionnels. Ces dernières sont utilisées pour traiter les caractéristiques temporelles du spectrogramme une fois que les caractéristiques spatiales ont été extraites par la couche de convolution. L'objectif de cette approche est de réduire les taux d'erreur par rapport aux modèles précédents en combinant à la fois des couches convolutionnelles et récurrentes (Voir les architectures des modèles ci-dessous 5.6).



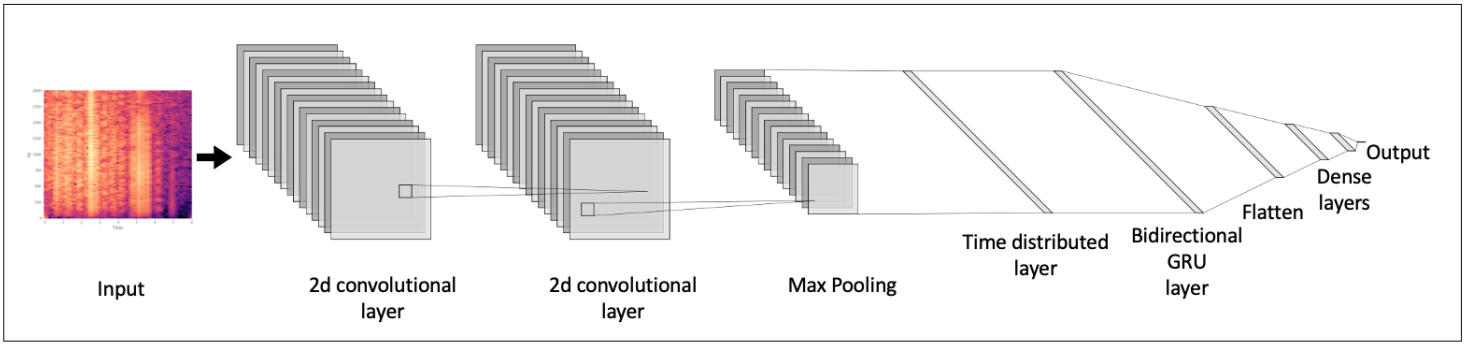


FIGURE 5.6 – **Architectures du 2DCNN spectrogram (haut) et du 2DCNN BiGRU - Hybrid spectrogram (bas)**

5.4.4 Les réseaux de neurones à convolution 2D pour les scalogrammes

De manière similaire, les réseaux de neurones de cette section sont employés pour le traitement de jeux d'images, impliquant ainsi la représentation bidimensionnelle des spectres MALDI-TOF en tant que données. Les modèles de réseaux de neurones utilisés comprennent les 2DCNN pour les scalogrammes (2DCNN scalogram) prenant en entrée un ensemble d'images, ainsi qu'une version hybride de ce modèle avec une composante récurrente (2DCNN BiGRU - Hybrid scalogram).

Le pré-traitement des spectres

Pour ce type de modèle, le prétraitement initial est semblable à celui précédemment décrit. Il débute par la soustraction de la ligne de base en utilisant la méthode des moindres carrés asymétriques, suivie d'une interpolation avec un espacement de 4 pour réduire la taille initiale du spectre de 20 000 à 5 000 points. Ensuite, le spectre est dérivé, puis, passe par la méthode de la transformée en ondelette continue (CWT - Continuous Wavelet Transform). Le signal obtenu est ensuite divisé en neuf parties, chacune des tranches du spectre est convertie en une image bidimensionnelle, puis transformée en une image RVB avant d'être normalisée à l'aide des packages Matplotlib et Skimage. Ainsi, le spectre, initialement un signal unidimensionnel de 20 000 points, est transformé en un ensemble de neuf images de taille 75x75 (Voir la figure 5.7 ci-dessous).

La transformée en ondelettes continues (CWT) suscite un vif intérêt dans de nombreuses études récentes en raison de sa précision et de sa capacité à gérer des informations à différentes échelles (Coombes et al., 2005, Yang et al., 2009). Elle se révèle particulièrement utile dans la détection et la sélection des pics dans les spectres de masse. De plus, elle est efficace pour réduire le bruit présent dans le signal du spectre de masse, ce qui la conduit à remplacer la méthode de lissage (Yang et al., 2009, Wijetunge et al. 2015). La CWT transforme le spectre en un scalogramme, qui peut être visualisé comme un tenseur tridimensionnel (Salles et al., 2023). Ce tenseur comporte des couches correspondant à des convolutions du signal à différentes échelles d'ondelettes, offrant ainsi une représentation multidimensionnelle des informations spectrales.

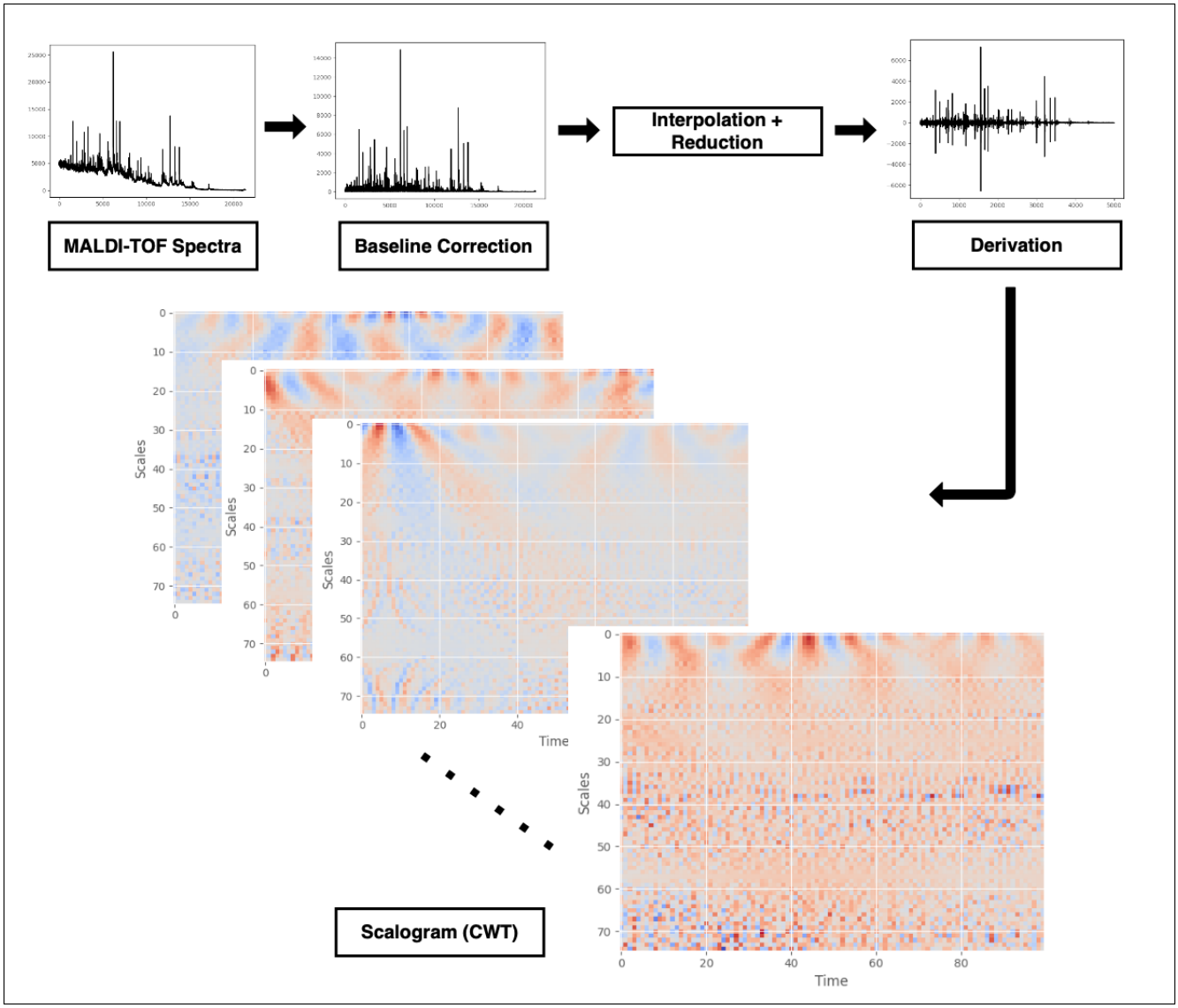
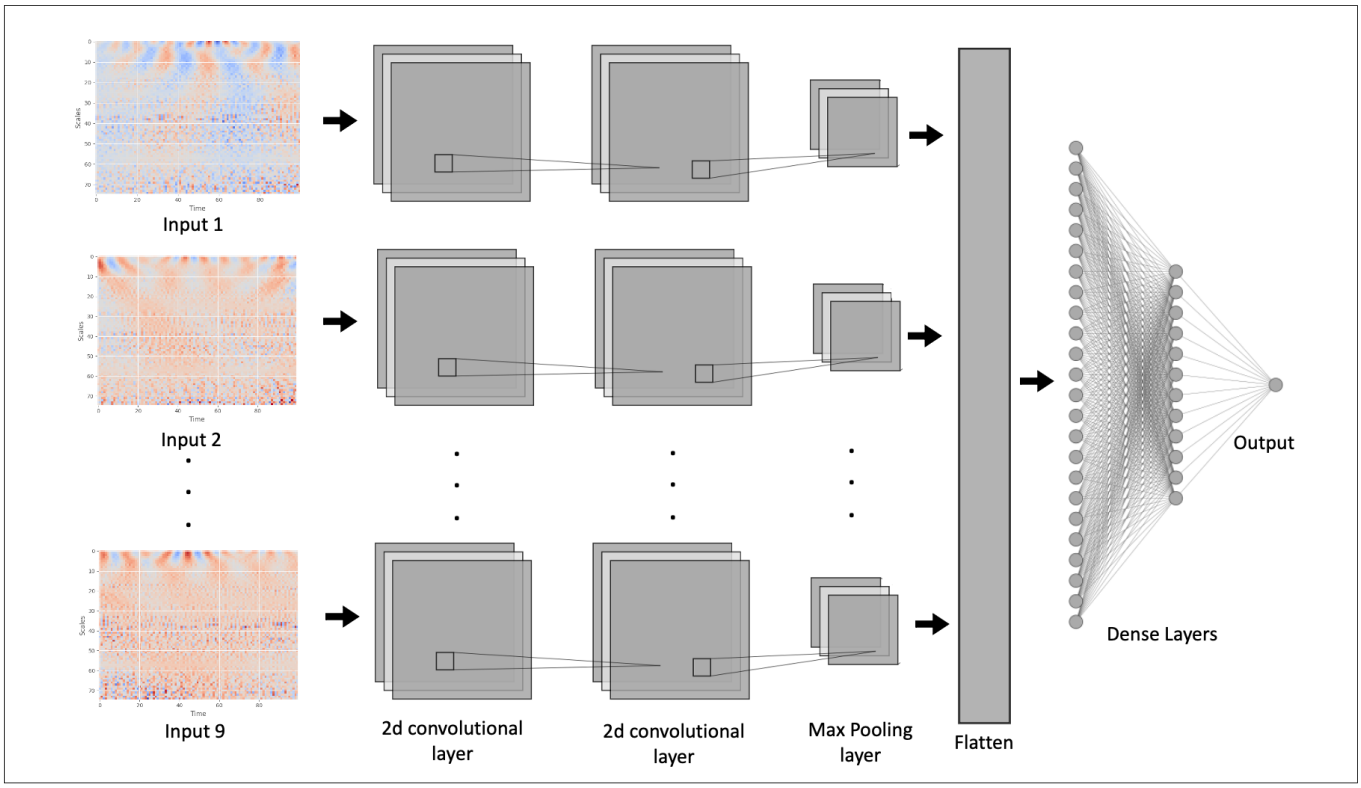


FIGURE 5.7 – **Pré-traitement des spectres sur 2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram.**

Les modèles

La représentation est analysée par un CNN 2D, qui capture les variations de couleur et de position en fonction des échelles de fréquence et du temps. Il filtre les motifs de couleurs générés par les différentes images du scalogramme, permettant d'identifier les informations relatives aux pics et aux variations d'intensité dans différentes zones du spectre. Cette approche facilite l'identification des similitudes ou des différences entre les spectres d'une même classe et contribue aux prédictions lors de l'entraînement. Le réseau de neurones applique une couche de convolution 2D à chaque tranche temporelle du spectre (représentée sous forme d'image), grâce à la couche TimeDistributed de Keras. Cette méthode permet un traitement et une extraction plus précise des caractéristiques de chaque tranche temporelle. En cumulant toutes ces extractions, le modèle dispose de davantage d'informations par rapport à l'utilisation du spectrogramme seul. Toutes les informations extraites sont ensuite traitées directement par la partie entièrement connectée du réseau.
En résumé, cette approche combine une analyse fine des données temporelles avec une extraction efficace des caractéristiques spatiales du scalogramme, ce qui est bénéfique pour les applications de classification et de prédiction (Bernitsas, 2021).

Un modèle hybride a également été développé en utilisant le modèle précédent, auquel une couche récurrente BiGRU a été ajoutée pour améliorer la précision des prédictions en tirant parti de l'information de la couche récurrente bidirectionnelle (Voir les architectures des modèles ci-dessous 5.8).



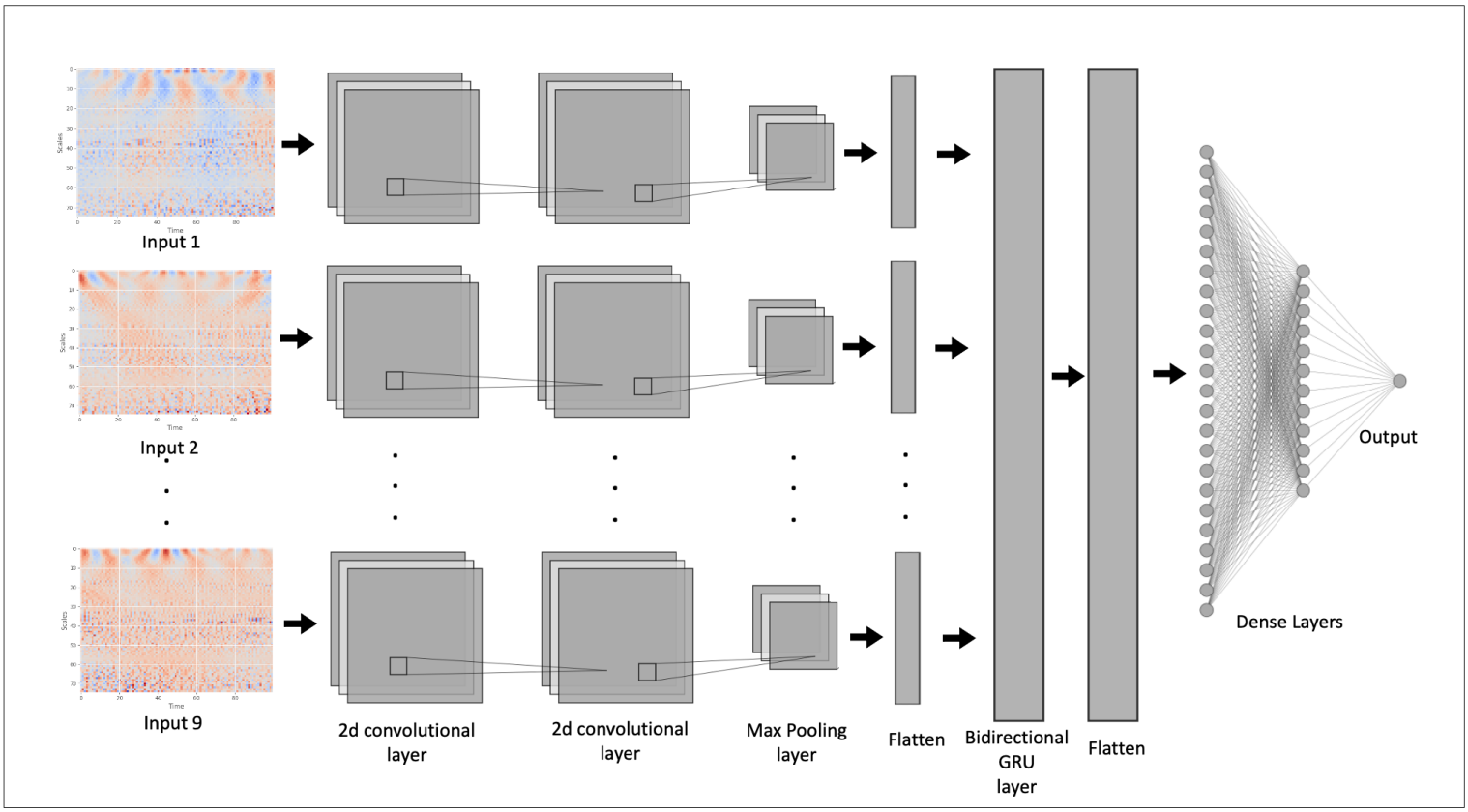


FIGURE 5.8 – **Architectures du 2DCNN scalogram (haut) et du 2DCNN BiGRU - Hybrid scalogram (bas)**

5.4.5 Les Autoencodeurs

Dans cette section, nous explorons les Autoencodeurs profonds (Bank et al., 2020), communément désignés sous l'acronyme DAE (Deep Auto Encoder). Il s'agit de réseaux de neurones artificiels qui opèrent de manière semi

supervisée. Leur rôle central est de réduire la redondance des données d'entrée, éliminer le bruit éventuel de ces données, et acquérir des caractéristiques avancées et abstraites des spectres. Les séquences générées à partir des spectres réduits des ensembles d'apprentissage et de test sont ensuite utilisées pour l'entraînement de différents modèles d'apprentissage automatique classiques, à la fois pour la classification et la régression.

Les Autoencodeurs évalués comprennent un Autoencodeur à couches entièrement connectées (DAE : Deep Auto Encoder) et un Autoencodeur à couches de convolution (DCAE : Deep Convolutional Auto Encoder).

Le pré-traitement des spectres

Un lissage du spectre brut a été réalisé en utilisant la méthode de la moyenne mobile, suivi d'une soustraction de la ligne de base à l'aide de la méthode des moindres carrés asymétriques. Ensuite, une interpolation a été effectuée avec un espacement de 5 pour réduire la taille du spectre initial de 20 000 à 4 000 points, suivi d'une étape de normalisation (Voir la figure 5.9 ci-dessous).

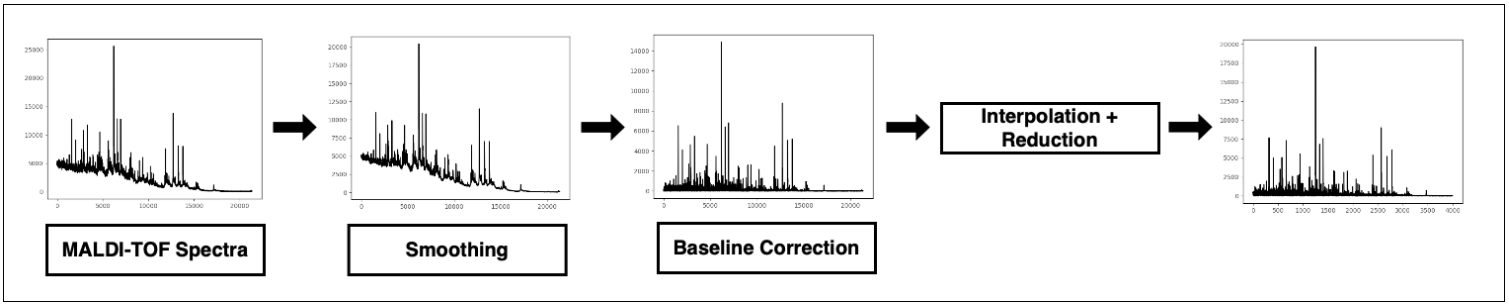


FIGURE 5.9 – **Pré-traitement des spectres sur les Autoencodeurs DAE et DCAE.**

Les modèles

Ce type de réseau de neurones se compose de deux composantes principales : l'encodeur et le décodeur. L'encodeur est responsable de l'extraction des éléments discriminants, de la capture des caractéristiques spécifiques du spectre et de l'évaluation du bruit éventuel. La sortie de cette partie a une dimension plus petite que celle de départ, ce qui permet de conserver un nombre d'informations sur le spectre dans une dimension réduite, appelée "bottleneck". À la sortie du bottleneck, les spectres représentés sous forme de séquences d'une centaine d'éléments sont ensuite exploités par divers modèles d'apprentissage automatique classiques pour l'entraînement en classification et en régression. Parmi ces modèles, on retrouve les K plus proches voisins (KNN), XGBoost, le classifieur Extra Trees, les forêts aléatoires, le classifieur Gradient Boosting, l'algorithme Naïf Bayésien Gaussien, LightGBM basé sur Gradient Boosting Machine, l'analyse discriminante linéaire, l'analyse discriminante quadratique, la machine à vecteurs de support (SVM) et la régression linéaire. La deuxième partie de ce modèle, le décodeur, n'a pas été exploitée dans notre étude, mais elle pourrait éventuellement servir à reconstruire le spectre sous une forme débruitée si nécessaire (Li et al., 2022).

Le premier Autoencodeur est un modèle à couches entièrement connectées (DAE fc), qui est utilisé pour traiter l'ensemble du spectre en reliant tous les neurones (Li et al., 2022). En revanche, le deuxième Autoencodeur utilise des couches de convolution (DCAE), pour réduire efficacement la dimension en employant des opérations de convolution et des filtres caractéristiques. Grâce à cette approche, une représentation latente des données MALDI-TOF peut être apprise, et leur dimension est ainsi significativement réduite (Zhou et al., 2020) (Voir les architectures des modèles ci-dessous 5.10).

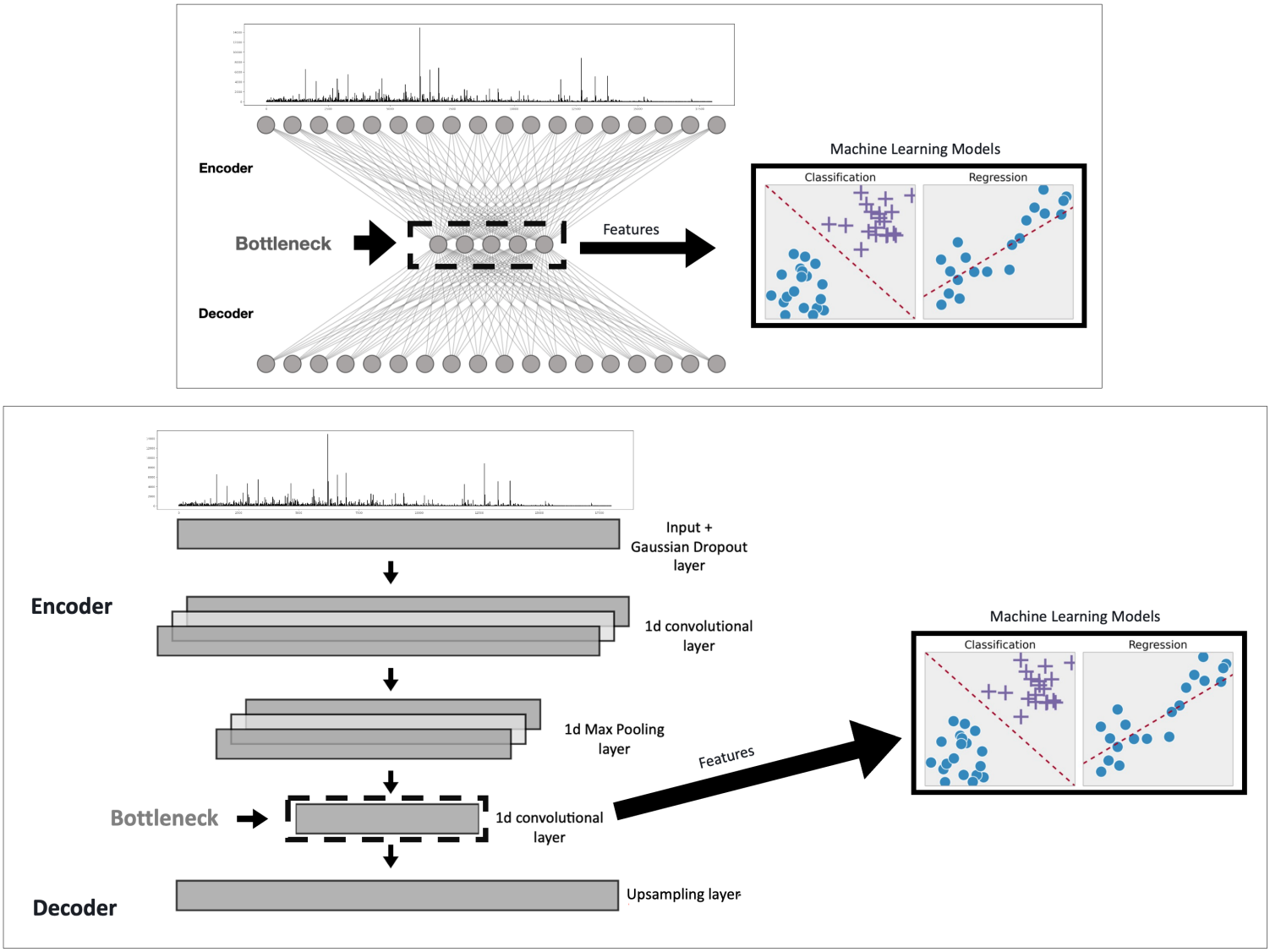


FIGURE 5.10 – **Architectures des Autoencodeurs DAE fc (haut) et DCAE (bas)**

5.4.6 Détails d'implémentation

Toutes les recherches de cette étude, depuis la collecte des spectres, ont été réalisées exclusivement sous Python. Chaque spectre a été considéré comme une entrée individuelle, mais pour éviter un surajustement indésirable, nous avons appliqué une stratégie de stratification. En particulier, nous avons stratifié les échantillons en fonction de la souche dans les données de MABSc, ce qui nous a permis de maintenir les distributions de classes d'une manière équilibrée. Pour les cohortes de moustiques, nous avons utilisé une stratification basée sur les spécimens. Cette approche nous a permis d'assigner les répliques d'un même spécimen de moustique (ou souche pour le MABSc) soit à l'ensemble d'entraînement (80 %) soit à l'ensemble de test (20 %), en utilisant la méthode "StratifiedKFold" du package Scikit Learn.

Les réseaux de neurones ont été implémentés avec l'aide de la librairie Tensorflow et des fonctions spécifiques de la librairie Keras tel que Keras-TCN pour le modèle TCN. Les expérimentations ont été réalisées en utilisant une machine équipée d'un GTX TITAN X Maxwell pour l'entraînement.
Tous les modèles ont été formés en utilisant l'optimiseur Adam. Pour les tâches de classification (binaire, trois classes et quatre classes), nous avons appliqué la fonction d'entropie croisée catégorielle (Categorical Cross-Entropy) à tous les modèles, à l'exception des Autoencodeurs. En ce qui concerne les tâches de régression, la fonction de perte utilisée sur tous les modèles, à l'exception des Autoencodeurs, était la fonction de Huber (Huber loss). Les Autoencodeurs, quant à eux, ont utilisé la fonction d'erreur quadratique moyenne (Mean Squared Error). Pour évaluer la perte, nous avons employé la fonction de perte de tous les modèles, sauf les Autoencodeurs, et avons mesuré l'exactique catégorielle (Categorical Accuracy) pour les problèmes de classification et l'erreur moyenne absolue (Mean Absolute Error) pour les problèmes de régression.

Certains modèles ont été formés en utilisant un générateur, une approche efficace pour gérer le déséquilibre des classes tout en augmentant l'échantillonnage pendant l'entraînement, ce qui est particulièrement utile pour les données en quantité limitée. De plus, l'utilisation du générateur permet de contrôler la quantité de données alimentée dans le modèle, évitant ainsi une surcharge de la mémoire de calcul et prévenant le surapprentissage. L'entraînement a été optimisé en utilisant deux techniques : l'arrêt précoce (EarlyStopping) avec une patience de 20 (Early Stopping), qui est un outil de régularisation pour éviter le surapprentissage, et la fonction de pondération des classes (Class Weigth), qui permet de traiter le déséquilibre des classes.
La taille des lots (Batch size) et le nombre d'époques (epochs) varient d'un modèle à l'autre en raison des exigences de prétraitement et de la gestion de la mémoire.

Nous avons utilisé le traqueur de carbone (CarbonTracker), un outil de suivi et de prévision de la consommation d'énergie et de l'empreinte carbone des modèles d'apprentissage profond (Anthony et al., 2020).

Les performances des modèles sont évaluées en utilisant plusieurs métriques (Voir la section 2.5 pour les définitions des métriques), notamment l'accuracy, le score F1, la précision, le rappel, le score AUC ROC et l'accuracy équilibrée (balanced accuracy) pour la classification et l'erreur moyenne absolue (MAE), l'AUROC et le R^2 pour la régression. Dans ce chapitre, nous simplifierons la présentation en ne montrant que l'accuracy équilibrée (Balanced Accuracy), car elle est mieux adaptée aux jeux de données déséquilibrés. Elle prend en compte la répartition des classes, offrant ainsi une évaluation plus équilibrée des performances du modèle. Cela la rend plus efficace pour repérer les erreurs de classification, en particulier pour les classes minoritaires, et plus pertinente dans de nombreux contextes d'application. Pour la régression nous présentons la MAE. Les performances des autres métriques sont disponibles en Annexe C (Tables S2). Pour toutes ces métriques, nous fournissons un intervalle de confiance à 95 % en utilisant la méthode bootstrap empirique (Dekking et al., 2005, page 275).

5.5 Résultats

5.5.1 Performances d'identification et de prédiction

Les tableaux 5.2, 5.3 et 5.4 présentent les performances des modèles sur les différentes cohortes décrites précédemment.

Dans l'ensemble, le RNN-BiGRU et le TCN affichent d'excellentes performances sur toutes les cohortes. Ce sont les modèles qui se distinguent en termes de classification et de prédiction. Pour l'identification des espèces d'anophèles par classification, les performances sont comparables, avec une accuracy équilibrée de 92 à 93 %, 81 % , et 84 à 85 % respectivement pour les pattes, les têtes et le thorax. En ce qui concerne la prédiction de l'âge par régression, le TCN se démarque avec une erreur moyenne absolue (MAE) allant de 1.94 pour le thorax à 2.49 pour les pattes. Pour la cohorte MABSc, le TCN est également le modèle le plus performant, atteignant une accuracy équilibrée de 97 % pour distinguer les souches résistantes des sensibles. Cependant, le RNN-BiGRU se distingue en obtenant la meilleure performance parmi tous les autres modèles des tableaux 5.2, 5.3 et 5.4, avec une accuracy équilibrée de 95 % pour l'identification des sous-espèces.

Les deux principaux modèles en tête de la compétition sont suivis par le 1DCNN, qui affiche également des performances satisfaisantes. Il excelle particulièrement dans la cohorte des anophèles en termes de régression, avec une MAE allant de 1,92 pour le thorax à 2,27 pour les pattes. De plus, il atteint une accuracy équilibrée de 99 % dans la cohorte MABSc pour la détection de la résistance. Ce modèle se distingue en étant le meilleur parmi tous les autres modèles sur ces deux dernières cohortes. Cependant, en ce qui concerne l'identification des espèces de moustiques, ses performances sont similaires, voire inférieures à celles des deux modèles précédents.

En ce qui concerne les Autoencodeurs des tableaux 5.3 et 5.4, certains modèles d'apprentissage automatique classique parviennent à identifier ou prédire avec des résultats équivalents à ceux du TCN et du RNN (Tableau 5.2). Parmi eux, on trouve le Gradient Boosting, suivi du LightGBM et du Linear Discriminant Analysis, qui affichent des performances comparables dans le tableau 5.3. Cependant, aucun modèle associé à l'Autoencodeur DAE fc n'arrive à classer les sous-espèces de la cohorte MABSc avec des performances dépassant les 68 %.

Il est à noter que le LightGBM utilisé avec l'Autoencodeur DCAE (Tableau 5.4) présente les meilleures performances par rapport à tous les autres modèles du Tableau 5.4, avec des résultats allant de 86 % sur le thorax à 90

% sur les pattes dans la cohorte Anophèles espèces. En ce qui concerne la régression avec l'âge des moustiques, le DCAE associé au LightGBM a la meilleure (MAE) parmi tous les autres modèles utilisés avec les Autoencodeurs des Tableaux 5.3 et 5.4, variant de 2.47 pour les pattes à 3.16 pour les têtes. Sur la cohorte MABSc, l'identification des sous-espèces et la détection de la résistance atteignent respectivement 87 % et 77 %, ce qui est satisfaisant mais ne dépasse pas les performances du RNN-BiGRU (Tableau 5.2).

Les performances sont presque équivalentes entre le LightGBM et les Random Forest du DCAE, bien que la classification des sous-espèces de MABSc atteigne 66 % de accuracy équilibrée. Il est intéressant de noter que le modèle XGBoost parvient, au contraire, à classer ces sous-espèces avec une accuracy équilibrée de 93 %, mais ses performances sont nettement inférieures aux autres modèles sur les autres cohortes, affichant une MAE de 4,61 pour l'âge des moustiques et une accuracy équilibrée de 27 à 32 % pour l'identification des espèces de moustiques ce qui est médiocre.

La plupart des autres modèles présentent des comportements similaires, c'est-à-dire qu'ils parviennent à classer ou prédire avec des performances satisfaisantes, mais cela ne s'applique pas à toutes les cohortes. Par exemple, l'ESN atteint des accuracy équilibrées de 82 % et 93 % dans la cohorte MABSc pour identifier les sous-espèces et la résistance, respectivement. Les modèles 2DCNN spectrogram, 2DCNN BiGRU - Hybrid spectrogram et 2DCNN scalogram affichent des performances équivalentes, atteignant même celles du RNN et du TCN, voire les dépassant avec 90 % sur le thorax dans la cohorte des espèces *Anopheles*. Cependant, leurs performances en termes de MAE pour la régression sont moins bonnes, et bien qu'ils atteignent 87 % pour l'identification de la résistance dans la cohorte MABSc, leurs performances globales dans cette cohorte sont moins bonnes. Les plus faibles performances sont observées avec les 2DCNN scalogram dans la cohorte MABSc, avec seulement 59 % et 50 % pour identifier les espèces et la résistance, respectivement. Ces résultats suggèrent que le modèle ne parvient pas à classer correctement (voir les détails des performances en Annexe C, Tables S2).

Le modèle 2DCNN BiGRU - Hybrid scalogram est en fin de course. Il atteint certes une accuracy équilibrée de 89 % sur les pattes des espèces *Anopheles*, mais ses performances sur les autres parties anatomiques sont inférieures à tous les autres modèles du Tableau 5.2 et n'excèdent pas celles des autres modèles sur les autres cohortes en termes de résultats. En ce qui concerne les Autoencodeurs, le modèle de régression linéaire utilisé pour estimer l'âge des moustiques avec la DCAE (Tableau 5.4) ne parvient pas à prédire avec précision, affichant une MAE allant de 7.89 sur les têtes à 16,23 sur le thorax.

TABLEAU 5.2 – **Performances des modèles sur les différentes cohortes.** La métrique utilisée sur la cohorte des MABSc pour l'identification des sous espèces et de la résistance ainsi que la cohorte des Anopheles espèces est l'accuracy équilibrée (balanced accuracy). Pour l'âge des moustiques il s'agit de l'erreur moyenne absolue (MAE).

	Anopheles 4 espèces			Anopheles âge			MABSc	
	Pattes	Tête	Thorax	Pattes	Tête	Thorax	3 sous espèces	R et S
1DCNN	**0.88** [*0.83,0.93*]	**0.80** [*0.76,0.85*]	**0.88** [*0.84,0.91*]	**2.27** [*2.0,2.54*]	**2.04** [*1.77,2.38*]	**1.92** [*1.67,2.17*]	**0.78** [*0.72,0.83*]	**0.99** [*0.98,1.0*]
TCN	**0.93** [*0.87,0.97*]	**0.81** [*0.76,0.86*]	**0.84** [*0.8,0.88*]	**2.49** [*2.19,2.84*]	**2.08** [*1.74,2.46*]	**1.94** [*1.65,2.23*]	**0.77** [*0.72,0.82*]	**0.97** [*0.93,1.0*]
RNN-BiGRU	**0.92** [*0.85,0.96*]	**0.81** [*0.77,0.85*]	**0.85** [*0.81,0.89*]	**2.77** [*2.39,3.14*]	**3.36** [*2.81,3.9*]	**2.80** [*2.27,3.34*]	**0.95** [*0.9,0.98*]	**0.83** [*0.79,0.87*]
ESN	0.84 [*0.75,0.92*]	0.84 [*0.8,0.89*]	0.84 [*0.79,0.87*]	2.99 [*2.65,3.33*]	3.35 [*2.87,3.78*]	2.85 [*2.44,3.31*]	0.82 [*0.76,0.87*]	0.93 [*0.9,0.96*]
2DCNN spectrogram	0.89 [*0.83,0.94*]	0.81 [*0.76,0.85*]	0.90 [*0.87,0.94*]	2.68 [*2.34,3.03*]	2.56 [*2.18,2.93*]	2.43 [*2.04,2.81*]	0.60 [*0.57,0.63*]	0.87 [*0.83,0.91*]
2DCNN BiGRU - Hybrid spectrogram	0.91 [*0.86,0.96*]	0.83 [*0.8,0.88*]	0.89 [*0.86,0.93*]	3.24 [*2.79,3.74*]	3.86 [*3.35,4.4*]	3.57 [*2.97,4.24*]	0.59 [*0.56,0.63*]	0.50 [*0.5,0.5*]
2DCNN scalogram	0.91 [*0.85,0.96*]	0.83 [*0.79,0.88*]	0.83 [*0.77,0.89*]	3.09 [*2.75,3.51*]	3.68 [*3.18,4.2*]	2.33 [*1.99,2.73*]	0.74 [*0.69,0.79*]	0.84 [*0.79,0.88*]
2DCNN BiGRU - Hybrid scalogram	0.89 [*0.83,0.95*]	0.79 [*0.74,0.84*]	0.76 [*0.69,0.83*]	3.25 [*2.9,3.66*]	3.99 [*3.12,4.91*]	2.20 [*1.85,2.58*]	0.82 [*0.77,0.87*]	0.76 [*0.7,0.84*]

TABLEAU 5.3 – **Performances de l'Autoencodeur DAE fc.** La métrique utilisée sur la cohorte des MABSc pour l'identification des sous espèces et de la résistance ainsi que la cohorte des Anopheles espèces est l'accuracy équilibrée (balanced accuracy). Pour l'âge des moustiques il s'agit de l'erreur moyenne absolue (MAE).

	Anopheles 4 espèces			Anopheles âge			MABSc	
DAE fc	Pattes	Tête	Thorax	Pattes	Tête	Thorax	3 sous espèces	R et S
KNeighborsClassifier/Regressor	0.86 [*0.8,0.92*]	0.83 [*0.79,0.87*]	0.91 [*0.88,0.95*]	3.25 [*2.85,3.7*]	2.97 [*2.55,3.4*]	2.26 [*1.82,2.78*]	0.65 [*0.64,0.67*]	0.59 [*0.51,0.68*]
XGBClassifier/Regressor	0.25 [*0.25,0.25*]	0.25 [*0.25,0.25*]	0.24 [*0.24,0.25*]	3.34 [*2.93,3.74*]	3.56 [*3.02,4.08*]	3.23 [*2.68,3.81*]	0.65 [*0.62,0.69*]	0.65 [*0.59,0.73*]
ExtraTreesClassifier/Regressor	0.93 [*0.89,0.98*]	0.82 [*0.78,0.87*]	0.86 [*0.83,0.91*]	2.64 [*2.33,2.97*]	3.39 [*2.93,3.83*]	2.25 [*1.89,2.64*]	0.65 [*0.64,0.67*]	0.59 [*0.51,0.67*]
RandomForestClassifier/Regressor	0.91 [*0.87,0.97*]	0.82 [*0.78,0.87*]	0.86 [*0.82,0.9*]	2.50 [*2.23,2.8*]	3.69 [*3.18,4.27*]	2.96 [*2.43,3.58*]	0.65 [*0.64,0.67*]	0.67 [*0.6,0.75*]
GradientBoostingClassifier/Regressor	**0.95** [*0.9,0.99*]	**0.78** [*0.74,0.83*]	**0.88** [*0.84,0.92*]	**2.78** [*2.48,3.1*]	**3.31** [*2.84,3.83*]	**2.67** [*2.22,3.18*]	**0.65** [*0.64,0.67*]	**0.88** [*0.82,0.93*]
LightGBMClassifier/Regressor	**0.93** [*0.88,0.98*]	**0.80** [*0.76,0.85*]	**0.91** [*0.87,0.94*]	**2.46** [*2.15,2.77*]	**3.52** [*3.09,3.99*]	**3.07** [*2.48,3.68*]	**0.66** [*0.67,0.67*]	**0.78** [*0.72,0.86*]
GaussianNB	0.65 [*0.55,0.75*]	0.60 [*0.51,0.69*]	0.56 [*0.51,0.63*]				0.59 [*0.55,0.63*]	0.72 [*0.67,0.77*]
LinearDiscriminantAnalysis	**0.91** [*0.86,0.96*]	**0.84** [*0.8,0.89*]	**0.90** [*0.87,0.94*]				**0.68** [*0.67,0.7*]	**0.79** [*0.73,0.85*]
QuadraticDiscriminantAnalysis	0.48 [*0.46,0.5*]	0.58 [*0.54,0.63*]	0.55 [*0.51,0.59*]				0.51 [*0.46,0.56*]	0.62 [*0.55,0.69*]
SVC	0.89 [*0.84,0.95*]	0.80 [*0.76,0.85*]	0.85 [*0.81,0.89*]				0.65 [*0.63,0.67*]	0.74 [*0.67,0.82*]
Linear Regression				3.28 [*2.96,3.62*]	2.93 [*2.43,3.42*]	2.78 [*2.39,3.23*]		

TABLEAU 5.4 – **Performances de l'Autoencodeur DCAE.** La métrique utilisée sur la cohorte des MABSc pour l'identification des sous espèces et de la résistance ainsi que la cohorte des Anopheles espèces est l'accuracy équilibrée (balanced accuracy). Pour l'âge des moustiques il s'agit de l'erreur moyenne absolue (MAE).

	Anopheles 4 espèces			Anopheles âge			MABSC	
DCAE	Pattes	Tête	Thorax	Pattes	Tête	Thorax	3 sous espèces	R et S
KNeighborsClassifier/Regressor	0.84 [*0.78,0.9*]	0.80 [*0.76,0.85*]	0.85 [*0.81,0.89*]	3.28 [*2.87,3.72*]	3.00 [*2.47,3.54*]	2.84 [*2.43,3.23*]	0.58 [*0.54,0.62*]	0.75 [*0.68,0.81*]
XGBClassifier/Regressor	0.32 [*0.28,0.39*]	0.27 [*0.26,0.3*]	0.87 [*0.82,0.91*]	3.30 [*2.86,3.76*]	4.61 [*4.1,5.14*]	3.25 [*2.64,3.96*]	0.93 [*0.89,0.97*]	0.72 [*0.64,0.8*]
ExtraTreesClassifier/Regressor	0.90 [*0.84,0.95*]	0.82 [*0.78,0.87*]	0.86 [*0.83,0.9*]	2.81 [*2.5,3.15*]	3.56 [*3.18,4.02*]	2.44 [*2.12,2.81*]	0.56 [*0.52,0.61*]	0.79 [*0.72,0.86*]
RandomForestClassifier/Regressor	0.90 [*0.84,0.95*]	0.84 [*0.8,0.89*]	0.87 [*0.83,0.91*]	2.54 [*2.14,2.97*]	3.72 [*3.31,4.18*]	2.79 [*2.29,3.35*]	0.66 [*0.67,0.67*]	0.78 [*0.71,0.85*]
GradientBoostingClassifier/Regressor	0.88 [*0.82,0.93*]	0.83 [*0.79,0.88*]	0.83 [*0.79,0.88*]	2.58 [*2.22,2.93*]	3.10 [*2.72,3.52*]	2.60 [*2.16,3.09*]	0.68 [*0.63,0.75*]	0.80 [*0.73,0.88*]
LightGBMClassifier/Regressor	**0.90** [*0.84,0.95*]	**0.87** [*0.83,0.91*]	**0.86** [*0.82,0.90*]	**2.47** [*2.1,2.82*]	**3.16** [*2.77,3.57*]	**2.52** [*2.07,3.05*]	**0.87** [*0.82,0.92*]	**0.77** [*0.7,0.85*]
GaussianNB	0.76 [*0.67,0.84*]	0.61 [*0.57,0.67*]	0.61 [*0.55,0.68*]				0.50 [*0.46,0.56*]	0.83 [*0.78,0.89*]
LinearDiscriminantAnalysis	0.84 [*0.78,0.91*]	0.76 [*0.72,0.81*]	0.82 [*0.78,0.86*]				0.70 [*0.65,0.77*]	0.62 [*0.54,0.71*]
QuadraticDiscriminantAnalysis	0.30 [*0.23,0.38*]	0.25 [*0.2,0.3*]	0.39 [*0.36,0.42*]				0.33 [*0.33,0.33*]	0.50 [*0.5,0.5*]
SVC	0.88 [*0.82,0.93*]	0.74 [*0.7,0.79*]	0.84 [*0.8,0.88*]				0.51 [*0.46,0.56*]	0.75 [*0.68,0.83*]
Linear Regression				9.79 [*7.84,12.24*]	7.89 [*6.77,9.09*]	16.23 [*12.68,20.06*]		

5.5.2 Temps d'execution et consommation en énergie lors de l'entraînement des modèles

La figure 5.11 présente le temps d'exécution des modèles. Pour simplifier la présentation et la comparaison des résultats, les temps de calcul lors de l'entraînement des pattes pour les cohortes Anopheles âge et Anopheles espèces sont affichés. Les exécutions par parties anatomiques ont été réalisées indépendamment à chaque fois, mais les temps de calcul sont équivalents pour chaque cohorte (voir en Annexe C Tables S2 le détail des résultats pour les autres parties anatomiques). En ce qui concerne les Autoencodeurs (DAE fc et DCAE), seul le temps de calcul de la réduction de dimension est présenté, et le temps de calcul est interrompu avant l'envoi des séquences réduites aux modèles d'apprentissage classique. Il est à noter que le temps requis pour les modèles classiques est très court, généralement inférieur à 5 secondes (voir les résultats en Annexe C Tables S2).

Dans l'ensemble, les modèles 2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram nécessitent significativement plus de temps d'exécution lors de l'entraînement, car ils traitent des images, ce qui implique une recherche de caractéristiques plus intensive sur des données bidimensionnelles. De plus, ces modèles travaillent sur des motifs abstraits, liés aux variations de couleurs plutôt qu'à des motifs clairement définis, ce qui explique le temps nécessaire à la recherche de ces modèles.

Le 2DCNN demande plus de temps encore, car il traite simultanément neuf images par spectre, entraînant un traitement d'informations plus complexe. Les RNN nécessitent également plus de temps que les CNN lors de l'entraînement, en grande partie en raison de la complexité de la composition des couches récurrentes.

Le temps de calcul observé pour la réduction de dimension des Autoencodeurs, notamment avec la composition des couches entièrement connectées du DAE fc et les convolutions réalisées dans le DCAE, s'explique par les opérations complexes impliquées dans ces modèles. Finalement, les 1DCNN, le TCN et l'ESN sont les trois modèles qui présentent un temps d'exécution équivalent entre eux et se distinguent par des temps de calcul nettement inférieurs par rapport aux autres modèles testés.

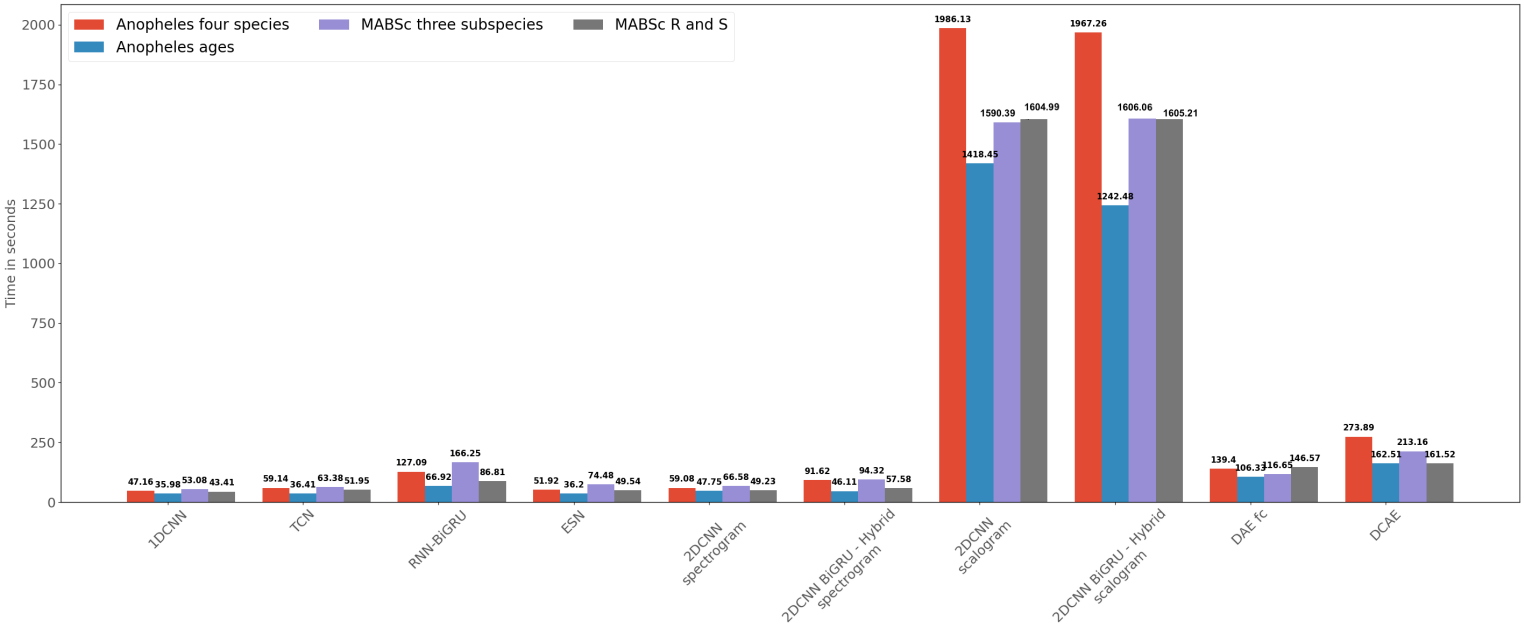


FIGURE 5.11 – **Temps de calculs des modèles** Pour simplifier la présentation des résultats nous nous contentons des présenter le temps de calcul lors de l'entraînement des pattes pour les cohortes des Anopheles âges et Anopheles espèces. Les exécutions par parties anatomiques ont été faites indépendamment à chaque fois mais les temps de calculs sont équivalents pour chaque cohorte (Voir l'Annexe C Tables S2 le détail des résultats pour les autres parties anatomiques).

La figure 5.12 présente la consommation d'énergie en kilowattheures lors de l'entraînement des modèles. L'allure des diagrammes en barre est similaire à celle de la figure 5.11. Il y a en effet une forte corrélation entre les résultats du temps de calcul et la consommation d'énergie lors de l'entraînement.

On remarque dans cette figure que les modèles 2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram consomment plus d'énergie que les autres modèles lors de l'entraînement, suivis des Autoencodeurs DAE fc et DCAE. La consommation d'énergie observée peut s'expliquer par les mêmes raisons que celles évoquées précédemment, à savoir la nature des données, la complexité du modèle, et le processus d'entraînement.

Également dans cette figure, on retrouve les 1DCNN, le TCN et l'ESN comme les modèles qui consomment le moins d'énergie. En revanche, le RNN-BiGRU présente une consommation d'énergie nettement plus importante.

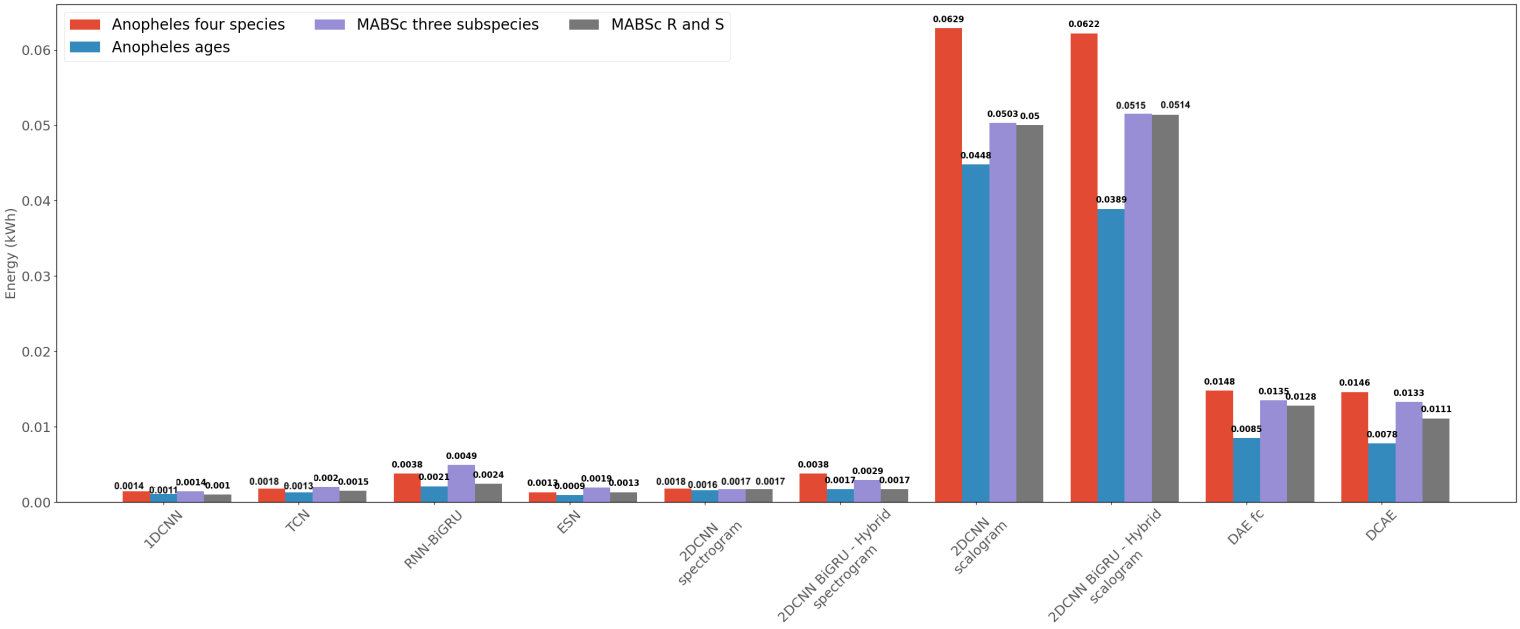


FIGURE 5.12 – **Consommation en énergie en Kilowatt-heure (kWh) des modèles** Pour simplifier la présentation des résultats nous nous contentons des présenter la consommation en énergie (kWh) lors de l'entraînement des pattes pour les cohortes des Anopheles âges et Anopheles espèces. Les exécutions par parties anatomiques ont été faites indépendamment à chaque fois mais les consommations en énergie sont équivalentes pour chaque cohortes (Voir l'Annexe C Tables S2 le détail des résultats pour les autres parties anatomiques).

5.6 Discussion

5.6.1 Points forts de l'étude

Cette étude explore de nouvelles possibilités en utilisant des modèles inexplorés pour traiter des données MALDI-TOF, telles que les réseaux de neurones récurrents, les convolutions bidimensionnelles avec l'utilisation de spectrogrammes et de scalogrammes, ainsi que les autoencodeurs. L'originalité de ce travail était de considérer les spectres de masse MALDI-TOF comme des données temporelles présentant des similitudes potentielles avec les signaux audio (Lemaire et al., 2019). Pour évaluer la capacité de généralisation des réseaux de neurones, nous avons utilisé des cohortes distinctes pour les ensembles d'apprentissage et de test indépendant.

De plus, cette étude apporte des éléments cruciaux pour l'intégration des modèles d'apprentissage statistique profonds dans les pratiques de laboratoire. Nous avons démontré que l'utilisation de modèles de convolution tels que les 1DCNN, le TCN, ou même un réseau récurrent comme le RNN-BiGRU, s'avèrent robustes pour résoudre diverses problématiques d'identification et de caractérisation. Dans nos expériences, ces modèles ont atteint de très fortes performances, atteignant 95 % pour l'identification des sous-espèces de *Mycobacterium abscessus* avec le réseau récurrent, et 97 % et 99 % pour l'identification de la résistance avec le 1DCNN. Ces résultats sont intéressants compte tenu de la forte similarité spectrale entre les souches de sous-espèces de *Mycobacterium abscessus* et du déséquilibre des classes dans la base de données. Ces modèles sont les seuls à atteindre de telles performances parmi les autres, ce qui suggère qu'ils peuvent surmonter le déséquilibre des classes, un problème courant dans les banques de spectres utilisées en laboratoire.

Nous avons observé qu'il était possible de prédire avec une précision moyenne de moins de deux jours l'âge des moustiques provenant de différentes origines et présentant une variabilité génétique et technique (Voir Chapitre précédent) en utilisant le CNN et le TCN (avec une MAE de 1,92 et 1,94, respectivement). Ces deux modèles se sont avérés être les plus performants pour la régression parmi toutes les options envisagées. Ces résultats revêtent une importance particulière dans le domaine entomologique pour la surveillance de l'épidémiologie du paludisme, car les méthodes actuelles sont imprécises et chronophages, comme exposé dans le chapitre précédent.
La précision de la prédiction de l'âge en jours, combinée à l'efficacité en termes de temps de calcul et de consommation énergétique, indique que l'utilisation en routine de laboratoire du 1DCNN, du TCN ou du RNN-BiGRU est envisageable, bien que ce dernier nécessite légèrement plus de temps de calcul et d'énergie. De surcroît, les cohortes ont été constituées par divers opérateurs lors de la préparation et de l'acquisition des spectres. Dans ce contexte, les performances remarquées sur les trois modèles mentionnés précédemment mettent en évidence la robustesse et la capacité de ces modèles à s'adapter à des données diverses.

Globalement, à des exceptions près, tous les modèles parviennent à classer ou prédire sur l'ensemble des cohortes, démontrant ainsi leur potentiel pour identifier des microorganismes ou caractériser des espèces vectrices de maladies à partir de spectres MALDI-TOF.
Cependant, les modèles à convolution bidimensionnelle (2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram) ainsi que les autoencodeurs (DAE fc et DCAE) sont moins favorisés en raison de leur temps d'exécution plus long, de leur consommation énergétique élevée, et de leurs performances inférieures, en particulier pour les modèles de convolution bidimensionnelle. En raison de leur architecture, ces modèles requièrent des ajustements pour gérer les données, les transformer en images, et incorporer des couches dédiées pour réduire la taille des données. Ces solutions permettraient d'optimiser les performances prédictives, réduire le temps de calcul, et économiser l'énergie requis lors de l'entraînement.

Au niveau des différentes problématiques abordées, les modèles testés ont révélé qu'il était possible d'identifier les sous-espèces de *Mycobacterium abscessus* en utilisant un ensemble d'apprentissage composé de seulement 814 spectres. Ce nombre de spectres semble suffisant pour atteindre des performances satisfaisantes dans l'identification des sous-espèces et la détection de la résistance.

Concernant l'identification des espèces étroitements liées de moustiques, on obtient des performances équivalentes entre le CNN, le TCN et les RNN-BiGRU et aussi plusieurs autres modèles moins performants sur les autres cohortes. En utilisant l'autoencodeur à convolutions, il est possible de classer les espèces d'*Anopheles* avec une précision maximale de 95 %. Il est à noter que cette performance maximale est un résultat notable, compte tenu de la diversité génétique de l'échantillonnage, qui comprend des moustiques de provenances variées, issus de colonies de laboratoire et du terrain. Malgré la complexité et le nombre de variables présentes dans la base de données, seuls deux spécimens de moustiques ont été mal classés, impliquant un *An. gambiae* identifié comme *An. coluzzii*, les deux espèces les plus proches parmi les quatre à identifier.

5.6.2 Comparaison avec des récentes études associant la technologie MALDI-TOF à l'apprentissage automatique

Cette étude confirme l'efficacité des réseaux de neurones pour l'identification des souches épidémiques, telles que pour les bactéries du MABSc. Les recherches précédentes ayant utilisé l'apprentissage automatique sur les spectres MALDI-TOF pour identifier les trois sous-espèces (*M. abscessus*, *M. bolletii*, *M. massiliense*) ont obtenu des performances d'identification ne dépassant pas 90,1 % lors de la validation externe, en utilisant des modèles d'apprentissage automatique classiques (Rodríguez-Temporal et al., 2023). Notre modèle RNN-BiGRU dépasse ces performances, même avec un nombre d'échantillons d'apprentissage moins élevé. Ces résultats restent à confirmer sur une plus large cohorte mais ils font de ce modèle, qui n'avait jamais été testé dans ce contexte, un choix potentiellement privilégié pour les laboratoires. À notre connaissance aucune étude n'avait exploré les techniques d'apprentissage automatique pour détecter des souches résistantes à la clarithromycine chez MABSc, notre étude retrouve des performances de 99 % et 97 % avec le CNN et le TCN respectivement sur le jeu de test indépendant.

Concernant l'identification d'espèces proches, nos performances se rapprochent de celles obtenues pour l'identification des espèces de *Listeria* avec des Autoencodeurs (Li et al., 2022), ainsi que pour l'identification de 18 espèces d'Anophèles avec des réseaux Siamois plus coûteux en termes de calcul (Merchan et al., 2023), atteignant des taux d'identification de 100 % et 99 % respectivement lors de la validation croisée. À la différence

de ces études, nos modèles ont été testés sur des ensembles de test indépendants, en dehors de la validation croisée, confirmant ainsi l'efficacité de ces modèles sur les spectres MALDI-TOF, ainsi que leur robustesse et leur généralisabilité (Weis et al., 2020). Les autres modèles testés démontrent également la possibilité d'utiliser différents types de réseaux de neurones pour résoudre ces problématiques.

Au sujet de la régression de l'âge des moustiques, notre étude confirme l'efficacité du TCN, du CNN, ainsi que du RNN-BiGRU, qui se révèlent être les meilleurs modèles de réseaux de neurones parmi ceux testés pour estimer l'âge des moustiques anophèles (comme indiqué dans le chapitre précédent).

5.6.3 Limites de l'étude

Les limites de cette étude sont liées à l'origine des machines d'acquisition. En effet, les cohortes des *Mycobacterium abscessus* et celle utilisée pour l'âge des moustiques sont chacune composées de spectres provenant d'une seule et même machine MALDI-TOF, ce qui limite notre capacité à évaluer la généralisation des modèles à des spectres provenant de différents laboratoires pour ces problématiques. Cependant, ce n'est pas le cas de la cohorte Anophèles espèces car une partie des spectres provient de Marseille. On peut donc se référer à cette cohorte quant à la généralisation des modèles pour identifier les espèces. Pour les tâches plus complexe comme l'identification des sous espèces ou l'estimation de l'âge, des résultats préliminaires issus d'une étude antérieure ont mis en évidence le potentiel des réseaux de neurones pour l'analyse de spectres provenant de différents appareils MALDI-TOF, atteignant des performances prometteuses dans l'identification de clones résistants de *Candida parapsilosis*, allant jusqu'à 93 % (Mohammad et al., 2023, Chapitre 3).

De plus, la taille de la cohorte de MABSc est limitée (41 souches au total), ce qui restreint notre capacité à évaluer la fiabilité des modèles sur un plus grand nombre de souches provenant de diverses origines (Weis et al., 2020). Les résultats doivent encore être confirmés en utilisant une plus grande variété de données indépendantes. D'autre part, les spectres de moustiques de terrain élevés en laboratoire et utilisés pour prédire l'âge des moustiques ne permettent pas de généraliser la fiabilité des modèles pour prédire l'âge des moustiques vieillissant en milieu naturel et/ou provenant d'origines différentes.

D'un point de vue méthodologique, malgré les recherches sur les transformations du signal MALDI-TOF (Yang et al., 2009, Whistler et al., 2007), telles que les transformations de Fourier et en ondelettes, celles-ci n'ont pas permis d'obtenir les meilleures performances en termes d'identification et de prédiction.
Par ailleurs, cette étude n'établit pas de comparaison directe avec des modèles conventionnels de machine learning, bien que ces derniers soient utilisés en conjonction avec des autoencodeurs de manière semi-supervisée. Cette limite est due au besoin des modèles conventionnels de recourir à un processus de sélection de caractéristiques en amont, contrairement aux réseaux de neurones.
D'un point de vue analytique, notre étude ne vise pas à identifier les pics d'intérêt sur les spectres, ni à découvrir les biomarqueurs responsables des différences entre les classes. Nous nous focalisons davantage sur la recherche de modèles de réseaux de neurones optimaux chargés de filtrer les éléments essentiels à l'identification, en fonction de la représentation spectrale.

Enfin, il convient de noter que cette étude se concentre sur l'utilisation de réseaux de neurones, ce qui implique une compréhension approfondie de cette forme avancée d'apprentissage automatique. Par conséquent, l'intégration de ces modèles dans les laboratoires nécessite une certaine expertise en la matière (Theodosiou et al., 2023).

5.6.4 Perspectives de l'étude

Parmi tous les modèles explorés, on observe des modèles théoriquement prometteurs mais finalement moins robuste que tous les autres comme le 2DCNN BiGRU – Hybrid scalogram pour l'identification de la résistance, ainsi que la classification des sous-espèces proches, et la régression linéaire regression avec l'Autoencodeur à convolution. Cela ouvre la porte à des possibilités de combiner ces modèles avec d'autres pour faire des modèles d'ensemble ou optimiser l'architecture des réseaux de neurones pour améliorer l'identification.

De plus, les représentations spectrales proposées pour permettre aux modèles de prédire offrent des perspectives intéressantes sur les spectres et les techniques de prétraitement qui peuvent être mises en place pour améliorer les performances de prédiction. Les représentations bidimensionnelles, telles que les spectrogrammes et les scalogrammes couramment utilisés pour les signaux audios, n'avaient jusqu'à présent jamais été utilisées sur les

spectres de masse de type MALDI-TOF. Bien que la transformée en ondelette continue n'ait pas été concluante, l'utilisation de la transformée en ondelette discrète, réputée pour son efficacité dans le traitement des spectres MALDI-TOF, pourrait améliorer les performances (Coombes et al., 2007, Starostin et al., 2020).

L'analyse des performances de modèles tels que les 2DCNN et les autoencodeurs, équivalentes ou inférieures à celles des CNN, TCN et RNN-BiGRU, notamment sur la cohorte des moustiques, révèle que l'identification ne peut dépasser un certain seuil maximal (95 % d'accuracy équilibrée) pour les espèces d'anophèles, et la prédiction de l'âge des moustiques présente systématiquement en moyenne une erreur de deux jours. Ces observations suggèrent qu'il serait pertinent d'examiner les données pour déterminer si des corrections peuvent être apportées afin d'améliorer les performances d'identification. Cela pourrait être réalisé par l'utilisation de méthode de contrôle qualité pour nettoyer la base de données des spectres dès la phase préliminaire, contribuant ainsi à réduire un éventuel biais présent dans les données.

Concernant les modèles robustes tels que les CNN, TCN et RNN-BiGRU, il est intéressant d'explorer leur efficacité sur un éventail plus large de données de test indépendantes. Cela inclut la caractérisation d'autres arthropodes ou espèces partageant un niveau de complexité similaire en matière d'identification et de caractérisation, au-delà des mycobactéries.

Naturellement, nous n'avons pas examiné tous les types de réseaux de neurones possibles. Cependant, cette étude ouvre des perspectives quant à la possibilité d'explorer diverses méthodes de représentation sur des modèles récents, tels que les réseaux de neurones Siamois, qui ont déjà démontré leur efficacité dans l'identification d'espèces à partir de spectres bruts (Merchan et al., 2023), ou des réseaux de neurones à apprentissage extrême (Orellana et al., 2022), ou même des modèles adaptés aux données temporelles, comme les Time Delay Neural Networks (Peddinti et al., 2015). Ce derniers est d'autant plus pertinent, car nous avons montré la faisabilité d'utiliser des réseaux récurrents avec des performances solides pour diverses problématiques.

5.7 Conclusion

Cette étude représente une avancée significative en démontrant la faisabilité de l'utilisation de réseaux de neurones sur des spectres de masse de type MALDI-TOF, tout en jetant les bases pour étendre cette méthodologie à d'autres types de spectres de masse. Elle offre un aperçu des avantages et des inconvénients de différentes architectures, ainsi que de leur potentiel pour l'analyse de spectres de masse.

L'objectif de cette étude est double : informer les chercheurs en spectrométrie de masse sur les possibilités de l'apprentissage statistique profond dans ce domaine tout en les inspirant à explorer davantage ces méthodes. Les architectures et les modes de représentation des données varient, chacun ayant ses propres caractéristiques distinctes. Cette diversité permet d'envisager des approfondissements, des combinaisons de composants de différentes architectures, et des améliorations potentielles des modèles.

De plus, certains des modèles et des modes de représentation introduits dans cette étude n'ont jamais été appliqué auparavant aux spectres MALDI-TOF pour des tâches d'identification, notamment des identifications allant au-delà de l'espèces. Ainsi, cette étude ouvre la voie à de nombreuses perspectives passionnantes pour l'utilisation de spectres MALDI-TOF en exploitant les opportunités offertes par l'apprentissage automatique profond, ouvrant ainsi un vaste territoire de recherche jusqu'ici peu étudié.

Chapitre 6

Conclusion et Perspectives

Cette thèse a été dédiée à l'exploration des possibilités offertes par l'utilisation de modèles d'apprentissage statistique profonds en combinaison avec la spectrométrie de masse (SM) de type MALDI-TOF pour relever les défis actuels de la surveillance épidémiologique de maladies infectieuse. Cette approche interdisciplinaire, intégrant la biologie, la bioinformatique et l'apprentissage statistique profond, a permis le développement de méthodes novatrices pour l'identification et la caractérisation des champignons, bactéries et d'arthoropodes. Nous avons également cherché à améliorer la capacité de la SM à individualiser des caractéristiques spécifiques au sein d'une espèce donnée.

6.1 Synthèse des études

Les trois études menées au cours de cette thèse suivent une séquence logique pour répondre aux problématiques initiales.

Dans un premier temps, nous avons évalué l'utilisation de réseaux de neurones pour détecter les clones fongiques épidémiques en milieu hospitalier. Nous avons examiné comment les méthodes de préparation des échantillons et l'analyse informatique des spectres de masse influencent l'apprentissage. Notre étude a révélé que les réseaux de neurones à convolution (CNN) avaient un fort potentiel pour identifier spécifiquement les spectres des clones de *Candida parapsilosis*. En optimisant des paramètres clés, tels que les conditions de culture, le temps de croissance et les équipements d'acquisition des spectres, nous avons atteint une accuracy de 94 %.
Nous avons souligné l'importance des méthodes de préparation des échantillons, de l'analyse des spectres et de leur alignement dans le processus d'identification. Comme le montre notre étude sur l'identification des clones fongiques (Chapitre 3), l'alignement se révèle essentiel pour permettre aux réseaux de neurones de généraliser leurs identifications à partir de données provenant de différentes machines MALDI-TOF. Cela ouvre la voie à des applications d'identification multicentriques de clones fongiques associées à des algorithmes d'apprentissage automatique.

De plus, nous avons démontré l'utilité des réseaux de neurones couplés à la spectrométrie de masse MALDI-TOF dans le contexte de la surveillance du paludisme, en se concentrant sur la population de moustiques vecteurs de cette maladie (Chapitre 4).
Dans une étude portant sur les anophèles, vecteurs du paludisme, nous avons exploré l'utilisation de la spectrométrie de masse MALDI-TOF sur le terrain, associée à l'apprentissage profond, pour prédire l'âge des moustiques anophèles sauvages adultes et modéliser la structure de leur population. En utilisant des modèles d'apprentissage profond optimisés, notamment un réseau de neurones à convolution (CNN) et un réseau de neurones à convolution temporelle (TCN), nous avons pu prédire avec précision l'âge des moustiques, avec une erreur moyenne d'environ deux jours. Cette approche permet une surveillance efficace de la structure d'âge des populations de moustiques anophèles sauvages.
De manière novatrice, nous avons introduit un modèle de réseau de neurones (TCN) qui n'avait jamais été testé auparavant sur les spectres de masse de type MALDI-TOF pour prédire l'âge des moustiques. Cette avancée a repoussé les limites des connaissances actuelles en permettant la prédiction précise de l'âge en jours des moustiques. En outre, nous avons ouvert la voie à l'utilisation des réseaux de neurones pour modéliser la population théorique des moustiques sur le terrain en combinant des outils statistiques, ainsi que pour simuler des scénarios de lutte, offrant ainsi une meilleure orientation des mesures de lutte antivectorielle. Cette approche innovante élargit les horizons des entomologistes dans l'analyse des populations de moustiques.

Enfin, au cœur de cette thèse, une étude méthodologique a exploré divers types de réseaux de neurones pour améliorer la capacité de la SM à discriminer des caractéristiques spécifiques au sein d'une espèce. Nous avons analysé les spectres de masse sous différentes modalités de représentation pour identifier celle qui maximise la performance du modèle en termes d'identification et de prédiction.
Ces méthodes ont été évaluées dans différentes cohortes couvrant divers problèmes épidémiologiques, tels que la prédiction de l'âge des moustiques anophèles, l'identification d'espèces étroitement apparentées, la distinction entre des sous-espèces proches, et la détection de souches résistantes de *Mycobacterium abscessus* Ces défis sont courants dans l'analyse de bases de données de routine, caractérisées par des déséquilibres entre les classes, des limitations de données, des origines variées des données, des étapes de préparation des données, ainsi que des variations entre les préparateurs d'échantillons et les aspects techniques.
Cette étude a révélé des modèles performants jusqu'ici non testés sur les spectres MALDI-TOF. Par exemple, nous avons réussi à classer avec une précision de 93 % les espèces de moustiques Anophèles (*An. arabiensis*, *An. coluzzii*, *An. gambiae* et *An. funestus*), à distinguer les sous-espèces de *Mycobacterium abscessus* avec une précision de 95 % grâce à un modèle récurrent, et à identifier des souches résistantes de *Mycobacterium abscessus* avec des performances exceptionnelles de 97 % et 99 % pour les modèles CNN et TCN respectivement. Ces trois modèles se sont avérés les plus robustes en termes de performances, de temps de calcul et de consommation d'énergie pendant l'entraînement, suggérant qu'ils seraient privilégiés pour une utilisation en routine dans l'identification et la caractérisation des espèces, en particulier dans le contexte des maladies hospitalières.

6.2 Originalité de cette thèse

À notre connaissance, cette thèse constitue la première exploration de l'utilisation de réseaux de neurones pour résoudre des problématiques liées à l'analyse des spectres de masse des micro-organismes et des vecteurs de maladies infectieuses, englobant la classification d'espèces, de sous-espèces, la détection de la résistance, la caractérisation de la clonalité et la description d'arthropodes.

Une autre originalité de cette étude réside dans son approche approfondie du prétraitement des spectres, explorant diverses façons de coupler les réseaux de neurones aux spectres de masse MALDI-TOF. Nous avons considéré les spectres non pas comme un simple signal unidimensionnel de 20 000 points, mais avons exploré des approches de réduction de dimension et de représentation dimensionnelle pour visualiser les spectres MALDI-TOF sous différentes perspectives, facilitant ainsi leur analyse par les réseaux de neurones. Cette thèse répond à la question de la meilleure approche de réseau de neurones pour l'analyse, la classification et la prédiction des spectres MALDI-TOF. Elle propose des modèles robustes avec des performances satisfaisantes, tout en offrant des méthodes de représentation essentielles pour une analyse approfondie des spectres par les modèles.

Cette thèse se démarque par l'exploration de méthodes provenant de divers domaines tels que le traitement du signal sonore, le traitement d'images, et même le traitement des séries temporelles, afin de répondre à notre problématique.
Au cœur de cette thèse, nous avons conçu une variété de modèles de réseaux de neurones en puisant dans ces domaines. Par exemple, nous avons traité les spectres comme des signaux sonores, utilisant le modèle récurrent RNN-BiGRU présenté dans la dernière étude. De même, les modèles 2DCNN spectrogram, 2DCNN BiGRU - Hybrid spectrogram, 2DCNN scalogram et 2DCNN BiGRU - Hybrid scalogram se sont basés sur des approches de traitement de signaux sonores à l'aide de représentations sous forme d'images grâce à la transformation du signal (Fourier et Ondelettes).
Même si les signaux MALDI-TOF présentent des caractéristiques distinctes, nous avons constaté que l'application de modèles généralement utilisés pour d'autres types de signaux pouvait améliorer les performances d'identification par des modèles robustes. De plus, l'utilisation d'auto-encodeurs, généralement associés au traitement de texte ou de signaux sonores pour filtrer le bruit, a été adaptée pour analyser les spectres de masse, permettant ainsi de réduire le bruit tout en préservant les informations relatives aux peptides et aux protéines.

Nous avons abordé les problématiques de cette étude en employant différentes approches de prédiction. Par exemple, nous avons utilisé des modèles de classification binaire pour l'étude des clones fongiques (Chapitre 3), identifié des souches résistantes de *Mycobacterium abscessus* dans la cohorte du dernier chapitre (Chapitre 5), prédit l'âge des moustiques (Chapitre 4), identifié des sous-espèces de *Mycobacterium abscessus* et classé les spectres de moustiques par espèce à l'aide de modèles de classification multiclasse. De plus, nous avons développé une méthode de prédiction originale basée sur une classification cohérente par rang, inspirée d'un article portant sur la prédiction de l'âge à partir de photos.

6.3 Limites

À travers les différentes études de cette thèse, plusieurs lacunes dans la recherche sont apparues. L'une des principales préoccupations concerne la quantité de données disponible. Nous avons constaté que le nombre de souches étudiées pour les clones fongiques se limitait à environ une centaine, de même pour la prédiction de l'âge des moustiques où seulement une centaine de spécimens ont été collectés, et le nombre de souches de *Mycobacterium abscessus* était extrêmement limité, avec seulement 41 souches. Souvent, les jeux de données utilisés se révèlent insuffisants pour parvenir à des conclusions significatives. En tant que chercheurs, nous devons faire preuve de rigueur à cet égard. Cependant, il convient de noter que cette limitation est largement attribuable à la rareté des souches épidémiques disponibles dans les laboratoires hospitaliers et les laboratoires de biologie entomologique, qui naturellement ne disposent pas de quantités massives d'échantillons biologiques.
Les résultats de cette recherche doivent être confirmés sur un échantillon plus large et des données indépendantes pour garantir la généralisation, la robustesse et la fiabilité des modèles. Les travaux présentés dans les chapitres de cette thèse constituent des études préliminaires, car il est essentiel de les valider sur différentes machines et par différents opérateurs afin de démontrer leur capacité à être automatisés. Bien que nous n'ayons pas encore atteint une performance de 100 %, les méthodes présentées incitent à réfléchir sur les démarches nécessaires pour atteindre des performances optimales.

Concernant les spectres de masse et leur analyse, cette thèse se concentre principalement sur l'exploration des réseaux de neurones pour évaluer leur applicabilité aux spectres MALDI-TOF. Nous avons proposé des approches visant à répondre aux problématiques en jouant sur la représentation spectrale de l'empreinte protéique des échantillons analysés. Cependant, nous n'avons pas encore évalué la meilleure représentation possible permettant à un réseau de neurones d'obtenir les meilleures prédictions.
D'un point de vue plus analytique, cette thèse n'aborde pas en profondeur le profilage protéique avec la recherche de biomarqueurs. L'analyse précise, basée sur la recherche exacte des pics discriminants dans les spectres et la recherche approfondie des protéines associées à ces pics, n'a pas été traitée ici, car elle constitue un vaste sujet qui ne peut être exploré de manière exhaustive. Comme mentionné dans le chapitre de l'État de l'art (Chapitre 2), l'approche motivée par l'utilisation des réseaux de neurones se limite au profil de masse peptidique (PMF), laissant le modèle filtrer les pics pertinents et effectuer des prédictions à partir des informations recueillies. Cependant, nous n'examinons pas les éléments filtrés qui sont utiles au modèle.

Les limites évoquées, en particulier les dernières, sont des aspects qui méritent d'être explorés à l'avenir, ouvrant ainsi des perspectives essentielles pour une meilleure évaluation du potentiel de l'utilisation des réseaux de neurones sur les spectres de masse de type MALDI-TOF.

6.4 Perspectives

6.4.1 Les spectres de masse MALDI-TOF

Il est primordial de reconsidérer la manière dont nous représentons les signaux spectraux pour identifier des souches ou des spécimens de micro-organismes, allant au-delà de la simple identification des espèces. Nous avons proposé divers traitements pour changer la perception du signal, mais il est envisageable d'explorer des traitements plus avancés en s'appuyant sur des outils mathématiques. Par exemple, nous avons évoqué l'utilisation de la transformée en ondelette continue pour convertir les spectres en scalogrammes, mais nous n'avons pas exploité la transformée en ondelette discrète, qui s'est avérée être une méthode très efficace pour traiter les spectres MALDI-TOF (Coombes et al., 2005 ; Starostin et al. 2020). Cette thèse ouvre la voie à l'utilisation de méthodes de prétraitement mentionnées dans la revue de la littérature, qui pourraient éventuellement être combinées avec les nouvelles approches utilisées dans la dernière étude. Des méthodes de prétraitement telles que les méthodes bayésiennes, la compression de données ou même des méthodes géométriques méritent d'être étudiées pour traiter le signal et proposer d'autres modes de représentation possibles (Starostin et al., 2020 ; Zhang et al., 2008 ; Conrad et al. ; 2017).
L'intérêt majeur du prétraitement des spectres réside principalement dans la gestion du bruit présent dans le signal. Une approche suggérée dans le dernier chapitre, mais non explorée dans ce contexte, est la reconstruction du signal spectral à l'aide d'autoencodeurs. En effet, un autoencodeur réduit un signal audio bruité en une représentation compacte, puis le décode pour reconstruire un signal propre, éliminant ainsi le bruit. Cette approche pourrait également être envisagée pour l'analyse des spectres MALDI-TOF, offrant la perspective de proposer une méthode permettant de récupérer un certain nombre de spectres de mauvaise qualité obtenus dans des conditions de routine.

6.4.2 Les modèles d'apprentissage statistiques profonds

Pour les perspectives futures dans le domaine de l'apprentissage automatique profond appliqué à la surveillance épidémiologique, il est essentiel de poursuivre la recherche en explorant plus en détail les capacités de cette approche.
Comme évoqué dans la discussion des études présentées dans ce manuscrit, il est pertinent d'explorer de nouveaux modèles de réseaux de neurones afin de mieux déterminer la meilleure approche en apprentissage automatique profond pour aborder les questions de surveillance épidémiologique. Il serait intéressant d'évaluer les performances de modèles plus complexes, sur des problématiques allant au-delà de la simple identification des espèces, comme cela a été étudié dans cette thèse. Cela pourrait inclure des approches plus spécifiques, telles que l'utilisation de réseaux siamois ou même des modèles multi-têtes (Merchan et al., 2023). Par ailleurs, l'exploration de modèles adaptés à des données temporelles, tels que le Time Delay Neural Network (Peddinti et al., 2015), est une piste prometteuse, compte tenu de notre expérience montrant la robustesse de cette approche en combinaison avec des modèles récurrents.

6.4.3 Les données

L'intégration de nouvelles données et de méthodes d'apprentissage profond plus avancées pourrait améliorer considérablement l'identification des agents pathogènes et renforcer la fiabilité des prévisions épidémiologiques. Le manque de données peut compromettre la validité des résultats et la généralisation des conclusions. Afin de consolider la robustesse de nos travaux, il est nécessaire d'élargir notre base de données en incluant un échantillon représentatif de la population ou du phénomène étudié.
De plus, il est essentiel de diversifier les ensembles de tests. Se limiter à un petit nombre d'échantillons ne suffit pas. Les variations naturelles et les nuances propres aux souches et aux spécimens requièrent une couverture plus complète. Un échantillon plus étendu nous permettra de mieux appréhender les tendances, d'évaluer la variabilité et de réduire les possibles biais.

En somme, pour tirer pleinement profit de nos études de thèse, il est impératif d'augmenter la quantité de données, d'élargir la palette des jeux de test et d'inclure diverses souches et spécimens, tout en respectant une méthodologie scientifique stricte. Cela renforcera la validité de nos conclusions et contribuera à l'avancement des connaissances dans notre domaine de recherche.

6.4.4 Les applications

Les résultats de cette thèse ont des implications majeures pour la surveillance épidémiologique et la lutte contre les maladies infectieuses, notamment les infections nosocomiales et les maladies transmises par les vecteurs. Ils fournissent des outils innovants pour prévoir et faire face aux menaces épidémiques à venir. Nous avons non seulement montré que l'apprentissage profond peut être appliqué avec succès dans le domaine de la spectrométrie de masse MALDI-TOF, mais nous avons également mis en évidence des méthodes permettant d'améliorer la précision de l'identification et de la caractérisation des agents pathogènes. Ces avancées sont cruciales pour la prévention et la gestion des maladies infectieuses, qui demeurent un défi majeur pour la santé mondiale.

De plus, la création d'outils conviviaux et d'interfaces utilisateur simplifiées pourrait rendre ces avancées plus accessibles aux professionnels de la santé et aux chercheurs. Enfin, il est essentiel de promouvoir la collaboration interdisciplinaire entre la biologie, la bioinformatique et l'apprentissage statistique pour ouvrir la voie à d'autres applications au-delà de la simple identification des espèces. Cela pourrait inclure la détection de clusters épidémiques de microorganismes résistants aux médicaments, la surveillance de la transmission des maladies bactériennes et fongiques, ainsi que l'évaluation de l'efficacité des interventions ciblées de lutte antivectorielle.

En conclusion, cette thèse a ouvert de nouvelles perspectives passionnantes dans la surveillance épidémiologique en exploitant la puissance de l'apprentissage automatique profond. Nous sommes convaincus que ces avancées joueront un rôle clé dans l'aide à la prévention et la maîtrise des maladies infectieuses à l'avenir.

Bibliographie

A Modern Introduction to Probability and Statistics : Understanding Why and How | SpringerLink. (s.d.). https ://link.springer.com/book/10.1007/1-84628-168-7

Aebersold, R., Mann, M. (2003). Mass spectrometry-based proteomics. Nature, 422(6928), 198-207. https ://doi.org/10.1038/nature01511

Agarap, A. F. (2019). Deep Learning using Rectified Linear Units (ReLU) (arXiv :1803.08375). arXiv. http ://arxiv.org/abs/1803.08375

An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing—Quintelas—2018—Biotechnology Journal—Wiley Online Library. (s. d.). https ://onlinelibrary.wiley.com/doi/10.1002/biot.201700449

Anthony, L. F. W., Kanding, B., Selvan, R. (2020). Carbontracker : Tracking and Predicting the Carbon Footprint of Training Deep Learning Models (arXiv :2007.03051). arXiv. http ://arxiv.org/abs/2007.03051

Antoniadis, A., Bigot, J., Lambert-Lacroix, S. (2010). Peaks detection and alignment for mass spectrometry data. 151(1).

Apicella, B., Bruno, A., Wang, X., Spinelli, N. (2013). Fast Fourier Transform and autocorrelation function for the analysis of complex mass spectra. International Journal of Mass Spectrometry, 338, 30-38. https ://doi.org/10.1016/j.ijms.2013.01.003

Armañanzas, R., Saeys, Y., Inza, I., García-Torres, M., Bielza, C., van de Peer, Y., Larrañaga, P. (2011). Peak-bin selection in mass spectrometry data using a consensus approach with estimation of distribution algorithms. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(3), 760-774. https ://doi.org/10.1109/TCBB.2010.18

Ba, J. L., Kiros, J. R., Hinton, G. E. (2016). Layer Normalization (arXiv :1607.06450). arXiv. http ://arxiv.org/abs/1607.06450

Baggerly, K. A., Morris, J. S., Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum : Comparing datasets from different experiments. Bioinformatics (Oxford, England), 20(5), 777-785. https ://doi.org/10.1093/bioinformatics/btg484

Balasubramanian, S., Andreani, M., Andrade, J. G., Saha, T., Sundaravinayagam, D., Garzón, J., Zhang, W., Popp, O., Hiraga, S.-I., Rahjouei, A., Rosen, D. B., Mertins, P., Chait, B. T., Donaldson, A. D., Di Virgilio, M. (2022). Protection of nascent DNA at stalled replication forks is mediated by phosphorylation of RIF1 intrinsically disordered region. eLife, 11, e75047. https ://doi.org/10.7554/eLife.75047

Bank, D., Koenigstein, N., Giryes, R. (2021). Autoencoders (arXiv :2003.05991). arXiv. http ://arxiv.org/abs/2003.05991

Bastian, S., Veziris, N., Roux, A.-L., Brossier, F., Gaillard, J.-L., Jarlier, V., Cambau, E. (2011). Assessment of clarithromycin susceptibility in strains belonging to the Mycobacterium abscessus group by erm(41) and rrl sequencing. Antimicrobial Agents and Chemotherapy, 55(2), 775-781.
https ://doi.org/10.1128/AAC.00861-10

Bayes Theorem and Information Gain Based Feature Selection for Maximizing the Performance of Classifiers | SpringerLink. (s.d.).
https://link.springer.com/chapter/10.1007/978-3-642-17857-3_49

Bianchi, F. M., Scardapane, S., Løkse, S., Jenssen, R. (2018). Bidirectional deep-readout echo state networks (arXiv :1711.06509). arXiv. http ://arxiv.org/abs/1711.06509

Bilecen, K., Yaman, G., Ciftci, U., Laleli, Y. R. (2015). Performances and Reliability of Bruker Microflex LT and VITEK MS MALDI-TOF Mass Spectrometry Systems for the Identification of Clinical Microorganisms. BioMed Research International, 2015, 516410. https ://doi.org/10.1155/2015/516410

Bizzini, A., Greub, G. (2010). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. Clinical Microbiology and Infection, 16(11), 1614-1619.
https ://doi.org/10.1111/j.1469-0691.2010.03311.x

Brown-Elliott, B. A., Vasireddy, S., Vasireddy, R., Iakhiaeva, E., Howard, S. T., Nash, K., Parodi, N., Strong, A., Gee, M., Smith, T., Wallace, R. J. (2015). Utility of sequencing the erm(41) gene in isolates of Mycobacterium abscessus subsp. Abscessus with low and intermediate clarithromycin MICs. Journal of Clinical Microbiology, 53(4), 1211-1215. https ://doi.org/10.1128/JCM.02950-14

Burckhardt, I., Zimmermann, S. (2018). Susceptibility Testing of Bacteria Using Maldi-Tof Mass Spectrometry. Frontiers in Microbiology, 9. https ://www.frontiersin.org/articles/10.3389/fmicb.2018.01744

Caputo, B., Dani, F. R., Horne, G. L., Petrarca, V., Turillazzi, S., Coluzzi, M., Priestman, A. A., Della Torre, A. (2005). Identification and composition of cuticular hydrocarbons of the major Afrotropical malaria vectorAnopheles gambiae s.s. (Diptera : Culicidae) : analysis of sexual dimorphism and age-related changes. Journal of Mass Spectrometry, 40(12), Article 12. https ://doi.org/10.1002/jms.961

Carnevale, P., Robert, V. (Éds.). (2009). Les anophèles : Biologie, transmission du Plasmodium et lutte antivectorielle. IRD Éditions.
https ://doi.org/10.4000/books.irdeditions.10374

Cassagne, C., Normand, A.-C., L'Ollivier, C., Ranque, S., Piarroux, R. (2016). Performance of MALDI-TOF MS platforms for fungal identification. Mycoses, 59(11), 678-690. https ://doi.org/10.1111/myc.12506

Chavy, A., Nabet, C., Normand, A. C., Kocher, A., Ginouves, M., Prévot, G., Vasconcelos dos Santos, T., Demar, M., Piarroux, R., de Thoisy, B. (2019). Identification of French Guiana sand flies using MALDI-TOF mass spectrometry with a new mass spectra library. PLOS Neglected Tropical Diseases, 13(2), Article 2.
https ://doi.org/10.1371/journal.pntd.0007031

Choi, Y. J., Kim, Y.-J., Yong, D., Byun, J.-H., Kim, T. S., Chang, Y. S., Choi, M. J., Byeon, S. A., Won, E. J., Kim, S. H., Shin, M. G., Shin, J. H. (2018). Fluconazole-Resistant Candida parapsilosis Bloodstream Isolates with Y132F Mutation in ERG11 Gene, South Korea. Emerging Infectious Diseases, 24(9), 1768-1770.
https ://doi.org/10.3201/eid2409.180625

Coetzee, M., Hunt, R. H., Wilkerson, R., Della Torre, A., Coulibaly, M. B., Besansky, N. J. (2013). Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. Zootaxa, 3619, 246-274.

Conrad, T. O. F., Genzel, M., Cvetkovic, N., Wulkow, N., Leichtle, A., Vybiral, J., Kutyniok, G., Schütte, C. (2017). Sparse Proteomics Analysis – a compressed sensing-based approach for feature selection and classifica-

tion of high-dimensional proteomics mass spectrometry data. BMC Bioinformatics, 18(1), 160. https ://doi.org/10.1186/s12859-017-1565-4

Cook, P. E., Hugo, L. E., Iturbe-Ormaetxe, I., Williams, C. R., Chenoweth, S. F., Ritchie, S. A., Ryan, P. A., Kay, B. H., Blows, M. W., O'Neill, S. L. (2007). Predicting the age of mosquitoes using transcriptional profiles. Nature Protocols, 2(11), Article 11. https ://doi.org/10.1038/nprot.2007.396

Cook, P. E., Sinkins, S. P. (2010). Transcriptional profiling of Anopheles gambiae mosquitoes for adult age estimation : Anopheles gambiae age grading. Insect Molecular Biology, 19(6), Article 6. https ://doi.org/10.1111/j.1365-2583.2010.01034.x

Coombes, K. R., Baggerly, K. A., Morris, J. S. (2007). Pre-Processing Mass Spectrometry Data. In W. Dubitzky, M. Granzow, D. Berrar (Éds.), Fundamentals of Data Mining in Genomics and Proteomics (p. 79-102). Springer US. https://doi.org/10.1007/978-0-387-47509-7_4

Coombes, K. R., Fritsche, H. A., Clarke, C., Chen, J., Baggerly, K. A., Morris, J. S., Xiao, L., Hung, M.-C., Kuerer, H. M. (2003). Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization. Clinical Chemistry, 49(10), 1615-1623. https ://doi.org/10.1373/49.10.1615

Coombes, K. R., Koomen, J. M., Baggerly, K. A., Morris, J. S., Kobayashi, R. (2005). Understanding the Characteristics of Mass Spectrometry Data through the use of Simulation. Cancer Informatics, 1, 117693510500100103. https ://doi.org/10.1177/117693510500100103

Croxatto, A., Prod'hom, G., Greub, G. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiology Reviews, 36(2), 380-407. https ://doi.org/10.1111/j.1574-6976.2011.00298.x

Cuénod, A., Foucault, F., Pflüger, V., Egli, A. (2021). Factors Associated With MALDI-TOF Mass Spectral Quality of Species Identification in Clinical Routine Diagnostics. Frontiers in Cellular and Infection Microbiology, 11, 646648. https ://doi.org/10.3389/fcimb.2021.646648

Deep metric learning for the classification of MALDI-TOF spectral signatures from multiple species of neotropical disease vectors. (2023). Artificial Intelligence in the Life Sciences, 3, 100071. https ://doi.org/10.1016/j.ailsci.2023.100071

Del Prete, E., Facchiano, A., Profumo, A., Angelini, C., Romano, P. (2021). GeenaR : A Web Tool for Reproducible MALDI-TOF Analysis. Frontiers in Genetics, 12. https ://www.frontiersin.org/articles/10.3389/fgene.2021.635814

Delavy, M., Cerutti, L., Croxatto, A., Prod'hom, G., Sanglard, D., Greub, G., Coste, A. T. (2020). Machine Learning Approach for Candida albicans Fluconazole Resistance Detection Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. Frontiers in Microbiology, 10. https ://doi.org/10.3389/fmicb.2019.03000

Desoubeaux, G., François, N., Poulain, D., Courcol, R., Chandenier, J., Sendid, B. (2010). Spectrométrie de masse MALDI-TOF, un nouvel outil que la mycologie médicale ne peut contourner. Exploration préliminaire d'une application concernant l'identification de levures isolées dans un CHU français. Journal de Mycologie Médicale, 20(4), 263-267. https ://doi.org/10.1016/j.mycmed.2010.10.003

Detinova, T. S. (1962). Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria. Monograph Series. World Health Organization, 47, 13-191.

Deulofeu, M., Peña-Méndez, E. M., Vaňhara, P., Havel, J., Moráň, L., Pečinka, L., Bagó-Mas, A., Verdú, E., Salvadó, V., Boadas-Vaello, P. (2023). Artificial Neural Networks Coupled with MALDI-TOF MS Serum

Fingerprinting To Classify and Diagnose Pathological Pain Subtypes in Preclinical Models. ACS Chemical Neuroscience, 14(2), 300-311. https ://doi.org/10.1021/acschemneuro.2c00665

Diab-Elschahawi, M., Forstner, C., Hagen, F., Meis, J. F., Lassnig, A. M., Presterl, E., Klaassen, C. H. W. (2020). Microsatellite Genotyping Clarified Conspicuous Accumulation of Candida parapsilosis at a Cardiothoracic Surgery Intensive Care Unit. Journal of Clinical Microbiology, 50(11), 3422-3426. https ://doi.org/10.1128/jcm.01179-12

Du, P., Kibbe, W. A., Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics, 22(17), 2059-2065. https ://doi.org/10.1093/bioinformatics/btl355

Dubrovkin, J. (s. d.). Smoothing of noisy spectra using the Fast Fourier Transform. Eidhammer, I., Flikka, K., Martens, L., Mikalsen, S.-O. (2008). Computational Methods for Mass Spectrometry Proteomics. John Wiley Sons.

Eilers, P. H. C., Boelens, H. F. M. (s. d.). Baseline Correction with Asymmetric Least Squares Smoothing. Elbehiry, A., Aldubaib, M., Abalkhail, A., Marzouk, E., ALbeloushi, A., Moussa, I., Ibrahem, M., Albazie, H., Alqarni, A., Anagreyyah, S., Alghamdi, S., Rawway, M. (2022). How MALDI-TOF Mass Spectrometry Technology Contributes to Microbial Infection Control in Healthcare Settings. Vaccines, 10(11), Article 11. https ://doi.org/10.3390/vaccines10111881

Eriksson, J. O., Sánchez Brotons, A., Rezeli, M., Suits, F., Markó-Varga, G., Horvatovich, P. (2020). MSIWarp : A General Approach to Mass Alignment in Mass Spectrometry Imaging. Analytical Chemistry, 92(24), 16138-16148. https ://doi.org/10.1021/acs.analchem.0c03833

Erlank, E., Koekemoer, L. L., Coetzee, M. (2018). The importance of morphological identification of African anopheline mosquitoes (Diptera : Culicidae) for malaria control programmes. Malaria Journal, 17(1), 43. https ://doi.org/10.1186/s12936-018-2189-5

Evangelista, A. J., Ferreira, T. L. (2022). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in the diagnosis of microorganisms. Future Microbiology, 17(17), Article 17. https ://doi.org/10.2217/fmb-2022-0067

Fangous, M.-S., Mougari, F., Gouriou, S., Calvez, E., Raskine, L., Cambau, E., Payan, C., Héry-Arnaud, G. (2014). Classification algorithm for subspecies identification within the Mycobacterium abscessus species, based on matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of Clinical Microbiology, 52(9), 3362-3369. https ://doi.org/10.1128/JCM.00788-14

Fernández, H. L., Jato, M. R., Peña, D. G., Riverola, F. F. (2014). A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification. International Journal of Data Mining and Bioinformatics, 10(4), 455. https ://doi.org/10.1504/IJDMB.2014.064897

Fiamanya, S., Cipolla, L., Prieto, M., Stelling, J. (2021). Exploring the value of MALDI-TOF MS for the detection of clonal outbreaks of Burkholderia contaminans. Journal of microbiological methods, 181, 106130. https ://doi.org/10.1016/j.mimet.2020.106130

Flores, H. A., O'Neill, S. L. (2018). Controlling vector-borne diseases by releasing modified mosquitoes. Nature Reviews Microbiology, 16(8), Article 8. https ://doi.org/10.1038/s41579-018-0025-0

Florio, W., Baldeschi, L., Rizzato, C., Tavanti, A., Ghelardi, E., Lupetti, A. (2020). Detection of Antibiotic-Resistance by MALDI-TOF Mass Spectrometry : An Expanding Area. Frontiers in Cellular and Infection Microbiology, 10, 572909. https ://doi.org/10.3389/fcimb.2020.572909

Forbes, B. A., Hall, G. S., Miller, M. B., Novak, S. M., Rowlinson, M.-C., Salfinger, M., Somoskövi, A., War-

shauer, D. M., Wilson, M. L. (2018). Practical Guidance for Clinical Microbiology Laboratories : Mycobacteria. Clinical Microbiology Reviews, 31(2), e00038-17. https ://doi.org/10.1128/CMR.00038-17

Foster, A. G. W. (2013). Rapid Identification of Microbes in Positive Blood Cultures by Use of the Vitek MS Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry System. Journal of Clinical Microbiology, 51(11), 3717-3719. https ://doi.org/10.1128/JCM.01679-13

Gautier, M., Ranque, S., Normand, A.-C., Becker, P., Packeu, A., Cassagne, C., L'Ollivier, C., Hendrickx, M., Piarroux, R. (2014). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry : Revolutionizing clinical laboratory diagnosis of mould infections. Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases, 20(12), 1366-1371. https ://doi.org/10.1111/1469-0691.12750

Geographic and Temporal Trends in Isolation and Antifungal Susceptibility of Candida parapsilosis : A Global Assessment from the ARTEMIS DISK Antifungal Surveillance Program, 2001 to 2005 | Journal of Clinical Microbiology. (s. d.). https ://journals.asm.org/doi/10.1128/jcm.02122-07

Gibb, S., Strimmer, K. (2012). MALDIquant : A versatile R package for the analysis of mass spectrometry data. Bioinformatics, 28(17), 2270-2271. https ://doi.org/10.1093/bioinformatics/bts447

Giebel, R., Worden, C., Rust, S. M., Kleinheinz, G. T., Robbins, M., Sandrin, T. R. (2010). Microbial fingerprinting using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) applications and challenges. Advances in Applied Microbiology, 71, 149-184.
https ://doi.org/10.1016/S0065-2164(10)71006-6

Gilles HM, David A. Warrell, Herbert M. Gilles. (2002). Essential malariology (4th ed). Arnold. Glish, G. L., Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. Nature Reviews Drug Discovery, 2(2), Article 2. https ://doi.org/10.1038/nrd1011

González Jiménez, M., Babayan, S. A., Khazaeli, P., Doyle, M., Walton, F., Reedy, E., Glew, T., Viana, M., Ranford-Cartwright, L., Niang, A., Siria, D. J., Okumu, F. O., Diabaté, A., Ferguson, H. M., Baldini, F., Wynne, K. (2019). Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. Wellcome Open Research, 4, 76.
https ://doi.org/10.12688/wellcomeopenres.15201.3

Govender, N. P., Patel, J., Magobo, R. E., Naicker, S., Wadula, J., Whitelaw, A., Coovadia, Y., Kularatne, R., Govind, C., Lockhart, S. R., Zietsman, I. L., TRAC-South Africa group. (2016). Emergence of azole-resistant Candida parapsilosis causing bloodstream infection : Results from laboratory-based sentinel surveillance in South Africa. The Journal of Antimicrobial Chemotherapy, 71(7), 1994-2004.
https ://doi.org/10.1093/jac/dkw091

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354-377.
https ://doi.org/10.1016/j.patcog.2017.10.013

Guajardo, C. E. A., López-Cortés, X. A., & Álvarez, S. H. (2022). Deep learning algorithm applied to bacteria recognition. 2022 IEEE International Conference on Automation/XXV Congress of the Chilean Association of Automatic Control (ICA-ACCA), 1-6.
https ://doi.org/10.1109/ICA-ACCA56767.2022.10005945

Guerra, C. A., Reiner, R. C., Perkins, T. A., Lindsay, S. W., Midega, J. T., Brady, O. J., Barker, C. M., Reisen, W. K., Harrington, L. C., Takken, W., Kitron, U., Lloyd, A. L., Hay, S. I., Scott, T. W., Smith, D. L. (2014). A global assembly of adult female mosquito mark-release-recapture data to inform the control of mosquito-borne pathogens. Parasites Vectors, 7(1), Article 1.
https ://doi.org/10.1186/1756-3305-7-276

Haegler, K., Mueller, N. S., Maccarrone, G., Hunyadi-Gulyas, E., Webhofer, C., Filiou, M. D., Zhang, Y., Turck, C. W. (2009). QuantiSpec—Quantitative mass spectrometry data analysis of 15N-metabolically labeled proteins. Journal of Proteomics, 71(6), 601-608. https ://doi.org/10.1016/j.jprot.2008.10.004

Han, S.-S., Jeong, Y.-S., Choi, S.-K. (2021). Current Scenario and Challenges in the Direct Identification of Microorganisms Using MALDI TOF MS. Microorganisms, 9(9), Article 9.
https ://doi.org/10.3390/microorganisms9091917

Harada, T., Akiyama, Y., Kurashima, A., Nagai, H., Tsuyuguchi, K., Fujii, T., Yano, S., Shigeto, E., Kuraoka, T., Kajiki, A., Kobashi, Y., Kokubu, F., Sato, A., Yoshida, S., Iwamoto, T., Saito, H. (2012). Clinical and Microbiological Differences between Mycobacterium abscessus and Mycobacterium massiliense Lung Diseases. Journal of Clinical Microbiology, 50(11), 3556-3561. https ://doi.org/10.1128/JCM.01175-12

Harbach, R. E. (2004). The classification of genus Anopheles (Diptera : Culicidae) : a working hypothesis of phylogenetic relationships. Bulletin of Entomological Research, 94(6), 537-553.
https ://doi.org/10.1079/ber2004321

He, Q. P., Wang, J., Mobley, J. A., Richman, J., Grizzle, W. E. (2011). Self-Calibrated Warping for Mass Spectra Alignment. Cancer Informatics, 10, 65-82. https ://doi.org/10.4137/CIN.S6358

He, S., Feng, Y., Grant, P. E., Ou, Y. (2022). Deep Relation Learning for Regression and Its Application to Brain Age Estimation. IEEE transactions on medical imaging, 41(9), Article 9.
https ://doi.org/10.1109/TMI.2022.3161739

Hospital Clonal Outbreak of Fluconazole-Resistant Candida parapsilosis Harboring the Y132F ERG11p Substitution in a French Intensive Care Unit | Antimicrobial Agents and Chemotherapy. (s. d.).
https ://journals.asm.org/doi/10.1128/aac.01130-22

Hospital Outbreak of Fluconazole-Resistant Candida parapsilosis : Arguments for Clonal Transmission and Long-Term Persistence | Antimicrobial Agents and Chemotherapy. (s. d.).
https ://journals.asm.org/doi/10.1128/aac.02036-20

Hou, T.-Y., Chiang-Ni, C., Teng, S.-H. (2019). Current status of MALDI-TOF mass spectrometry in clinical microbiology. Journal of Food and Drug Analysis, 27(2), 404-414.
https ://doi.org/10.1016/j.jfda.2019.01.001

House, L. L., Clyde, M. A., Wolpert, R. L. (2011). Bayesian nonparametric models for peak identification in MALDI-TOF mass spectroscopy. The Annals of Applied Statistics, 5(2B).
https ://doi.org/10.1214/10-AOAS450

Huang, Y.-C., Liu, M.-F., Shen, G.-H., Lin, C.-F., Kao, C.-C., Liu, P.-Y., Shi, Z.-Y. (2010). Clinical outcome of Mycobacterium abscessus infection and antimicrobial susceptibility testing. Journal of Microbiology, Immunology, and Infection = Wei Mian Yu Gan Ran Za Zhi, 43(5), 401-406.
https ://doi.org/10.1016/S1684-1182(10)60063-1

Hugo, L. E., Kay, B. H., Eaglesham, G. K., Holling, N., Ryan, P. A. (2006). Investigation of cuticular hydrocarbons for determining the age and survivorship of australasian mosquitoes. The American Journal of Tropical Medicine and Hygiene, 74(3), Article 3.

Hugo, L. E., Monkman, J., Dave, K. A., Wockner, L. F., Birrell, G. W., Norris, E. L., Kienzle, V. J., Sikulu, M. T., Ryan, P. A., Gorman, J. J., Kay, B. H. (2013). Proteomic Biomarkers for Ageing the Mosquito Aedes aegypti to Determine Risk of Pathogen Transmission. PLoS ONE, 8(3), Article 3.
https ://doi.org/10.1371/journal.pone.0058656

Hung, J., Goodman, A., Ravel, D., Lopes, S. C. P., Rangel, G. W., Nery, O. A., Malleret, B., Nosten, F.,

Lacerda, M. V. G., Ferreira, M. U., Rénia, L., Duraisingh, M. T., Costa, F. T. M., Marti, M., Carpenter, A. E. (2020). Keras R-CNN : Library for cell detection in biological images using deep neural networks. BMC Bioinformatics, 21(1), 300. https ://doi.org/10.1186/s12859-020-03635-x

Iacovidou, M. A., Barreaux, P., Spencer, S. E. F., Thomas, M. B., Gorsich, E. E., Rock, K. S. (2022). Omitting age-dependent mosquito mortality in malaria models underestimates the effectiveness of insecticide-treated nets. PLOS Computational Biology, 18(9), Article 9. https ://doi.org/10.1371/journal.pcbi.1009540

Imbert, S., Normand, A. C., Gabriel, F., Cassaing, S., Bonnal, C., Costa, D., Lachaud, L., Hasseine, L., Kristensen, L., Schuttler, C., Raberin, H., Brun, S., Hendrickx, M., Stubbe, D., Piarroux, R., Fekkar, A. (2019). Multi-centric evaluation of the online MSI platform for the identification of cryptic and rare species of Aspergillus by MALDI-TOF. Medical Mycology, 57(8), 962-968. https ://doi.org/10.1093/mmy/myz004

Invasive fungal infections following liver transplantation : Incidence, risk factors, survival, and impact of fluconazole-resistant Candida parapsilosis (2003-2007)—Raghuram—2012—Liver Transplantation—Wiley Online Library. (s. d.).
https ://aasldpubs.onlinelibrary.wiley.com/doi/10.1002/lt.23467

Jabet, A., Normand, A.-C., Moreno-Sabater, A., Guillot, J., Risco-Castillo, V., Brun, S., Demar, M., Blaizot, R., Nabet, C., Packeu, A., Piarroux, R. (2022). Investigations upon the Improvement of Dermatophyte Identification Using an Online Mass Spectrometry Application. Journal of Fungi (Basel, Switzerland), 8(1), 73.
https ://doi.org/10.3390/jof8010073

Jeffries, N. (2005). Algorithms for alignment of mass spectrometry proteomic data. Bioinformatics, 21(14), 3066-3073. https ://doi.org/10.1093/bioinformatics/bti482

Johnson, B. J., Hugo, L. E., Churcher, T. S., Ong, O. T. W., Devine, G. J. (2020). Mosquito Age Grading and Vector-Control Programmes. Trends in Parasitology, 36(1), Article 1.
https ://doi.org/10.1016/j.pt.2019.10.011

K, D. B., B, S., W, W., P, V., B, D. B., P, V. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Systematic and Applied Microbiology, 34(1).
https ://doi.org/10.1016/j.syapm.2010.11.003

Kiranyaz. (2023, mars 29). 1D convolutional neural networks and applications : A survey | Elsevier Enhanced Reader. https ://doi.org/10.1016/j.ymssp.2020.107398

Koh, W.-J., Jeong, B.-H., Kim, S.-Y., Jeon, K., Park, K. U., Jhun, B. W., Lee, H., Park, H. Y., Kim, D. H., Huh, H. J., Ki, C.-S., Lee, N. Y., Kim, H. K., Choi, Y. S., Kim, J., Lee, S.-H., Kim, C. K., Shin, S. J., Daley, C. L., ... Kwon, O. J. (2017). Mycobacterial Characteristics and Treatment Outcomes in Mycobacterium abscessus Lung Disease. Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 64(3), 309-316. https ://doi.org/10.1093/cid/ciw724

La Scola, B., Raoult, D. (2009). Direct identification of bacteria in positive blood culture bottles by matrix-assisted laser desorption ionisation time-of-flight mass spectrometry. PloS One, 4(11), e8041.
https ://doi.org/10.1371/journal.pone.0008041

Lachaud, L., Fernández-Arévalo, A., Normand, A.-C., Lami, P., Nabet, C., Donnadieu, J. L., Piarroux, M., Djenad, F., Cassagne, C., Ravel, C., Tebar, S., Llovet, T., Blanchet, D., Demar, M., Harrat, Z., Aoun, K., Bastien, P., Muñoz, C., Gállego, M., Piarroux, R. (2017). Identification of Leishmania by Matrix-Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF) Mass Spectrometry Using a Free Web-Based Application and a Dedicated Mass-Spectral Library. Journal of Clinical Microbiology, 55(10), 2924-2933.
https ://doi.org/10.1128/JCM.00845-17

Lambert, B., North, A., Godfray, H. C. J. (2022). A Meta-analysis of Longevity Estimates of Mosquito Vectors of Disease [Preprint]. Ecology. https ://doi.org/10.1101/2022.05.30.494059

Lambert, B., Sikulu-Lord, M. T., Mayagaya, V. S., Devine, G., Dowell, F., Churcher, T. S. (2018). Monitoring the Age of Mosquito Populations Using Near-Infrared Spectroscopy. Scientific Reports, 8(1), Article 1. https ://doi.org/10.1038/s41598-018-22712-z

Lea, C., Vidal, R., Reiter, A., Hager, G. D. (2023). Temporal Convolutional Networks : A Unified Approach to Action Segmentation. http ://arxiv.org/abs/1608.08242

Lemaire, Q., Holzapfel, A. (2019). Temporal convolutional networks for speech and music dectetion in radio broadcast. 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands. 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands. Li, Y., Gan, Z., Zhou, X., Chen, Z. (2022). Accurate classification of Listeria species by MALDI-TOF mass spectrometry incorporating denoising autoencoder and machine learning. Journal of Microbiological Methods, 192, 106378. https ://doi.org/10.1016/j.mimet.2021.106378

Ling, J., Li, G., Shao, H., Wang, H., Yin, H., Zhou, H., Song, Y., Chen, G. (2020). Helix Matrix Transformation Combined With Convolutional Neural Network Algorithm for Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry-Based Bacterial Identification. Frontiers in Microbiology, 11, 565434. https ://doi.org/10.3389/fmicb.2020.565434

Liu, W., Wen, Y., Yu, Z., Yang, M. (2017). Large-Margin Softmax Loss for Convolutional Neural Networks (arXiv :1612.02295). arXiv. http ://arxiv.org/abs/1612.02295

López-Cortés, X. A., Astudillo, C. A., González, C., & Maldonado, S. (2021). Semi-supervised learning for MS MALDI-TOF data. 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 1-4. https ://doi.org/10.1109/LA-CCI48322.2021.9769825

López-Fernández, H., Santos, H. M., Capelo, J. L., Fdez-Riverola, F., Glez-Peña, D., Reboiro-Jato, M. (2015). Mass-Up : An all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics, 16(1), 318. https ://doi.org/10.1186/s12859-015-0752-4

Lynn, H. M., Pan, S. B., Kim, P. (2019). A Deep Bidirectional GRU Network Model for Biometric Electrocardiogram Classification Based on Recurrent Neural Networks. IEEE Access, 7, 145395-145405. https ://doi.org/10.1109/ACCESS.2019.2939947

Macdonald, G. (1956). Epidemiological basis of malaria control. Bull. Wld Hlth Org. 15, 613-626.

Mallat, S., Zhong, S. (1992). Characterization of signals from multiscale edges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(7), 710-732. https ://doi.org/10.1109/34.142909

Malyarenko, D. I., Cooke, W. E., Adam, B.-L., Malik, G., Chen, H., Tracy, E. R., Trosset, M. W., Sasinowski, M., Semmes, O. J., Manos, D. M. (2005). Enhancement of Sensitivity and Resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric Records for Serum Peptides Using Time-Series Analysis Techniques. Clinical chemistry, 51(1), 65-74. https ://doi.org/10.1373/clinchem.2004.037283

Manguin, S. (Éd.). (2013). Anopheles mosquitoes—New insights into malaria vectors. InTech. https ://doi.org/10.5772/3392

Marinach, C., Alanio, A., Palous, M., Kwasek, S., Fekkar, A., Brossas, J.-Y., Brun, S., Snounou, G., Hennequin, C., Sanglard, D., Datry, A., Golmard, J.-L., Mazier, D. (2009). MALDI-TOF MS-based drug susceptibility testing of pathogens : The example of Candida albicans and fluconazole. PROTEOMICS, 9(20), 4627-4631. https ://doi.org/10.1002/pmic.200900152

Matthews, J., Bethel, A., Osei, G. (2020). An overview of malarial Anopheles mosquito survival estimates in

relation to methodology. Parasites Vectors, 13(1), Article 1. https ://doi.org/10.1186/s13071-020-04092-4

Mayagaya, V. S., Michel, K., Benedict, M. Q., Killeen, G. F., Wirtz, R. A., Ferguson, H. M., Dowell, F. E. (2009). Non-destructive Determination of Age and Species of Anopheles gambiae s.l. Using Near-infrared Spectroscopy. The American Journal of Tropical Medicine and Hygiene, 81(4), Article 4.
https ://doi.org/10.4269/ajtmh.2009.09-0192

McGrath, E. E., Blades, Z., McCabe, J., Jarry, H., Anderson, P. B. (2010). Nontuberculous mycobacteria and the lung : From suspicion to treatment. Lung, 188(4), 269-282.
https ://doi.org/10.1007/s00408-010-9240-9

Merchan, F., Contreras, K., Gittens, R. A., Loaiza, J. R., Sanchez-Galan, J. E. (2023). Deep metric learning for the classification of MALDI-TOF spectral signatures from multiple species of neotropical disease vectors. Artificial Intelligence in the Life Sciences, 3, 100071. https ://doi.org/10.1016/j.ailsci.2023.100071

Microsatellite Genotyping Clarified Conspicuous Accumulation of Candida parapsilosis at a Cardiothoracic Surgery Intensive Care Unit | Journal of Clinical Microbiology. (s. d.).
https ://journals.asm.org/doi/10.1128/jcm.01179-12

Mohammad, N., Normand, A.-C., Nabet, C., Godmer, A., Brossas, J.-Y., Blaize, M., Bonnal, C., Fekkar, A., Imbert, S., Tannier, X., Piarroux, R. (2023). Improving the Detection of Epidemic Clones in Candida parapsilosis Outbreaks by Combining MALDI-TOF Mass Spectrometry and Deep Learning Approaches. Microorganisms, 11, 1071.

Monchamp, P., Andrade-Cetto, L., Zhang, J. Y., Henson, R. (s. d.). Signal Processing Methods for Mass Spectrometry. 24.

Moreno-Camacho, J. L., Calva-Espinosa, D. Y., Leal-Leyva, Y. Y., Elizalde-Olivas, D. C., Campos-Romero, A., Alcántar-Fernández, J. (2018). Transformation From a Conventional Clinical Microbiology Laboratory to Full Automation. Laboratory Medicine, 49(1), e1-e8. https ://doi.org/10.1093/labmed/lmx079

Mortier, T., Wieme, A. D., Vandamme, P., Waegeman, W. (2021). Bacterial species identification using MALDI-TOF mass spectrometry and machine learning techniques : A large-scale benchmarking study. Computational and Structural Biotechnology Journal, 19, 6157-6168. https ://doi.org/10.1016/j.csbj.2021.11.004

Mougari, F., Guglielmetti, L., Raskine, L., Sermet-Gaudelus, I., Veziris, N., Cambau, E. (2016). Infections caused by Mycobacterium abscessus : Epidemiology, diagnostic tools and treatment. Expert Review of Anti-Infective Therapy, 14(12), 1139-1154. https ://doi.org/10.1080/14787210.2016.1238304

Müller, P., Pflüger, V., Wittwer, M., Ziegler, D., Chandre, F., Simard, F., Lengeler, C. (2013). Identification of Cryptic Anopheles Mosquito Species by Molecular Protein Profiling. PLOS ONE, 8(2), e57486.
https ://doi.org/10.1371/journal.pone.0057486

Nabet, C., Chaline, A., Franetich, J.-F., Brossas, J.-Y., Shahmirian, N., Silvie, O., Tannier, X., Piarroux, R. (2020). Prediction of malaria transmission drivers in Anopheles mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry. Scientific Reports, 10(1), Article 1. https ://doi.org/10.1038/s41598-020-68272-z

Nabet, C., Kone, A. K., Dia, A. K., Sylla, M., Gautier, M., Yattara, M., Thera, M. A., Faye, O., Braack, L., Manguin, S., Beavogui, A. H., Doumbo, O., Gay, F., Piarroux, R. (2021). New assessment of Anopheles vector species identification using MALDI-TOF MS. Malaria Journal, 20(1), Article 1.
https ://doi.org/10.1186/s12936-020-03557-2

Nii-Trebi, N. I. (2017). Emerging and Neglected Infectious Diseases : Insights, Advances, and Challenges. BioMed Research International, 2017, 5245021. https ://doi.org/10.1155/2017/5245021

Nirthika, R., Manivannan, S., Ramanan, A., Wang, R. (2022). Pooling in convolutional neural networks for medical image analysis : A survey and an empirical study. Neural Computing and Applications, 34(7), 5321-5347. https ://doi.org/10.1007/s00521-022-06953-8

Nomura, F. (2015). Proteome-based bacterial identification using matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) : A revolutionary shift in clinical diagnostic microbiology. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1854(6), 528-537.
https ://doi.org/10.1016/j.bbapap.2014.10.022

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling : The MI-CLAIM checklist. Nature Medicine, 26(9), Article 9.
https ://doi.org/10.1038/s41591-020-1041-y

Normand, A.-C., Chaline, A., Mohammad, N., Godmer, A., Acherar, A., Huguenin, A., Ranque, S., Tannier, X., Piarroux, R. (2022). Identification of a clonal population of Aspergillus flavus by MALDI-TOF mass spectrometry using deep learning. Scientific Reports, 12(1), Article 1.
https ://doi.org/10.1038/s41598-022-05647-4

Normand, A.-C., Moreno-Sabater, A., Jabet, A., Hamane, S., Cremer, G., Foulet, F., Blaize, M., Dellière, S., Bonnal, C., Imbert, S., Brun, S., Packeu, A., Bretagne, S., Piarroux, R. (2022). MALDI-TOF Mass Spectrometry Online Identification of Trichophyton indotineae Using the MSI-2 Application. Journal of Fungi, 8(10), Article 10. https ://doi.org/10.3390/jof8101103

Orellana, M. U., López-Cortès, X. A., Zabala-Blanco, D., Játiva, P. P., & Datta, J. (2022). Extreme Learning Machine for Mass Spectrometry Data Analysis. 2022 IEEE Colombian Conference on Communications and Computing (COLCOM), 1-6.
https ://doi.org/10.1109/Colcom56784.2022.10107875

Pais, R. J., Iles, R. K., Zmuidinaite, R. (2021). MALDI-ToF Mass Spectra Phenomic Analysis for Human Disease Diagnosis Enabled by Cutting-Edge Data Processing Pipelines and Bioinformatics Tools. Current Medicinal Chemistry, 28(32), 6532-6547. https ://doi.org/10.2174/0929867327666201027154257

Papagiannopoulou, C., Parchen, R., Rubbens, P., Waegeman, W. (2020). Fast Pathogen Identification Using Single-Cell Matrix-Assisted Laser Desorption/Ionization-Aerosol Time-of-Flight Mass Spectrometry Data and Deep Learning Methods. Analytical Chemistry, 92(11), 7523-7531. https ://doi.org/10.1021/acs.analchem.9b05806

Papagiannopoulou, C., Parchen, R., Waegeman, W. (s. d.). Investigating Time Series Classification Techniques for Rapid Pathogen Identification with Single-Cell MALDI-TOF Mass Spectrum Data. Parker, C. E., Mocanu, V., Mocanu, M., Dicheva, N., Warren, M. R. (2010). Mass Spectrometry for Post-Translational Modifications. In O. Alzate (Éd.), Neuroproteomics. CRC Press/Taylor Francis.
http ://www.ncbi.nlm.nih.gov/books/NBK56012/

Patel, R. (2019). A Moldy Application of MALDI : MALDI-ToF Mass Spectrometry for Fungal Identification. Journal of Fungi (Basel, Switzerland), 5(1), 4. https ://doi.org/10.3390/jof5010004

Peddinti, V., Povey, D., Khudanpur, S. (2015b). A time delay neural network architecture for efficient modeling of long temporal contexts. Interspeech 2015, 3214-3218. https ://doi.org/10.21437/Interspeech.2015-647

Pinhati, H. M. S., Casulari, L. A., Souza, A. C. R., Siqueira, R. A., Damasceno, C. M. G., Colombo, A. L. (2016). Outbreak of candidemia caused by fluconazole resistant Candida parapsilosis strains in an intensive care unit. BMC Infectious Diseases, 16(1), 433. https ://doi.org/10.1186/s12879-016-1767-9

Pizzato, J., Tang, W., Bernabeu, S., Bonnin, R. A., Bille, E., Farfour, E., Guillard, T., Barraud, O., Cattoir, V.,

Plouzeau, C., Corvec, S., Shahrezaei, V., Dortet, L., Larrouy-Maumus, G. (2022). Discrimination of Escherichia coli, Shigella flexneri, and Shigella sonnei using lipid profiling by MALDI-TOF mass spectrometry paired with machine learning. MicrobiologyOpen, 11(4), e1313. https ://doi.org/10.1002/mbo3.1313

Popa, S. L., Pop, C., Dita, M. O., Brata, V. D., Bolchis, R., Czako, Z., Saadani, M. M., Ismaiel, A., Dumitrascu, D. I., Grad, S., David, L., Cismaru, G., Padureanu, A. M. (2022). Deep Learning and Antibiotic Resistance. Antibiotics, 11(11), Article 11. https ://doi.org/10.3390/antibiotics11111674

Presente, S., Bonnal, C., Normand, A.-C., Gaudonnet, Y., Fekkar, A., Timsit, J.-F., Kernéis, S. (2023). Hospital Clonal Outbreak of Fluconazole-Resistant Candida parapsilosis Harboring the Y132F ERG11p Substitution in a French Intensive Care Unit. Antimicrobial Agents and Chemotherapy, 67(3), e01130-22. https ://doi.org/10.1128/aac.01130-22

Ranson, H., Lissenden, N. (2016). Insecticide Resistance in African Anopheles Mosquitoes : A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. Trends in Parasitology, 32(3), Article 3. https ://doi.org/10.1016/j.pt.2015.11.010

Rapid Detection of COVID-19 Using MALDI-TOF-Based Serum Peptidome Profiling | Analytical Chemistry. (s. d.). https ://pubs.acs.org/doi/10.1021/acs.analchem.0c04590

Rapid identification of bacteria from bioMerieux BacT/ALERT blood culture bottles by MALDI-TOF MS : British Journal of Biomedical Science : Vol 70, No 4. (s. d.). https ://www.tandfonline.com/doi/abs/10.1080/09674845.2013.11669949

Rashidi, H. H., Pepper, J., Howard, T., Klein, K., May, L., Albahra, S., Phinney, B., Salemi, M. R., Tran, N. K. (2022). Comparative performance of two automated machine learning platforms for COVID-19 detection by MALDI-TOF-MS. PLoS ONE, 17(7), e0263954. https ://doi.org/10.1371/journal.pone.0263954

Redeker, V., Vinh, J., Le Caer, J. P. (1998). Characterization of posttranslational modifications of proteins by MALDI-TOF MS : Application to the study of tubulin. Analusis, 26(10), 22-25. https ://doi.org/10.1051/analusis :1998260022

Ressom, H. W., Varghese, R. S., Drake, S. K., Hortin, G. L., Abdel-Hamid, M., Loffredo, C. A., Goldman, R. (2007). Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics, 23(5), 619-626. https ://doi.org/10.1093/bioinformatics/btl678

Rockwood, A. (2003). Isotopic compositions and accurate masses of single isotopic peaks. Journal of the American Society for Mass Spectrometry. https ://doi.org/10.1016/S1044-0305(03)00631-7

Rockwood, A. L. (1995). Relationship of Fourier transforms to isotope distribution calculations. Rapid Communications in Mass Spectrometry, 9(1), 103-105. https ://doi.org/10.1002/rcm.1290090122

Rodríguez-Temporal, D., Herrera, L., Alcaide, F., Domingo, D., Héry-Arnaud, G., van Ingen, J., Van den Bossche, A., Ingebretsen, A., Beauruelle, C., Terschlüsen, E., Boarbi, S., Vila, N., Arroyo, M. J., Méndez, G., Muñoz, P., Mancera, L., Ruiz-Serrano, M. J., Rodríguez-Sánchez, B. (2023). Identification of Mycobacterium abscessus Subspecies by MALDI-TOF Mass Spectrometry and Machine Learning. Journal of Clinical Microbiology, 61(1), e0111022. https ://doi.org/10.1128/jcm.01110-22

Sabino, R., Sampaio, P., Rosado, L., Stevens, D. A., Clemons, K. V., Pais, C. (2010). New Polymorphic Microsatellite Markers Able To Distinguish among Candida parapsilosis Sensu Stricto Isolates. Journal of Clinical Microbiology, 48(5), 1677-1682. https ://doi.org/10.1128/jcm.02151-09

Salles, R. S., Ribeiro, P. F. (2023). The use of deep learning and 2-D wavelet scalograms for power quality disturbances classification. Electric Power Systems Research, 214, 108834.

https ://doi.org/10.1016/j.epsr.2022.108834

Sevestre, J., Diarra, A. Z., Laroche, M., Almeras, L., Parola, P. (2021). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry : An emerging tool for studying the vectors of human infectious diseases. Future Microbiology, 16, 323-340. https ://doi.org/10.2217/fmb-2020-0145

Shaw, W. R., Catteruccia, F. (2018). Vector biology meets disease control : Using basic research to fight vector-borne diseases. Nature Microbiology, 4(1), Article 1.
https ://doi.org/10.1038/s41564-018-0214-7

Shi, X., Cao, W., Raschka, S. (s. d.). Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities. https ://doi.org/10.48550/arXiv.2111.08851

Shin, H., Mutlu, M., Koomen, J. M., Markey, M. K. (2007a). Parametric Power Spectral Density Analysis of Noise from Instrumentation in MALDI TOF Mass Spectrometry. Cancer Informatics, 3, 117693510700300.
https ://doi.org/10.1177/117693510700300019

Shin, H., Mutlu, M., Koomen, J. M., Markey, M. K. (2007b). Parametric Power Spectral Density Analysis of Noise from Instrumentation in MALDI TOF Mass Spectrometry. Cancer Informatics, 3, 117693510700300019.
https ://doi.org/10.1177/117693510700300019

Sikulu, M., Killeen, G. F., Hugo, L. E., Ryan, P. A., Dowell, K. M., Wirtz, R. A., Moore, S. J., Dowell, F. E. (2010). Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. Parasites Vectors, 3(1), Article 1. https ://doi.org/10.1186/1756-3305-3-49

Sikulu, M. T., Monkman, J., Dave, K. A., Hastie, M. L., Dale, P. E., Kitching, R. L., Killeen, G. F., Kay, B. H., Gorman, J. J., Hugo, L. E. (2015). Proteomic changes occurring in the malaria mosquitoes Anopheles gambiae and Anopheles stephensi during aging. Journal of Proteomics, 126, 234-244. https ://doi.org/10.1016/j.jprot.2015.06.008

Singhal, N., Kumar, M., Kanaujia, P. K., Virdi, J. S. (2015). MALDI-TOF mass spectrometry : An emerging technology for microbial identification and diagnosis. Frontiers in Microbiology, 6.
https ://doi.org/10.3389/fmicb.2015.00791

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C. M., Hemingway, J., Patil, A. P., Temperley, W. H., Gething, P. W., Kabaria, C. W., Burkot, T. R., Harbach, R. E., Hay, S. I. (2012). A global map of dominant malaria vectors. Parasites Vectors, 5, 69.
https ://doi.org/10.1186/1756-3305-5-69

Siria, D. J., Sanou, R., Mitton, J., Mwanga, E. P., Niang, A., Sare, I., Johnson, P. C. D., Foster, G. M., Belem, A. M. G., Wynne, K., Murray-Smith, R., Ferguson, H. M., González-Jiménez, M., Babayan, S. A., Diabaté, A., Okumu, F. O., Baldini, F. (2022). Rapid age-grading and species identification of natural mosquitoes for malaria surveillance. Nature Communications, 13(1), Article 1.
https ://doi.org/10.1038/s41467-022-28980-8

Smith, K. P., Kang, A. D., Kirby, J. E. (2018). Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network. Journal of Clinical Microbiology, 56(3), 10.1128/jcm.01521-17.
https ://doi.org/10.1128/jcm.01521-17

Sogawa, K., Watanabe, M., Ishige, T., Segawa, S., Miyabe, A., Murata, S., Saito, T., Sanda, A., Furuhata, K., Nomura, F. (2017). Rapid Discrimination between Methicillin-Sensitive and Methicillin-Resistant Staphylococcus aureus Using MALDI-TOF Mass Spectrometry. Biocontrol Science, 22(3), 163-169.
https ://doi.org/10.4265/bio.22.163

Sorace, J. M., Zhan, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics, 4, 24. https ://doi.org/10.1186/1471-2105-4-24

Starostin, K. V., Demidov, E. A., Ershov, N. I., Bryanskaya, A. V., Efimov, V. M., Shlyakhtun, V. N., Peltek, S. E. (2020). Creation of an Online Platform for Identification of Microorganisms : Peak Picking or Full-Spectrum Analysis. Frontiers in Microbiology, 11. https ://www.frontiersin.org/articles/10.3389/fmicb.2020.609033

Stevenson, L. G., Drake, S. K., Shea, Y. R., Zelazny, A. M., Murray, P. R. (2010). Evaluation of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of clinically important yeast species. Journal of Clinical Microbiology, 48(10), 3482-3486. https ://doi.org/10.1128/JCM.00687-09

Sun, C., Song, M., Hong, S., Li, H. (2020). A Review of Designs and Applications of Echo State Networks (arXiv :2012.02974). arXiv. http ://arxiv.org/abs/2012.02974

Sun, C., Zhang, M., Wu, R., Lu, J., Xian, G., Yu, Q., Gong, X., Luo, R. (2021). A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. Scientific Reports, 11(1), 1434. https ://doi.org/10.1038/s41598-020-80713-3

Tadec, L., Talarmin, J.-P., Gastinne, T., Bretonnière, C., Miegeville, M., Le Pape, P., Morio, F. (2016). Epidemiology, risk factor, species distribution, antifungal resistance and outcome of Candidemia at a single French hospital : A 7-year study. Mycoses, 59(5), 296-303. https ://doi.org/10.1111/myc.12470

Tadros, M., Petrich, A. (2013). Evaluation of MALDI-TOF Mass Spectrometry and Sepsityper Kit™ for the Direct Identification of Organisms from Sterile Body Fluids in a Canadian Pediatric Hospital. Canadian Journal of Infectious Diseases and Medical Microbiology, 24(4), 191-194. https ://doi.org/10.1155/2013/701093

Theodosiou, A. A., Read, R. C. (2023). Artificial intelligence, machine learning and deep learning : Potential resources for the infection clinician. Journal of Infection, 87(4), 287-294.
https ://doi.org/10.1016/j.jinf.2023.07.006

Thomaz, D. Y., de Almeida, J. N., Lima, G. M. E., Nunes, M. de O., Camargo, C. H., Grenfell, R. de C., Benard, G., Del Negro, G. M. B. (2018). An Azole-Resistant Candida parapsilosis Outbreak : Clonal Persistence in the Intensive Care Unit of a Brazilian Teaching Hospital. Frontiers in Microbiology, 9.
https ://www.frontiersin.org/articles/10.3389/fmicb.2018.02997

Tong, D. L., Boocock, D. J., Coveney, C., Saif, J., Gomez, S. G., Querol, S., Rees, R., Ball, G. R. (2011). A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis : Stem cell and melanoma cancer studies. Clinical proteomics, 8(1), 14. https ://doi.org/10.1186/1559-0275-8-14

Torres-Sangiao, E., Leal Rodriguez, C., García-Riestra, C. (2021). Application and Perspectives of MALDI–TOF Mass Spectrometry in Clinical Microbiology Laboratories. Microorganisms, 9(7), Article 7.
https ://doi.org/10.3390/microorganisms9071539

van der Werf, M. J., Ködmön, C., Katalinić-Janković, V., Kummik, T., Soini, H., Richter, E., Papaventsis, D., Tortoli, E., Perrin, M., van Soolingen, D., Žolnir-Dovč, M., Østergaard Thomsen, V. (2014). Inventory study of non-tuberculous mycobacteria in the European Union. BMC Infectious Diseases, 14(1), 62.
https ://doi.org/10.1186/1471-2334-14-62

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2023). Attention Is All You Need (arXiv :1706.03762). arXiv. http ://arxiv.org/abs/1706.03762

Vervier, K., Mahé, P., Veyrieras, J.-B., Vert, J.-P. (2015). Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data (arXiv :1506.07251). arXiv.
http ://arxiv.org/abs/1506.07251

Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J., Tempst, P. (2005). Correcting Common Errors in Identifying Cancer-Specific Serum Peptide Signatures. Journal

of proteome research, 4(4), 1060-1072. https ://doi.org/10.1021/pr050034b

Vrioni, G., Tsiamis, C., Oikonomidis, G., Theodoridou, K., Kapsimali, V., Tsakris, A. (2018). MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance : Current achievements and future perspectives. Annals of Translational Medicine, 6(12), 240. https ://doi.org/10.21037/atm.2018.06.28

Wagner, I., Grigoraki, L., Enevoldson, P., Clarkson, M., Jones, S., Hurst, J. L., Beynon, R. J., Ranson, H. (2023). Rapid identification of mosquito species and age by mass spectrometric analysis. BMC Biology, 21, 10. https ://doi.org/10.1186/s12915-022-01508-8

Wang, H.-Y., Hsieh, T.-T., Chung, C.-R., Chang, H.-C., Horng, J.-T., Lu, J.-J., Huang, J.-H. (2022). Efficiently Predicting Vancomycin Resistance of Enterococcus Faecium From MALDI-TOF MS Spectra Using a Deep Learning-Based Approach. Frontiers in Microbiology, 13, 821233.
https ://doi.org/10.3389/fmicb.2022.821233

Wang, M.-H., Marinotti, O., Zhong, D., James, A. A., Walker, E., Guda, T., Kweka, E. J., Githure, J., Yan, G. (2013). Gene Expression-Based Biomarkers for Anopheles gambiae Age Grading. PLoS ONE, 8(7), Article 7.
https ://doi.org/10.1371/journal.pone.0069439

Weems, J. J. (1992). Candida parapsilosis : Epidemiology, Pathogenicity, Clinical Manifestations, and Antimicrobial Susceptibility. Clinical Infectious Diseases, 14(3),756-766. https ://doi.org/10.1093/clinids/14.3.756

Weeraratne, T. C., Karunaratne, S. H. P. P., Reimer, L., de Silva, W. A. P. P., Wondji, C. S. (2021). Use of transcriptional age grading technique to determine the chronological age of Sri Lankan Aedes aegypti and Aedes albopictus females. Parasites Vectors, 14(1), Article 1. https ://doi.org/10.1186/s13071-021-04994-x

Weis, C., Cuénod, A., Rieck, B., Dubuis, O., Graf, S., Lang, C., Oberle, M., Brackmann, M., Søgaard, K. K., Osthoff, M., Borgwardt, K., Egli, A. (2022). Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nature Medicine, 28(1), 164-174.
https ://doi.org/10.1038/s41591-021-01619-9

Weis, C. V., Jutzeler, C. R., Borgwardt, K. (2020). Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra : A systematic review. Clinical Microbiology and Infection, 26(10), 1310-1317. https ://doi.org/10.1016/j.cmi.2020.03.014

Whistler, T., Rollin, D., Vernon, S. D. (2007). A method for improving SELDI-TOF mass spectrometry data quality. Proteome Science, 5, 14.
https ://doi.org/10.1186/1477-5956-5-14

Wijetunge, C. D., Saeed, I., Boughton, B. A., Roessner, U., Halgamuge, S. K. (2015). A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model. BMC Genomics, 16(Suppl 12), S12. https ://doi.org/10.1186/1471-2164-16-S12-S12

Wilkins, E. E., Howell, P. I., Benedict, M. Q. (2006). IMP PCR primers detect single nucleotide polymorphisms for Anopheles gambiae species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in Anopheles arabiensis. Malaria Journal, 5(1), Article 1.
https ://doi.org/10.1186/1475-2875-5-125

Wong, J. W. H., Durante, C., Cartwright, H. M. (2005a). Application of Fast Fourier Transform Cross-Correlation for the Alignment of Large Chromatographic and Spectral Datasets. Analytical Chemistry, 77(17), 5655-5661. https ://doi.org/10.1021/ac050619p

Wong, J. W. H., Durante, C., Cartwright, H. M. (2005b). Application of Fast Fourier Transform Cross-Correlation for the Alignment of Large Chromatographic and Spectral Datasets. Analytical Chemistry, 77(17), 5655-5661. https ://doi.org/10.1021/ac050619p

Wyse, L. (2017). Audio Spectrogram Representations for Processing with Convolutional Neural Networks (arXiv :1706.09559). arXiv. http ://arxiv.org/abs/1706.09559

Xu, B., Wang, N., Chen, T., Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. http ://arxiv.org/abs/1505.00853

Yang, C., He, Z., Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics, 10(1), 4. https ://doi.org/10.1186/1471-2105-10-4

Yssouf, A., Almeras, L., Raoult, D., Parola, P. (2016). Emerging tools for identification of arthropod vectors. Future Microbiology, 11(4), Article 4. https ://doi.org/10.2217/fmb.16.5

Zhang, S.-Q., Zhou, X., Wang, H., Suffredini, A., Gonzales, D., Ching, W.-K., Ng, M. K., Wong, S. (s. d.). Peak Detection with Chemical Noise Removal Using Short-Time FFT for a Kind of MALDI Data.

Zhang, Z., Sabuncu, M. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. Advances in Neural Information Processing Systems, 31.

Zhou, Q., Yong, B., Lv, Q., Shen, J., Wang, X. (2020). Deep Autoencoder for Mass Spectrometry Feature Learning and Cancer Detection. IEEE Access, 8, 45156-45166.
https ://doi.org/10.1109/ACCESS.2020.2977680

Zhu, Y., Girault, H. H. (2023). Algorithms push forward the application of MALDI–TOF mass fingerprinting in rapid precise diagnosis. VIEW, 4(2), 20220042.
https ://doi.org/10.1002/VIW.20220042

Production scientifique dans le cadre de la thèse

Publications déjà publiées

Noshine Mohammad*, A. Huguenin*, A. Lefebvre, L. Menvielle, D. Toubas, S. Ranque, I. Villena, X. Tannier, A-C Normand, R. Piarroux, 9 Janvier 2024, Medical Mycology - Oxford Academic.
Nosocomial transmission of Aspergillus flavus in a neonatal intensive care unit : long term persistence in environment and interest of MALDI-ToF Mass-Spectrometry coupled with Convolutional Neural Network (CNN) for rapid clone recognition

*Co-premiers auteurs

Noshine Mohammad, A-C Normand, C. Nabet, A. Godmer, J-Y Brossas, M. Blaize, C. Bonnal, A. Fekkar, S. Imbert, X. Tannier, R. Piarroux, 17 avril 2023, MDPI – Microorganisms.
Improving the Detection of Epidemic Clones in Candida parapsilosis Outbreaks by Combining MALDI-TOF Mass Spectrometry and Deep Learning Approaches

A-C Normand , A. Chaline, **Noshine Mohammad**, A. Godmer, A. Acherar, A. Huguenin, S. Ranque, X. Tannier, R. Piarroux, 28 Janvier 2022, Scientific Reports - Nature.
Identification of a clonal population of Aspergillus flavus by MALDI-TOF mass spectrometry using deep learning

Publications soumises

Noshine Mohammad*, P. Naudion*, A. K. Dia, P-Y Boëlle, A. Konaté, L. Konate, E-H A. Niang, R. Piarroux, X. Tannier and C. Nabet, en cours de révision, Science Advances.
Accurate mosquito age prediction for epidemiological monitoring of malaria transmission

*Co-premiers auteurs

A. Godmer, L. Bigey, Q. Giai-Gianetto, G. Pierrat, **Noshine Mohammad**, E. Cambau, R. Piarroux, N. Veziris, A. Aubry, article soumis, Journal of Clinical Microbiology.
Contribution of Machine Learning for subspecies identification from Mycobacterium abscessus complex with MALDI-TOF MS in solid and liquid media

Publication en cours d'écriture

Noshine Mohammad, A. Godmer, R. Piarroux, C. Nabet, X. Tannier, en cours d'écriture (soumise courant 2024).
Exploring deep learning models coupled with MALDI-TOF mass spectrometry to improve epidemiological surveillance of infectious diseases

Communications

2023 - Séminaire de l'Ecole doctorale 393 Pierre Louis de Santé Publique à Saint Malo (France) du 6 au 8 février
Improving age prediction of field malaria mosquitoes using deep learning models coupled to proteomics. Noshine Mohammad, Pauline Naudion, Noemie Shahmirian, Abdoulaye Kane Dia, Pierre-Yves Boëlle, Renaud Piarroux, Xavier Tannier, Cécile Nabet. Poster.

2023 - Webinar de la JM-SFM : L'intelligence artificielle au service de la microbiologie
Oratrice : L'IA et microbiologie : quelques outils en recherche pour une future application au diagnostic

2022 - DMU BioGem - 2e séminaire virtuel portant sur l'Intelligence Artificielle : «IA et diagnostic biologique»
IA et diagnostic biologique : Amélioration des techniques et leurs limites

2021 - 31st European Congress of Clinical Microbiology & Infectious Diseases (ECCMID)
A Deep Learning model for Mycobacterium abscessus complex subspecies identification with MALDI-TOF MS. Alexandre Godmer, Noshine Mohammad, Yahia Benzerara, Anne-Cecile Normand, Nicolas Veziris, Renaud Piarroux, Alexandra Aubry. Poster.

Enseignements

2021 à 2023 - UE : La médecine à l'heure des « OMIQUES » et de l'IA
Deux cours enseignés : Introduction à l'apprentissage statistique (Cours 1) et Apprentissage statistique et Protéomique au service des maladies infectieuses (Cours 2).
Responsable de l'UE : Antonin Lamazière (PU-PH à l'Hôpital Saint Antoine, AP-HP/Sorbonne Université, Paris, France)

2022 à 2023 - UE : Intelligence artificielle et médecine
Cours enseignés : Apprentissage automatique et spectrométrie de masse. Enseignements dirigés : Prise en main des outils informa- tiques et initiation au code Python sur l'utilisation de réseau de neurone dans la reconnaissance d'image. (2022) Initiation au langage de programmation R : exemples sur des données cliniques et la découverte de l'analyse en composantes principales. (2023).
Responsable de l'UE : Cécile Nabet (MCU-PH à l'Hôpital de la Pitié-Salpétrière, AP-HP/Sorbonne Université, Paris, France)

2022 - La Journée Bio-Info du ReJMiC - Jussieu 2022
Cours et enseignement dirigé : Rappels sur les statistiques, initiation au langage de programmation R et la découverte de l'analyse en composantes principales.
Responsable du programme de formation : Alexandre Godmer (APH à l'Hôpital Saint Antoine, AP-HP/Sorbonne Université, Paris, France)

Publications issues de la thèse et présentations de posters

Veuillez consulter le fichier Annexes.pdf fourni sous format zip.

Annexe A

Annexe B

Annexe C