République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia Farès de Médéa

Faculté de Technologie

Département de Génie des Procèdes et Environnement

Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT)

THESE

EN VUE DE L'OBTENTION DU DIPLOME DE DOCTORAT

Filière : Génie des Procédés

Option : Génie des Procédés

Thème

Prédiction des Affections Transmissibles et non Transmissibles par Intelligence Artificielle

Présentée par : TARED Samira

Devant le jury composé de :

Mr. LAIDI Maamar	Professeur	UYFM, Médéa	Président
Mr. RAHAL Soufiane	MCA	UYFM, Médéa	Examinateur
Mr. KERCHICH Yacine	Professeur	ENP, Alger	Examinateur
Mme. KHAOUANE Latifa	Professeur	UYFM, Médéa	Directrice de thèse
Mr. HANINI Salah	Professeur	UYFM, Médéa	Co-Directeur de thèse

Remerciements

Je tiens en premier lieu et avant tout à remercier **ALLAH** qui m'a permis d'atteindre cette étape de vie.

Je tiens à remercier ma directrice de thèse, Pr. KHAOUANE Latifa, et mon co-directeur de thèse, Pr. HANINI Salah, pour leur encadrement du projet de thèse.

Je tiens également à présenter ma profonde gratitude aux membres du jury :

Pr. Maamar LAIDI, professeur à l'UYFM pour présider mon jury de thèse.

Dr. Sofiane RAHAL Maître de Conférences (A) à l'UYFM pour avoir accepté de réviser ce travail et de participer à mon jury de thèse.

Pr. Yacine KERCHICHE Professeur à ENP, Alger pour avoir accepté de réviser ce travail et de participer à mon jury de thèse.

Merci à tous ceux qui m'ont aidé de près ou de loin durant ces années de recherche et d'études.

Résumé : Prédiction des affections transmissibles et non transmissibles par intelligence artificielle

Cette étude se concentre sur le développement des modèles prédictifs tirant parti de l'intelligence artificielle (IA) pour prédire les affections transmissibles virales et non transmissibles. La recherche est structurée en trois contributions distinctes, les deux premières centrées sur la prédiction des affections transmissibles et la troisième dédiée à la prédiction des affections non transmissibles.

Dans la première contribution, la thèse aborde la prédiction rapide du COVID-19 à l'aide d'un modèle de réseau neuronal artificiel à perceptrons multicouches (RNA-PMC). Le modèle appliqué a atteint une exactitude de 83,45 %, avec une aire sous la courbe de caractéristique de fonctionnement du récepteur (AUC-ROC) de 88,24 %, soulignant ainsi son efficacité dans la prédiction de cette maladie transmissible.

La deuxième contribution concerne la prédiction de l'hépatite virale en utilisant la même approche RNA-PMC. Le modèle développé a démontré une grande exactitude, atteignant une valeur de 95 %. De plus, il présentait une AUC-ROC de 95,12 %. Ces mesures d'évaluation complètes soulignent la robustesse du modèle dans la prédiction de l'hépatite virale, soulignant ainsi son utilité potentielle dans les établissements de soins de santé.

La troisième contribution se penche sur la prédiction des affections non transmissibles. Nous avons proposé une nouvelle méthode de prédiction du cancer du poumon utilisant des techniques d'intelligence artificielle. Nous avons évalué et comparé divers modèles d'apprentissage automatique, notamment RNA, K-plus proches voisins (KNN) et les machines à vecteurs de support (SVM), pour identifier le modèle le plus efficace. Le modèle SVM a atteint l'exactitude la plus élevée de 96,8 %, ainsi qu'une sensibilité de 98,33 %. Cependant, pour améliorer encore l'exactitude, nous avons introduit une approche d'optimisation hybride connue sous le nom Crow Search Algorithm- Artificial Bee Colony - Support Vector Machine (CSA-ABC-SVM). L'algorithme CSA- ABC-SVM combine les atouts de CSABC Optimiser et de SVM pour concevoir un modèle prédictif puissant. Notre étude a donné une exactitude remarquable de 98,38 %, ainsi qu'une sensibilité de 100 %. Ce résultat indique l'efficacité du modèle hybride CSA-ABC-SVM dans la prédiction du cancer du poumon, démontrant ses performances supérieures par rapport aux modèles d'apprentissage automatique traditionnels.

Cette étude met l'accent sur l'adaptabilité et l'efficacité de l'IA dans la prédiction de diverses affections, notamment la COVID-19, l'hépatite et le cancer du poumon. Cette exactitude remarquable souligne le potentiel de l'IA en tant qu'outil précieux pour l'anticipation et la prédiction précoces des affections. Ces résultats contribuent à des interventions de soins de santé proactives et basées sur les données pour gérer les futurs défis de santé.

Mots-clés : Affections transmissibles ; Affections non transmissibles ; réseau neuronal artificiel ; Machines à vecteurs de support ; Algorithme de recherche de corbeau ; Colonie d'abeilles artificielle.

Abstract: Prediction of communicable and non-communicable diseases by artificial intelligence

This study focuses on the development of predictive models leveraging artificial intelligence (AI) to predict viral communicable and non-communicable diseases. The research is structured into three distinct contributions, the first two focused on the prediction of communicable diseases and the third dedicated to the prediction of non-communicable diseases.

In the first contribution, the thesis addresses rapid prediction of COVID-19 using a multi-layer perceptron artificial neural network (RNA-PMC) model. The applied model achieved an accuracy of 83.45%, with an Area Under the Curve Receiver Operating Characteristic (AUC ROC) of 88.24%, highlighting its effectiveness in predicting this communicable disease.

The second contribution concerns the prediction of viral hepatitis using the same RNA-PMC approach. The developed model demonstrated high accuracy, reaching a value of 95%. Additionally, it had an AUC-ROC of 95.12%. These comprehensive evaluation metrics highlight the robustness of the model in predicting viral hepatitis, thus highlighting its potential utility in healthcare settings. This achievement confirms the ability of artificial intelligence to predict various infectious conditions.

The third contribution focuses on the prediction of non-communicable diseases. We proposed a new lung cancer prediction method using artificial intelligence techniques. We evaluated and compared various machine learning models, including ANN, K-nearest neighbors (KNN), and support vector machines (SVM), to identify the most effective model. The SVM model achieved the highest accuracy of 96.8%, along with a sensitivity of 98.33%. However, to further improve the accuracy, we introduced a hybrid optimization approach known as Crow Search Algorithm- Artificial Bee Colony - Support Vector Machine (CSA-ABC-SVM). The CSA-ABC-SVM algorithm combines the strengths of CSA-ABC Optimizer and SVM to design a powerful predictive model. Our study yielded a remarkable accuracy of 98.38%, along with a sensitivity of 100%. This result indicates the effectiveness of the hybrid CSA-ABC-SVM model in lung cancer prediction, demonstrating its superior performance compared to traditional machine learning models.

This study focuses on the adaptability and effectiveness of AI in predicting various conditions, including COVID-19, hepatitis, and lung cancer. This remarkable accuracy highlights the potential of AI as a valuable tool for early anticipation and prediction of conditions. These findings contribute to proactive, data-driven healthcare interventions to manage future healthcare challenges.

Keywords: Communicable diseases; Non-communicable diseases; Neural networks; Support vector machines; Crow search algorithm; Artificial bee colony

ملخص : التنبؤ بالأمراض المعدية وغير المعدية بواسطة الذكاء الاصطناعي

تركز هذه الدراسة على تطوير النماذج التنبؤية التي تستفيد من الذكاء الاصطناعي (AI) للتنبؤ بالامراض المعدية الفيروسية وغير المعدية. تم تنظيم البحث إلى ثلاث مساهمات متميزة، ركزت أول مساهمتين على التنبؤ بالامراض المعدية والثالثة مخصصة للتنبؤ بالامراض غير المعدية.

في المساهمة الأولى، تتناول الأطروحة التنبؤ السريع لكوفيد-19 باستخدام نموذج الشبكة العصبية الاصطناعية الإدراكية متعددة الطبقات (RNA-PMC). وحقق النموذج المطبق دقة بلغت 83.45%، مع مساحة المنطقة الواقعة تحت المنحنى المميزة لتشغيل جهاز الاستقبال (AUC-ROC) تبلغ 88.24%، مما يسلط الضوء على فعاليته في التنبؤ بهذا المرض المعدي.

وتتعلق المساهمة الثانية بالتنبؤ بالتهاب الكبد الفيروسي باستخدام نفس نهج RNA-PMC. وقد أظهر النموذج الذي تم تطويره دقة عالية، حيث وصلت إلى قيمة 95%. بالإضافة إلى ذلك، كان لديها المنطقة الواقعة أسفل المنحنى - خاصية تشغيل جهاز الاستقبال AUC-ROC بنسبة 95.12٪. تسلط مقاييس التقييم الشاملة هذه الضوء على قوة النموذج في التنبؤ بالتهاب الكبد الفيروسي، وبالتالي تسليط الضوء على فائدته المحتملة في أماكن الرعاية الصحية. ويؤكد هذا الإنجاز قدرة الذكاء الاصطناعي على التنبؤ بالامراض المعدية المختلفة.

وتركز المساهمة الثالثة على التنبؤ بالأمراض غير المعدية. اقترحنا طريقة جديدة للتنبؤ بسرطان الرئة باستخدام تقنيات الذكاء الاصطناعي. قمنا بتقييم ومقارنة نماذج التعلم الآلي المختلفة، بما في ذلك ANN، وأقرب جيران (KNN) K، وأجهزة ناقل الدعم (SVM)، لتحديد النموذج الأكثر فعالية. حقق نموذج SVM أعلى دقة بلغت 96.8%، إلى جانب حساسية قدرها 98.33%. ومع ذلك، لتحسين الدقة بشكل أكبر، قدمنا نهج تحسين مختلط يُعرف باسم خوارزمية بحث الغراب - مستعمرة النحل الاصطناعية - آلة ناقل الدعم (CSABC-SVM). تجمع خوارزمية CSABC-SVM بين نقاط قوة محسن CSABC وSVM لتصميم نموذج تنبؤي قوي. أسفرت دراستنا عن دقة ملحوظة بلغت 98.38%، إلى جانب حساسية قدرها 100%. تشير هذه النتيجة إلى فعالية نموذج CSABC-SVM الهجين في التنبؤ بسرطان الرئة، مما يدل على أدائه المتفوق مقارنة بنماذج التعلم الآلي التقليدية.

تركز هذه الدراسة على قدرة الذكاء الاصطناعي على التكيف وفعاليته في التنبؤ بالامراض المختلفة، بما في ذلك كوفيد-19 والتهاب الكبد وسرطان الرئة. تسلط هذه الدقة الرائعة الضوء على إمكانات الذكاء الاصطناعي كأداة قيمة للتوقع المبكر والتنبؤ بالامراض. تساهم هذه النتائج في تدخلات الرعاية الصحية الاستباقية والمبنية على البيانات لإدارة تحديات الرعاية الصحية المستقبلية.

الكلمات المفتاحية: الأمراض المعدية؛ الامراض غير المعدية؛ الشبكات العصبية الاصطناعية؛ آلة ناقل الدعم؛ خوارزمية بحث الغراب؛ مستعمرة النحل الاصطناعية

Table des matières

Remerciements

Résumé (Français, Anglais, Arabe)

Abréviations

Nomenclature et symbole

Liste des tableaux

Liste des figures

Introduction générale…….. 1

I LES AFFECTIONS TRANSMISSIBLES ET NON TRANSMISSIBLES..... 4

I.1 Introduction 4

I.2 I.2 Les affections transmissibles 4
I.2.1 I.2.1 Définitions des affections transmissibles..... 4
I.2.2 Les affections transmissibles les plus fréquentes..... 4
I.2.2.1 La grippe..... 4
I.2.2.2 La tuberculose..... 5
I.2.2.3 L'hépatite virale..... 6
I.2.2.4 COVID-19 6

I.3 Les affections non transmissibles 6
I.3.1 Définitions des affections non transmissibles..... 6
I.3.1.1 Diabètes 7
I.3.1.2 Les maladies cardiovasculaires..... 8
I.3.1.3 Le cancer du poumon 10
I.3.1.4 Le cancer du du foie 11
I.3.1.5 Cancer du pancréas..... 12

II CHAPITRE II. L'INTELLIGENCE ARTIFICIELLE 13

II.1 Introduction 13

II.2 Histoire de l'intelligence artificielle..... 13

II.3 Définition de l'intelligence artificielle 14

II.4 Principes fondamentaux de l'apprentissage automatique 14

II.4.1 Les types d'apprentissage automatique 15
II.4.1.1 Apprentissage supervise 15
II.4.1.1.1 Les Algorithmes d'apprentissage supervisé 16
II.4.1.1.1.1 Régression logistique 16
II.4.1.1.1.2 Réseaux de neurones artificiels 17
II.4.1.1.1.3 Algorithme de k-plus proches voisins 22
II.4.1.1.1.4 Les arbres de décision 23
II.4.1.1.1.5 Les forêts aléatoires 23
II.4.1.1.1.6 Machines à vecteurs de support (SVM) 24
II.4.1.2 L'apprentissage non supervisé 25
II.4.1.3 Apprentissage par renforcement 26

II.5 Apprentissage en profondeur 26
II.5.1 Réseau neuronal convolutif 26
II.5.2 Réseau neuronal récurrent 27

II.6 Algorithmes d'optimisation 27

II.7 Applications de l'intelligence artificielle en santé 28
II.7.1 Diagnostic médical amélioré 29
II.7.2 Prédiction des maladies 29
II.7.3 Recherche médicale 30
II.7.4 Télémédecine 30

II.8 Conclusion 31

III REVUE BIBLIOGRAPHIQUE SUR L'UTILISATION DE L'INTELLIGENCE ARTIFICIELLE DANS LA PREDICTION MEDICALE 32

III.1 Introduction 32

III.2 Les revues de littérature portant sur l'application de l'intelligence artificielle dans la prédiction médicale 32

III.3 Revue critique de la littérature sur la Prédiction des affection transmissible et non transmissible par l'Intelligence Artificielle 33

IV MATERIELS ET METHODES 39

IV.1 Introduction 39

IV.2 Logiciels utilisés 40

IV.3 Mesures d'évaluation des performances des modèles de classification 41

IV.4 Prédictions des affections transmissibles par IA 45
IV.4.1 Contribution I : Prédiction de COVID-19 par RNA-PMC 45
IV.4.1.1 Collecte de la base de données 45

IV.4.1.2 Prétraitement et analyse des données 46
IV.4.1.3 Méthode de classification 47
IV.4.1.4 Mesures d'évaluation des performances et le choix du meilleur modèle 48
IV.4.1.5 Analyse de sensibilité par la méthode de poids 48
IV.4.2 Contribution II : Prédiction de l'hépatite par PMC-RNA 49
IV.4.2.1 Collecte de la base de données 50
IV.4.2.2 Prétraitement et analyse des données 51
IV.4.2.3 Méthode de classification 52
IV.4.2.4 Analyse de sensibilité par la méthode de poids 52

IV.5 Prédictions des affections non transmissibles par IA 53
IV.5.1 Contribution III : Prédiction de cancer du poumon par CSABC-SVM 53
IV.5.1.1 Collecte de la base de données 53
IV.5.1.2 Prétraitement et analyse des données 54
IV.5.1.3 Méthode de classification 55
IV.5.1.4 Optimisation 55
IV.5.1.5 Importance des variables 57

IV.6 Conclusion 57

V RESULTATS ET DISCUSSION 59

V.1 Introduction 59

V.2 Prédiction des affections transmissibles 59
V.2.1 Contribution I : Prédiction de Covid-19 par RNA-PMC 59
V.2.1.1 Résultat de l'analyse exploratoire des données 59
V.2.1.2 Résultats du modèle proposé RNA- PMC 62
V.2.1.3 Analyse de sensibilité par méthode de poids 67
V.2.2 Contribution II : prédiction de l'hépatite virale par RNA- PMC 68
V.2.2.1 Résultats de l'analyse exploratoire des données 68
V.2.2.2 Résultats du modèle proposé RNA -PMC 70
V.2.2.3 Analyse de sensibilité par méthode de poids 74
V.2.2.4 Analyse comparative entre notre modèle et les modèles existants dans la littérature. 75

V.3 Prédiction des affections non transmissibles 76
V.3.1 Contribution III : Prédiction de cancer du poumon par CSABC-SVM 76
V.3.1.1 Résultat de l'analyse exploratoire des données 76
V.3.1.2 Résultats des modèles proposés 80
V.3.1.3 Résultats d'optimisation 84
V.3.1.4 Importance des variables 86
V.3.1.5 Interface graphique pour la prédiction de cancer du poumon 87
V.3.1.6 Comparaison avec d'autres études portant sur la prédiction du cancer du poumon 89

V.4 Conclusion 91
Conclusion générale et perspectives 95

Références bibliographiques ……. 97

Annexes …….……. 108

Valorisations scientifique : articles et séminaires …….……..….111

Liste des abréviations

ABC	Artificial Bee Colony
Ac	Accuracy
ACP – SRIA	Principal Component Analysis- Artificial immune recognition system
Alk	L'alkaline
ARN	Acide Ribonucléique
AUC	Area Under the Curve
BFGS	Broyden, Fletcher, Goldfarb et Shanno
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CPNPC	Cancer du Poumon Non à Petites Cellules
CPPC	Cancer du Poumon à Petites Cellules
CSA	Crow Search Algorithm
CT	Computed Tomography
DL	Deep Learning
DT	Decision Trees
FDA	Food and Drug Administration
FN	Faux Négatifs
FP	Faux Positifs
AG	Algorithme Génétique
GPU	Graphics Processing Units
IA	Intelligence Artificielle
IBM	International Business Machines
IR	Importance Relative

KDE	Kernel Density Estimation
K-NN	K-Nearest Neighbors
MCC	Coefficient de Corrélation de Matthews
MCV	Maladies Cardio-Vasculaires
NB	Naive Bayes
ACP – AIRS	Principal Component Analysis and Artificial Immune Recognition System
PCA–LSSVM	Principal Component Analysis and Least Square Support Vector Machine
PID	Pima Indians Diabetes
PMC	Perceptron Multi-Couches
Pr	Précision
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RF	Random Forest
RNA	Réseaux de Neurones Artificiels
RNN	Recurrent Neural Network
RNRG	Réseau Neuronal de Régression Générale
ROC	Receiver Operating Characteristic
RotF	Forêt de Rotation
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
Se	Sensibilité
SIDA	Syndrome d'Immuno Déficience Acquise
SGD	Stochastic Gradient Descent
SGOT	Serum Glutamic Oxaloacetic Transaminase
SMOTE	Synthetic Minority Over-Sampling Technique
SOS	Sum Of Squares

Sp	Spécificité
SVM	Support Vector Machines
UCI	Université de Californie à Irvine
VHA	Virus de l'Hépatite A
VHB	Virus de l'Hépatite B
VHC	Virus de l'Hépatite C
VHD	Virus de l'Hépatite D
VHE	Virus de l'Hépatite E
VIH	Virus de l'Immunodéficience Humaine
VN	Vrais Négatifs
VP	Vrais Positifs
VPN	Valeur Prédictive Négative
VPP	Valeur Prédictive Positive
WOA	Whale Optimization Algorithms

Nomenclature et symbole

a_i	Potentiel du neurone
$\mathbf{b_i}$	Biais d'un neurone artificiel (i)
C	Paramètres de pénalité
e	Fonction exponentielle
Eq	Equation
f	Activation function
F	Fonction
i	Elément i d'un vecteur
n	Variables des modèles ANN
W	Weight (poids synaptique des connexions inter-neuronales)
$\boldsymbol{X}$	Entrée de réseau (variables d'entrée)
Y	Sortie de réseau
γ	la bande passante du noyau gaussien

Liste des tableaux

Chapitre II

Tableau. II.1. Fonctions d'activation usuelles existantes dans le logiciel Statistica …… 20

Chapitre IV

Tableau IV. 1 Les logiciels utilisés……40

Tableau IV. 2 Description statistique des attributs de la base de données COVID-19……..46

Tableau IV. 3 Description statistique des attributs de la base de données d'hépatite……51

Tableau IV. 4 Description des attributs de la base de données de cancer du poumon ……. 54

Chapitre V

Tableau V.1 Architecture du meilleur modèle RNA-PMC pour la prédiction de COVID-19 …… 64

Tableau V. 2 Performances du meilleur modèle de la prédiction de Covid-19 par RNA-PMC 66

Tableau V.3 Architecture du meilleur modèle RNA-PMC pour la prédiction de l'hépatite virale ..72

Tableau V. 4 Performances du meilleur modèle de la prédiction de l'hépatite virale par RNA-PMC ……... 74

Tableau V. 5. Comparaison de la prédiction de l'hépatite avec d'autres études de littérature …… 76

Tableau V. 6 Performances des modèles de classification ……..… 81

Tableau V. 7. Performances de prédiction de cancer du poumon par CSABC-SVM …….. 84

Tableau V. 8 Paramètres de modèle CSABC- SVM via la prédiction de cancer du poumon…….86

Tableau V. 9. Comparaison du présent travail via la littérature ……90

Liste des figures

Chapitre I

Figure I.1 Illustration des différents types de diabète …….. 8

Figure I.2 Illustration visuelle des altérations vasculaires : Comparaison entre une artère normal et une artère rétrécie par l'athérome …..…….…… 10

Figure I.3 Facteurs de risque des maladies cardiovasculaire …….…….… 10

Figure I.4 Evolution du cancer des poumons …..……. 11

Chapitre II

Figure II.1 Illustration de l'apprentissage supervisé …..……..… 16

Figure II. 2 Neurone biologique……..…….…… 18

Figure II. 3 Neurone artificiel ……..……...…… 18

Figure II 4 Comparaison entre un neurone biologique et un neurone artificiel …..…… 18

Figure II.5 Architecture d'un réseau neuronal perceptron multicouches …..…….. 22

Figure II.6 Algorithme des k-plus proches voisins…..…… 23

Figure II.7 Algorithme de SVM …..…… 25

Figure II.8 Apprentissage non supervisé ……..26

Chapitre III

Figure III.1 Nombre de publications sur la prédiction des maladies par l'intelligence artificielle

(Science direct 14 / 02 /2024) ……33

Chapitre IV.

Figure IV.1 Matrice de confusion pour une classification binaire ……41

Figure IV. 2 Illustration de la courbe ROC ……44

Figure IV. 3 Diagramme de la procédure appliquée pour la prédiction de COVID-19 par RNA-PMC ……...45

Figure IV. 4. Capture d'écran des fonctions d'activation de STATISTICA ……..47

Figure IV. 5 Organigramme via la méthode d'analyse de sensibilité par la méthode de poids ...49

Figure IV. 6 Diagramme de la procédure appliquée pour la prédiction de l'hépatite virale par RNA-PMC ……..…….…….…….....50

Figure IV. 7 Structure organisationnelle de la procédure appliquée pour la prédiction de cancer du poumon ……..53

Chapitre V.

Figure V. 1 Matrice de corrélation de la prédiction de COVID-19 par RNA-PMC ……..…....60

Figure V. 2 Analyse de la distribution des classes (infecté, non infecté) de COVID-19 ……....61

Figure V. 3 Répartition par âge des patients avec et sans COVID-19 ……..……61

Figure V. 4 Relation de différentes variables catégorielles avec le COVID-19 …..……..…62

Figure V. 5 Présentation de l'exactitude pour chaque division de la base de données ……..…… 63

Figure V. 6 Représentation des meilleurs pourcentages de division de la base de données de COVIDE-19…….. 63

Figure V. 7 Matrice de confusion de la prédiction de Covid-19 par RNA-PMC ……..…….65

Figure V. 8 Courbe ROC pour la prédiction de COVID-19 par RNA-PMC ……..……..….66

Figure V. 9 Histogramme d'importance relative pour le COVID-19….……..……..67

Figure V.10 Matrice de corrélation de la prédiction de l'hépatite virale par RNA-PMC …..…...68

Figure V. 11 Analyse de la distribution des classes (oui, non) d'hépatite virale ……..…….....69

Figure V.12 Répartition par âge des deux classes de patients atteints d'hépatite virale ……..... 69

Figure V.13 Répartition par sexe des deux classes de patients atteints d'hépatite virale …..... 70

Figure V. 14 Présentation de l'exactitude pour chaque division de la base de données ……..... 71

Figure V. 15 Présentation des meilleurs pourcentages de division de la base de données …….. 71

Figure V.16 Matrice de confusion pour la prédiction de l'hépatite virale par RNA-PMC ……..... 73

Figure V.17 Courbe ROC pour la prédiction de l'hépatite virale par modèle RNA-PMC …..….. 74

Figure V.18 Histogramme d'importance relative pour la prédiction de l'hépatite virale par RNA-PMC ……. 75

Figure. V.19 Matrice de corrélation montrant les corrélations entre tous les attributs de l'ensemble de données. ……..……. 77

Figure. V. 20 Analyse de la distribution des classes (oui, non) de cancer du poumon …… 78

Figure. V. 21 Graphiques KDE pour les patients avec et sans cancer du poumon en fonction de la répartition d'âge …….……. 78

Figure. V. 22 Analyse de la relation entre différentes variables catégorielles et la variable Cancer de poumon ……..……. 80

Figure V. 23 Comparaison de l'exactitude entre les trois modèles (RNA, KNN et SVM) …….. 82

Figure V. 24 Courbe ROC via RNA ……83

Figure V. 25 Courbe ROC via KNN……. 83

Figure V. 26 Courbe ROC via SVM……. 84

Figure V. 27 Comparaison des performances de SVM et CSABC-SVM pour la prédiction de cancer du poumon …… 85

Figure V. 28 Courbe ROC pour CSABC-SVM…… 86

Figure V. 29 Illustration de l'importance des variables via CSABC-SVM…….. 87

Figure V. 30 Interface graphique de cancer du poumon via CSABC-SVM…… 88

Figure V. 31 Interface graphique de prédiction positive de cancer du poumon via CSABC-SVM ……. 89

Introduction générale

Introduction générale

La santé, en tant que fondement essentiel du bien-être humain, transcende les frontières individuelles pour devenir une préoccupation primordiale à l'échelle mondiale [1]. Dans ce contexte, la compréhension profonde des mécanismes de préservation de la santé et de la prévention des affections revêt une importance cruciale.

Au cœur des préoccupations de la santé publique mondiale se trouvent les affections, une catégorie vaste et complexe comprenant à la fois les affections transmissibles et non transmissibles. Les affections transmissibles, telles que les infections virales ou bactériennes, ont toujours constitué une menace constante pour la santé mondiale [2]. L'incidence croissante de ces maladies, exacerbée par la mobilité accrue des populations, souligne La nécessité d'adopter des approches innovantes pour prédire, surveiller et contrôler leur propagation est indispensable [3]. D'un autre côté, les affections non transmissibles, telles que les maladies cardiovasculaires, le diabète et certains cancers, et les affections chroniques des voies respiratoires sont devenues des défis majeurs de santé publique, nécessitant une gestion proactive pour réduire leur prévalence et minimiser leur impact sur la qualité de vie [4].

L'avènement de l'intelligence artificielle (IA) a révolutionné de nombreux domaines [5], et le secteur de la santé n'échappe pas à cette transformation. Les progrès spectaculaires de l'IA, rendus possibles par l'essor du "Big Data" et l'augmentation de la puissance de calcul des processeurs [6]. À l'échelle globale du système de soins, l'IA pourrait générer des économies et améliorer les performances médicales en personnalisant les prises en charge, améliorant la qualité des soins, et renforçant la sécurité et la traçabilité des soins [7]. Les possibilités s'étendent également à la recherche clinique, à la formation médicale, et à la gestion des bases de données, ouvrant la voie à une médecine plus préventive et prédictive. L'utilisation de l'IA dans la prédiction des affections, qu'elles soient transmissibles ou non transmissibles, représente une avancée significative dans la manière dont nous abordons la prévention, le diagnostic précoce et la gestion des maladies. Des travaux ont démontré son efficacité dans la détection précoce de maladies comme le cancer et les maladies cardiaques, ainsi que dans la prévision des épidémies de maladies transmissibles telles que la grippe et la COVID-19. Ces avancées permettent des interventions médicales plus ciblées et efficaces.

L'objectif fondamental de cette thèse est d'explorer le potentiel de l'intelligence artificielle dans la prédiction des affections, et la détection précoce. En examinant les avancées récentes en matière de modélisation prédictive, d'apprentissage automatique et de traitement des données

massives, cette recherche vise à éclairer les possibilités offertes par l'IA pour relever ces défis complexes.

Le choix d'utiliser des réseaux de neurones artificiels , des machines à vecteurs de support et le k plus proches voisins, algorithme de colonie d'abeilles, et l'algorithme de recherche de corbeau dans notre étude découle des avantages principaux à ces approches. Les réseaux de neurones artificiels offrent une capacité exceptionnelle d'apprentissage complexe, simulant le fonctionnement du cerveau humain et permettant la modélisation. Les machines à vecteurs de support excellent dans la gestion des données complexes et la classification précise, tandis que l'utilisation du KNN garantit une adaptation flexible aux schémas de données changeants. L'algorithme de colonie d'abeilles, et l'algorithme de recherche de corbeau, introduisent des stratégies d'optimisation novatrices, idéales pour résoudre des problèmes complexes. Ensemble, ces choix méthodologiques visent à exploiter au mieux les forces spécifiques de chaque approche, créant ainsi une base méthodologique solide pour notre étude de prédiction des affections transmissibles et non transmissibles.

La méthodologie utilisée dans cette thèse se compose de cinq chapitres, chacun élaboré de manière à explorer systématiquement et à contribuer à notre compréhension de la prédiction des affections transmissibles et non transmissibles par l'intelligence artificielle.

- Le premier chapitre de cette thèse offre une exploration approfondie des affection transmissibles et non transmissibles
- Deuxième chapitre offre une plongée approfondie dans l'évolution historique, les définitions essentielles et les diverses techniques d'IA. Il propose une analyse approfondie des méthodologies clés, notamment les réseaux de neurones artificiels, les machines à vecteurs de support, les procédures K-plus proches voisins, et les algorithmes d'optimisation. Cette exploration méthodologique vise à établir une base solide pour l'application pratique de ces techniques dans les chapitres subséquents.
- Le chapitre 3 s'engage dans une revue critique des études existantes relatives à la prédiction médicale par l'IA, extraites de la littérature spécialisée. En examinant de manière approfondie les travaux antérieurs, cette revue vise à identifier les tendances émergentes, les défis persistants et les avancées significatives dans le domaine de la prédiction médicale.
- Le passage à l'application pratique de cette étude se réalise dans le chapitre 4, qui détaille également les matériaux utilisés et explicite les différentes techniques mises en œuvre.

Cette phase englobe la collecte minutieuse de données, le prétraitement et l'analyse des données, tout en mettant en lumière les étapes impliquées dans la construction du modèle.

- Le chapitre 5 termine ce parcours de recherche en présentant les résultats issus de nos études approfondies. La comparaison des performances des modèles développés souligne leur efficacité dans la prédiction des affections transmissibles et non transmissibles. De plus, une comparaison rigoureuse avec les modèles existants dans la littérature permet de d'identifier les avancées spécifiques et les contributions singulières que nos travaux apportent au domaine de la prédiction médicale par l'IA. Ce chapitre clé consolide ainsi l'impact de notre recherche et offre des perspectives significatives pour l'évolution future de ce domaine en constante évolution.

Chapitre I

Les Affections Transmissibles et non Transmissibles

I Les affections transmissibles et non transmissibles

I.1 Introduction

Le présent chapitre se consacre à l'étude approfondie des affections transmissibles et non transmissibles, deux catégories majeures de maladies qui jouent un rôle crucial dans la compréhension de la santé publique et des défis auxquels sont confrontés les systèmes de soins à l'échelle mondiale. La dichotomie entre ces deux types d'affections reflète la diversité des mécanismes et des facteurs qui influent sur la santé humaine, allant des agents pathogènes microscopiques aux déterminants génétiques et environnementaux.

I.2 I.2 Les affections transmissibles

I.2.1 I.2.1 Définitions des affections transmissibles

Les affections transmissibles, également appelées maladies infectieuses, sont des maladies causées par la présence et la propagation d'agents pathogènes spécifiques, tels que des bactéries, des virus, des parasites ou des champignons, d'une personne ou d'un organisme infecté à une autre [8]. Ces maladies se propagent principalement par divers mécanismes de transmission, notamment le contact direct entre individus, le contact indirect par l'intermédiaire d'objets contaminés, l'inhalation d'aérosols infectés, la transmission par vecteurs (comme les moustiques pour le paludisme) [8]. Les affections transmissibles sont généralement caractérisées par leur potentiel de contagion et leur capacité à provoquer des épidémies ou des pandémies si elles ne sont pas contrôlées efficacement.

Exemples courants d'affections transmissibles : Grippe (influenza), Tuberculose, COVID-19, VIH/SIDA, Hépatite B et C, Rougeole, Varicelle, Choléra, Paludisme, Typhoïde, Peste, Méningite, Fièvre typhoïde, etc.

La prévention des affections transmissibles repose souvent sur des stratégies telles que la vaccination, l'hygiène personnelle, la quarantaine, les mesures de contrôle des infections et, dans certains cas, des traitements antiviraux ou antibiotiques spécifiques

I.2.2 Les affections transmissibles les plus fréquentes

I.2.2.1 La grippe

La grippe est une infection respiratoire causée par le virus de l'influenza. Elle se caractérise par une inflammation des voies respiratoires supérieures [9]. La grippe est différente du rhume commun et peut être plus grave, surtout chez les populations vulnérables, comme les jeunes enfants, les personnes âgées et les individus atteints de problèmes de santé. La grippe de type A se

divise davantage en plusieurs sous-types en fonction des propriétés antigéniques de l'hémagglutinine et de la neuraminidase, deux glycoprotéines de la surface [10].

- **Symptômes** : Les symptômes de la grippe peuvent varier en gravité, mais ils incluent généralement :

  Fièvre élevée soudaine, frissons et tremblements, maux de tête, fatigue sévère, toux , gorge douloureuse, congestion nasale, douleurs musculaires et articulaires, écoulement nasal, faiblesse générale [9].

I.2.2.2 La tuberculose

La tuberculose est une maladie contagieuse causée par la bactérie Mycobacterium tuberculoses [11,12]. Elle affecte principalement les poumons, mais lorsqu'elle se présente sous forme extra pulmonaire, la tuberculose peut affecter d'autres organes ou systèmes, en particulier chez ceux dont le système immunitaire est affaibli [12].

La tuberculose se propage principalement par inhalation de gouttelettes respiratoires contenant la bactérie émises par une personne infectée lorsqu'elle tousse, éternue, parle ou respire. Elle est généralement transmise d'une personne à une autre par des contacts étroits et prolongés [11].

De nombreuses recherches ont souligné comment une insuffisance ou un manque de vitamine D peut augmenter le risque de contracter la tuberculose [13].

- **Symptômes**

Les symptômes de la tuberculose peuvent varier en fonction de la forme de la maladie et de la personne infectée, mais ils peuvent inclure

Toux persistante, Fièvre et sueurs nocturnes, Perte de poids non intentionnelle, Fatigue et faiblesse générale, Douleurs thoraciques, Essoufflement, En cas de tuberculose extra pulmonaire, des symptômes spécifiques à l'organe affecté.

- **Les facteurs de risque pour la tuberculose comprennent**

Le contact étroit avec une personne infectée.

Un système immunitaire affaibli (par le VIH/SIDA, la malnutrition, etc.).

Des conditions de vie surpeuplées.

L'abus de drogues ou d'alcool.

Les voyages dans des zones où la tuberculose est fréquente

I.2.2.3 L'hépatite virale

C'est une inflammation du foie qui peut être causée par divers facteurs de santé. Le virus de l'hépatite se décline en cinq souches principales, appelées virus de l'hépatite A (VHA), virus de l'hépatite B (VHB), virus de l'hépatite C (VHC), virus de l'hépatite D (HDV) et virus de l'hépatite E (VHE) [14]. Les hépatites B et C sont davantage associées aux maladies chroniques et sont considérées comme les principales causes de décès par hépatite virale [15].

– **Symptômes**

Fatigue, jaunisse (coloration jaune de la peau et des yeux), douleurs abdominales, perte d'appétit, nausées, vomissements, fièvre, etc. Cependant, dans de nombreux cas, l'hépatite peut être asymptomatique.

I.2.2.4 COVID-19

C'est une maladie respiratoire qui peut varier en gravité, allant de cas bénins à des formes graves, voire mortelles. La COVID-19 est causée par le SARS-CoV-2, un virus de la famille des coronavirus [16]. Il s'agit d'un virus à ARN [17], qui peut muter au fil du temps, conduisant à l'émergence de nouvelles souches.

Les principales voies de transmission du virus sont la respiration de gouttelettes infectieuses, le contact direct avec une personne infectée lors de la toux, de l'éternuement et la respiration.

- **Symptômes**

Fièvre ou frissons, toux sèche, essoufflement ou difficultés respiratoires, fatigue, douleurs musculaires ou corporelles, maux de gorge, perte du goût ou de l'odorat, congestion nasale ou nez qui coule, maux de tête, nausées ou vomissements, diarrhée, une pneumonie virale, une insuffisance respiratoire.

I.3 Les affections non transmissibles

I.3.1 Définitions des affections non transmissibles

Les affections non transmissibles, également appelées maladies chroniques, sont des maladies qui ne se propagent pas d'une personne à une autre par un agent pathogène spécifique. Elles résultent principalement de facteurs internes ou externes qui affectent la santé d'un individu, tels que la génétique, les habitudes de vie, l'environnement et le vieillissement. Contrairement aux affections transmissibles, les affections non transmissibles se développent souvent

progressivement sur une longue période et peuvent persister tout au long de la vie d'un individu. Ces maladies sont responsables d'une grande partie de la morbidité et de la mortalité à l'échelle mondiale, et elles imposent un fardeau important sur les systèmes de santé.

Exemples courants d'affections non transmissibles : maladies cardiovasculaires, cancer, diabète, maladies respiratoires chroniques, maladies neurologiques, obésité, maladies rénales chroniques, maladies métaboliques, maladies oculaires.

I.3.1.1 Diabètes

C’est une maladie chronique caractérisée par un taux élevé de glucose dans le sang (hyperglycémie) [18], résultant d'une production insuffisante d'insuline par le pancréas ou d'une résistance des cellules à l'action de l'insuline. L'insuline est une hormone essentielle qui régule la concentration de glucose dans le sang et permet aux cellules d'utiliser le glucose comme source d'énergie. Il existe plusieurs types de diabète, les principaux étant le diabète de type 1, le diabète de type 2 et le diabète gestationnel. Le diabète de type 1 (autrefois appelé diabète insulinodépendant) survient lorsqu’il y a un manque de sécrétion d’insuline [19]. Le diabète de type 2 modifie la façon dont l’organisme utilise le glucose comme source d’énergie. Il perturbe la capacité du corps à utiliser correctement l'insuline, pouvant entraîner une élévation du taux de sucre dans le sang s'il n'est pas traité [19]. Le diabète gestationnel survient pendant la grossesse [19]. L'hyperglycémie entraîne des complications à long terme, notamment la rétinopathie diabétique, qui entraîne une perte de vision, et la neuropathie, qui provoque des lésions nerveuses avec un risque d'ulcères du pied. Amputation, néphropathie entraînant une insuffisance rénale, des symptômes cardiovasculaires et finalement une mort prématurée [20, 21].

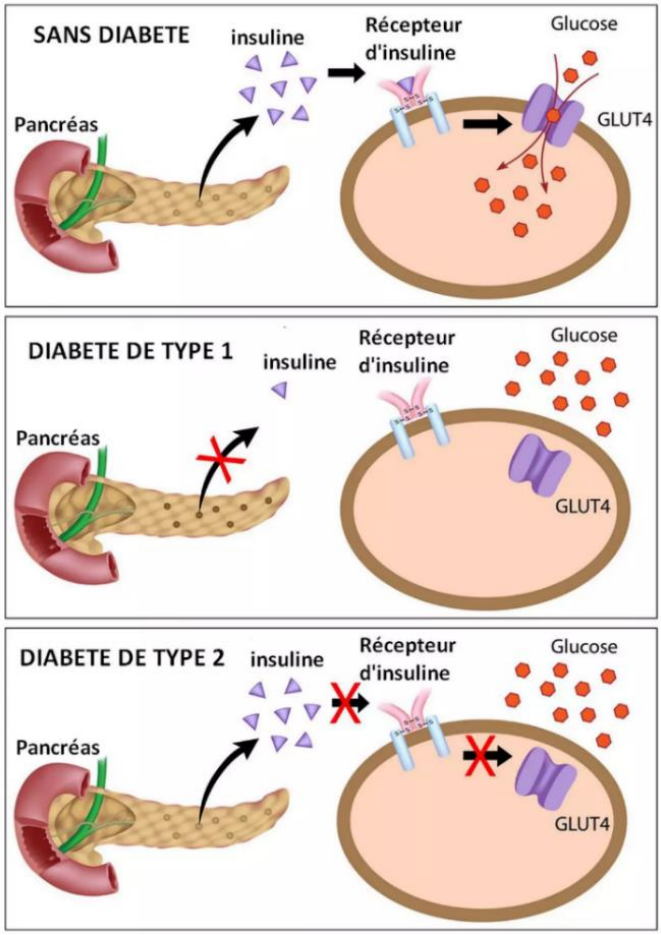


Figure I.1 Illustration des différents types de diabète [22]

- **Facteurs de risque**

  Les antécédents familiaux, l'obésité, un mode de vie sédentaire, une mauvaise alimentation, le vieillissement, et des facteurs génétiques.

- **Symptômes**

  Une soif excessive, le besoin d'uriner plus souvent que d'habitude, une fatigue persistante, une perte de poids involontaire, et des infections fréquentes [19].

I.3.1.2 Les maladies cardiovasculaires

Sont un groupe de troubles affectant le cœur et les vaisseaux sanguins [23] Elle est généralement liée à l'accumulation de dépôts graisseux dans les artères, appelée athérosclérose, et à un risque élevé de formation de caillots sanguins [23]. Elle peut aussi être liée à des lésions des artères d'organes tels que le cerveau, le cœur, les yeux, et les reins [18]. Elles représentent l'une des principales causes de décès dans le monde [24,25,26]. Voici quelques-unes des maladies cardiovasculaires les plus courantes

- La maladie coronarienne survient lorsqu'il y a un blocage ou une réduction du flux sanguin riche en oxygène vers les muscles du cœur [23].
- Les maladies cérébro-vasculaires sont des affections qui affectent les vaisseaux sanguins du cerveau [27]. Ces maladies impliquent généralement des problèmes de circulation sanguine dans le cerveau
- Les cardiopathies rhumatismales surviennent lorsque le muscle cardiaque et les valves sont affectés à la suite d'un rhumatisme articulaire aigu, causé par une bactérie streptocoque [27].

– **Symptômes**

Essoufflement, fatigue, arythmie cardiaque, douleur thoracique et syncope pour une cardiopathie rhumatismale

- Les malformations cardiaques congénitales sont des anomalies structurelles du cœur présentes dès la naissance.
- Les thromboses veineuses profondes et les embolies pulmonaires se caractérisent par le blocage des veines des jambes causé par la présence de caillots sanguins, pouvant se détacher et migrer vers le cœur ou les poumons.
- Les thromboses veineuses profondes et les embolies pulmonaires se caractérisent par le blocage des veines des jambes causé par la présence de caillots sanguins, qui peuvent se détacher et migrer vers le cœur ou les poumons

– **Facteurs de risque**

Le tabagisme, une alimentation riche en graisses saturées et en cholestérol, l'inactivité physique (la sédentarité), l'obésité, l'hypertension artérielle, le diabète, la consommation d'alcool l'âge, et les antécédents familiaux de maladies cardiaques comme illustré dans la Figure I.3.

- **Symptômes**

Douleur thoracique, essoufflement, fatigue, palpitations, vertiges ou étourdissements, gonflement des extrémités, douleurs dans les bras, le dos, la mâchoire ou le cou, sensation de brûlure à l'estomac, nausées, perte de conscience ou évanouissements.

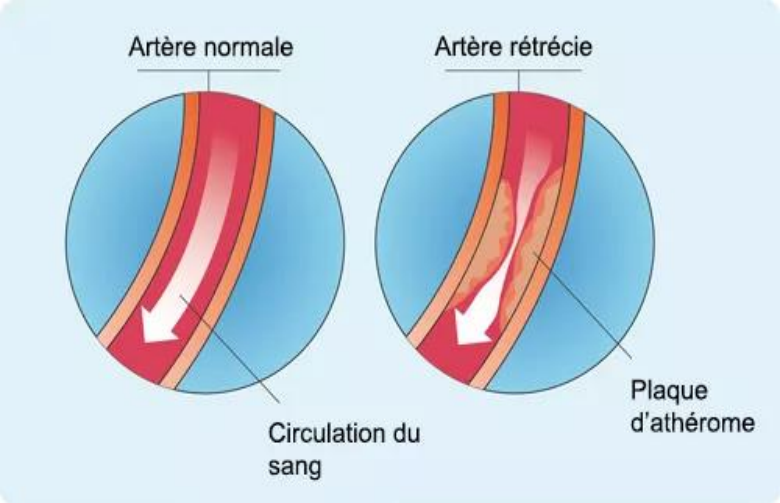


Figure I.2 Illustration visuelle des altérations vasculaires : Comparaison entre une artère normal et une artère rétrécie par l'athérome [28]

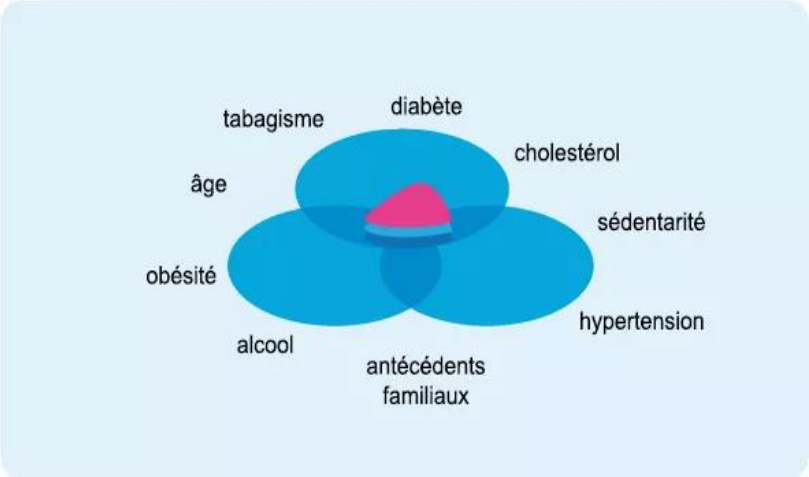


Figure I.3 Facteurs de risque des maladies cardiovasculaire [28]

I.3.1.3 Le cancer du poumon

Se caractérise par le développement incontrôlé de cellules anormales qui débute dans les tissus pulmonaires comme illustré dans la figure I.2. Généralement dans les cellules qui tapissent les bronches. Il est l'une des formes les plus courantes de cancer et est souvent associé au tabagisme.

Comme il prend naissance dans la muqueuse des bronches, le cancer du poumon obstrue les échanges gazeux dans les poumons et provoque des difficultés respiratoires [29]. Difficile à évaluer et à découvrir dès les premiers stades car cela ne crée pas de douleur ni de symptômes dans certaines situations [30].

Il existe deux principales catégories de cancer du poumon : le cancer du poumon à petites cellules (CPPC) et le cancer du poumon non à petites cellules (CPNPC). Le CPNPC est le type le plus fréquent, représentant environ 85 % des cas de cancer du poumon.

- **Les facteurs de risque**

Le tabagisme (le principal facteur de risque), l'exposition à des substances cancérigènes comme l'amiante et le radon, des antécédents familiaux de cancer du poumon, ainsi que d'autres conditions pulmonaires préexistantes.

- **Les symptômes**

Une toux persistante, des crachats de sang, un essoufflement, une douleur thoracique, une perte de poids inexpliquée et une fatigue importante, Syndrome de Pancoast (douleur à l'épaule), enrouement (paralysie des cordes vocales), faiblesse [30].

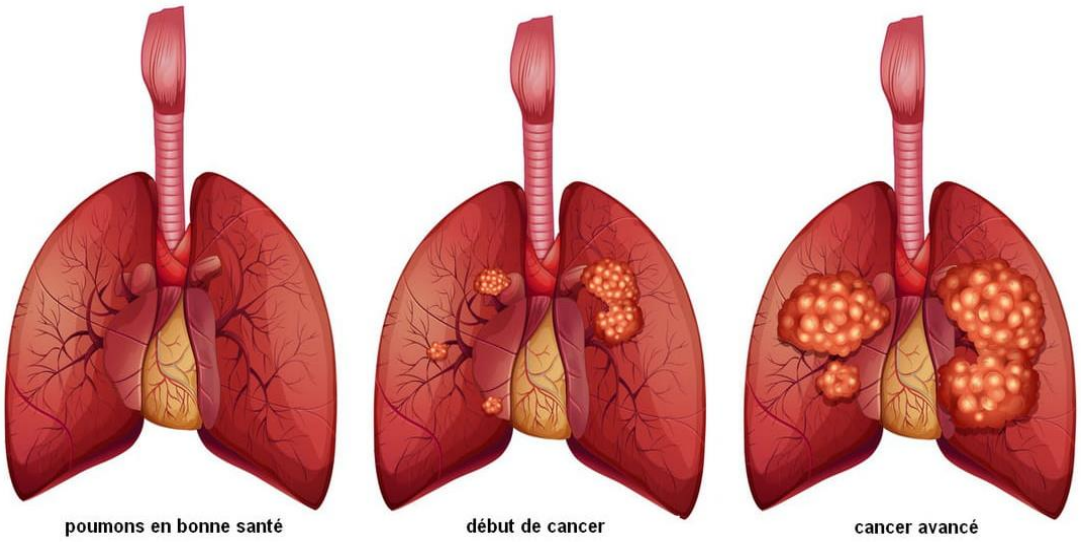


Figure I.4 Evolution du cancer des poumons [31]

I.3.1.4 Le cancer du foie

Le cancer du foie, ou carcinome hépatocellulaire (CHC), est l'un des types les plus courants de cancer primitif du foie, également connu sous le nom d'hépatome. Les principales causes du CHC sont les infections par le virus de l'hépatite B (virus à ADN) ou de l'hépatite C (virus à ARN), ainsi que la consommation excessive d'alcool [32].

Le type expansif est la forme la plus courante de CHC. Il se manifeste par une masse unique, volumineuse, de couleur jaune-brun, généralement située dans le lobe droit du foie [32].

Le type multifocal est une forme moins courante de CHC caractérisée par la présence de plusieurs masses tumorales, de 3 à 5 cm de diamètre, dispersées dans tout le foie [32].

Le type infiltrant est une forme rare de carcinome hépatocellulaire CHC où les cellules tumorales se diffusent à travers l'ensemble du foie [32].

- **Les facteurs de risque**

Hépatite B et C chroniques cirrhose, consommation excessive d'alcool, stéatohépatite non alcoolique (NASH), exposition aux aflatoxines, diabète, obésité, facteurs génétiques, tabagisme

- **Les symptômes**

Douleurs abdominales, perte de poids, perte d'appétit, nausées et vomissements, fatigue, jaunisse, enflure de l'abdomen, démangeaisons.

I.3.1.5 Cancer du pancréas

La majorité des tumeurs malignes du pancréas sont des adénocarcinomes. Parmi les néoplasmes pancréatiques rares, on trouve les tumeurs neuroendocrines, qui peuvent sécréter des hormones comme l'insuline ou le glucagon, et les carcinomes acineux, qui peuvent libérer des enzymes digestives dans la circulation. Des néoplasmes encore moins fréquents incluent les carcinomes colloïdes, les pancréatoblastomes et les néoplasmes pseudopapillaires solides. Plus précisément, l'adénocarcinome canalaire est le type de tumeur maligne pancréatique le plus répandu. [33].

- **Les facteurs de risque**

Les facteurs de risque de cette maladie maligne comprennent le tabagisme, les antécédents familiaux de pancréatite chronique, le diabète ,le sexe masculin, l'âge avancé, l'obésité, le groupe sanguin non O, les expositions professionnelles, l'origine ethnique afro-américaine, un régime riche en viande et pauvre en légumes et en folate, un régime alimentaire riche en graisses, et éventuellement une infection à Helicobacter pylori et une maladie parodontale [34].

- **Les symptômes**

Douleur abdominale, perte de poids, jaunisse, nausées et vomissements, appétit réduit, changements dans les selles, diabète de novo, fatigue, problèmes digestifs

Conclusion

L'étude des affections transmissibles et non transmissibles révèle la complexité inhérente à la santé humaine. La coexistence de ces deux catégories souligne la nécessité d'une approche intégrée en matière de santé publique, capable de faire face simultanément aux épidémies infectieuses et aux défis persistants des maladies chroniques. En comprenant les mécanismes de transmission et les facteurs sous-jacents, nous sommes mieux outillés pour élaborer des stratégies

préventives et curatives efficaces, contribuant ainsi à améliorer la santé mondiale dans son ensemble.

Chapitre II

L'intelligence Artificielle

II Chapitre II. L'intelligence artificielle

II.1 Introduction

Le terme Intelligence Artificielle s'est généralisé parmi le grand public à mesure que les progrès de la technologie informatique se propagent dans divers aspects de nos vies. L'IA a désormais pris une place prépondérante, comme en témoigne la profusion d'articles et d'ouvrages qui lui sont consacrés, reflétant ainsi son influence directe ou indirecte sur de nombreuses domaines d'activités humaines. Les techniques d'intelligence artificielle représentent des méthodes numériques qui exploitent les connaissances. Les principales techniques de l'IA comprennent la représentation des connaissances, la recherche de solutions, le raisonnement et l'apprentissage (ou formation). Le présent chapitre vise à décrire les notions fondamentales liées à l'intelligence artificielle (IA)

II.2 Histoire de l’intelligence artificielle

En 1956, lors d'un séminaire organisé à Dartmouth aux États-Unis, le terme "intelligence artificielle" a été formellement introduit, et c’est à John McCarthy qu’on attribue la paternité du nom de cette nouvelle discipline [35]. Cependant Warren McCulloch et Walter Pitts sont considérés comme des précurseurs dans les travaux appartenant à l’IA [35]. Leur inspiration découle de la physiologie et du fonctionnement des neurones cérébraux, de l'analyse formelle de la logique propositionnelle établie par Russell et Whitehead, et de la théorie du calcul de Turing [36].

En 1943, Un modèle de neurones artificiels a été défini par McCulloch et Pitts [36].

En 1957, Frank Rosenblatt a introduit le perceptron, un algorithme d’apprentissage qui simulait le fonctionnement des neurones [37]. L’objectif premier du perceptron était de réaliser la classification d'images [37].

L'IA a connu des périodes de désintérêt et de désengagement financier, notamment dans les années 1970, à la suite des échecs de l’IA dans le domaine du langage, de la traduction automatique marquant ce que l'on a appelé "l'hiver de l'IA". Après un regain d'intérêt lié à la montée des ordinateurs personnels et à l'augmentation de la puissance de calcul, l'IA a subi une deuxième période de déclin à la fin des années 1980. Cependant, l'avènement d'Internet au milieu des années 1990 et la profusion de données qui en a découlé ont relancé l'IA [37].

En 2001, l’apparition des premières cartes graphiques dédiées pour Graphics Processing Units (GPU) a simultanément renforcé la capacité de calcul des ordinateurs [37].

II.3 Définition de l’intelligence artificielle

L'intelligence artificielle est définie comme un ensemble de technologies permettant aux ordinateurs de simuler et d'imiter l'intelligence humaine [37]. L'intelligence artificielle confère à une machine la compétence d'effectuer des traductions de texte, de reconnaître des images, d'automatiser la conduite de véhicules, ainsi que de superviser le pilotage de procédés industriels [37]. L'approche cognitive vise à explorer les mécanismes des raisonnements humains et à les traduire en algorithmes [38]. Dans l'approche pragmatiste, l'intelligence artificielle est perçue comme une boîte noire [38]. Elle prend en entrée les données du problème et fournit en sortie les résultats, sans nécessité de comprendre le fonctionnement interne de cette boîte [38].

L'intelligence artificielle est constituée de plusieurs composants ou briques scientifiques qui travaillent de concert pour permettre aux systèmes d'IA d'accomplir des tâches intelligentes. Ces briques comprennent, entre autres, l'apprentissage automatique, les mathématiques, les statistiques, la planification, la neurobiologie, la psychologie, la philosophie, et bien d'autres domaines de l'informatique et des sciences cognitives.

La polyvalence de l'IA en matière d'innovation provient de son caractère multidisciplinaire. [37]. La distinction entre l'IA forte et l'IA faible réside dans le niveau de capacité cognitive et d'autonomie des systèmes d'intelligence artificielle. L'IA forte se réfère à des systèmes capables de reproduire complètement les capacités intellectuelles humaines [37]. L’IA faible est conçue pour effectuer des tâches spécifiques, souvent de manière très performante, mais sans avoir la capacité de généraliser son intelligence à d'autres domaines [37].

Cette brève définition de l'intelligence artificielle nous offre une introduction aux principes fondamentaux de l'apprentissage automatique.

II.4 Principes fondamentaux de l'apprentissage automatique

L'apprentissage automatique, également connu sous le nom de machine learning ou apprentissage statistique, est une branche de l'intelligence artificielle qui se concentre sur le développement de techniques permettant aux ordinateurs d'apprendre à partir de données et d'algorithmes, imitant ainsi la façon dont les humains apprennent. L'apprentissage automatique permet à un système d'être piloté ou assisté par ordinateur, tel qu'un programme ou un robot, d'adapter ses réponses aux situations rencontrées en se fondant sur l'analyse de données empiriques passées issues de bases de données, de capteurs ou du web [39]. L'objectif de l'apprentissage automatique est de déterminer une fonction f qui établit une relation entre les données d'entrée X

et les résultats Y. La nature de cette fonction dépend de l'algorithme d'apprentissage spécifique employé [40].

II.4.1 Les types d'apprentissage automatique

Les caractéristiques et la quantité des données utilisées pour construire les algorithmes d'apprentissage automatique contribuent à déterminer les différents types dans ce domaine. On peut classifier ces types en trois approches principales : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

II.4.1.1 Apprentissage supervise

L'apprentissage supervisé représente l'une des formes les plus répandues de l'apprentissage automatique. Il s'agit d'une approche où un algorithme est entraîné sur un ensemble de données étiquetées, signifiant que chaque exemple d'entraînement est associé à une étiquette ou à une réponse connue [41]. L'objectif de l'apprentissage supervisé est d'apprendre à partir de ces données étiquetées afin de réaliser des prédictions précises sur de nouvelles données non étiquetées.

L'apprentissage supervisé est utilisé dans de nombreuses applications, telles que la classification d'images, la détection d'objets, la prédiction de prix, la reconnaissance vocale, la traduction automatique, la recommandation de produits, et bien d'autres. Il est essentiel dans le domaine de l'intelligence artificielle pour résoudre une grande variété de problèmes où la prédiction ou la classification est nécessaire.

On peut classifier les problèmes d'apprentissage supervisé en deux catégories principales en fonction du type d'étiquettes associées : la classification et la régression, comme illustré dans la Figure II.1

a. La classification

La classification est une tâche visant à prédire une classe ou une catégorie à partir d'un ensemble de variables. L'objectif est de déterminer une fonction approximative qui associe les variables d'entrée à la variable qualitative de sortie. Cette fonction est ensuite utilisée pour catégoriser de nouvelles données. Les méthodes de classification se divisent généralement en trois types :

- **Classification binaire** : Dans la classification binaire, la valeur à prédire est de nature binaire, ce qui signifie qu'elle peut prendre l'une de deux classes possibles (par exemple, oui/non, positif/négatif, spam/non-spam).

- **Classification multi classe** : concerne des situations où la variable de sortie peut appartenir à l'une de plusieurs classes. Contrairement à la classification binaire, il existe plus de deux catégories possibles. Par exemple, cela peut être appliqué à la classification des espèces d'animaux ou à la catégorisation de documents en plusieurs catégories.
- **Classification multi label** : implique la prédiction de plusieurs étiquettes ou catégories simultanément pour une seule observation. Chaque observation peut être associée à plusieurs classes en même temps. Par exemple : la classification de tags sur une image, où une image peut être associée à plusieurs thèmes ou catégories.

b. La régression

Dans le domaine de la régression, l'objectif est de découvrir le modèle ou la fonction permettant de distinguer les données en valeurs réelles continues ou d'identifier le mouvement de distribution en fonction des données historiques.

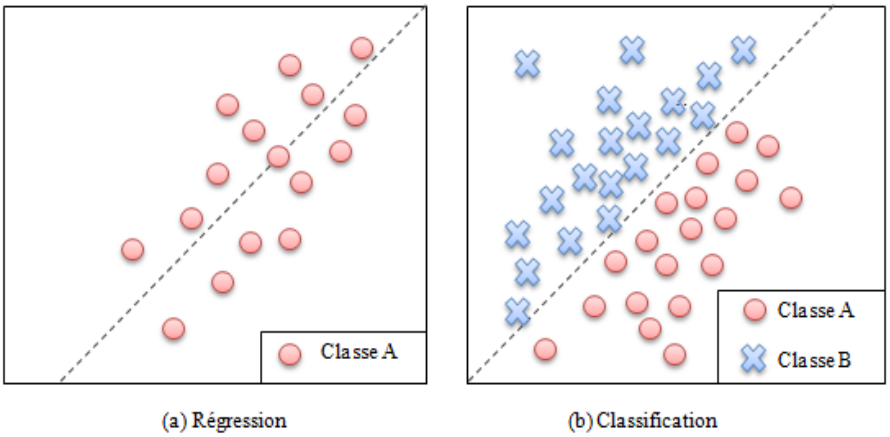


Figure II.1 Illustration de l'apprentissage supervisé [42].

II.4.1.1.1 Les Algorithmes d'apprentissage supervisé

Il existe plusieurs algorithmes dans le domaine de l'apprentissage supervisé, parmi lesquels les plus largement reconnus et utilisés sont :

II.4.1.1.1.1 Régression logistique

La régression logistique (RL), étendue à partir d'une méthode de régression conventionnelle, est une technique de classification largement utilisée [43-45]. RL est conçue pour modéliser un classificateur ou une variable binaire à l'aide d'un seuil discriminant et est appliquée pour différencier deux classes (par exemple, en bonne santé et malade).

II.4.1.1.1.2 Réseaux de neurones artificiels

Un réseau de neurones artificiels (RNA) est un modèle mathématique informatique inspiré de la structure et de la fonction des réseaux neuronaux biologiques du cerveau. Le RNA est composé de neurones artificiels. Ces neurones sont connectés les uns aux autres par des poids synaptiques, souvent simplement appelés poids [46]. Lorsqu'un réseau de neurones artificiels est utilisé pour une tâche d'apprentissage, il commence par ajuster ces poids lors du processus d'apprentissage [47, 48].

A. Analogie entre le neurone biologique et le neurone artificiel

L'apprentissage dans le cerveau humain se fait grâce à un réseau de neurones reliés par des axones, des synapses et des dendrites (voire Figure II. 2). Le neurone biologique est une structure cellulaire constituée d'un corps cellulaire et d'un noyau. Les dendrites partent du corps cellulaire et servent de conduits aux informations provenant de l’environnement externe pour atteindre le soma. Après avoir traité cette information, le neurone la transmet le long de l'axone pour communiquer avec d'autres neurones. Le point de connexion entre deux neurones est appelé synapse. Les dendrites collectent des signaux électrochimiques provenant d'autres cellules ou de l'extérieur via des synapses et les transmettent au corps cellulaire, ou soma.

Le soma, outre son rôle dans le métabolisme cellulaire, accumule et consolide ces informations. Si le potentiel somatique dépasse un seuil spécifique, un potentiel d'action est généré, représentant une brève oscillation électrique. L'axone, une extension cellulaire, est responsable du relais des informations nerveuses vers d'autres cellules via ses ramifications terminales au niveau des synapses.

Le neurone biologique et le neurone artificiel présentent une analogie fascinante dans le domaine de l'intelligence artificielle. Un neurone biologique, l'unité de base du cerveau humain, transmet des signaux électrochimiques via ses axones, synapses et dendrites, établissant ainsi des connexions complexes. De même, le neurone artificiel, qui est l’élément de base d’un réseau neuronal artificiel. Il reçoit un nombre variable d'entrées, et chacune de ces entrées est associée à un poids (en anglais *weight* ($\boldsymbol{w}$)) qui représente l’élément clé de la connexion [46]**.**

Chaque neurone artificiel émet une seule sortie qui se ramifie ensuite pour alimenter un nombre variable de neurones comme illustrées dans la figure II. 3 [47, 49].

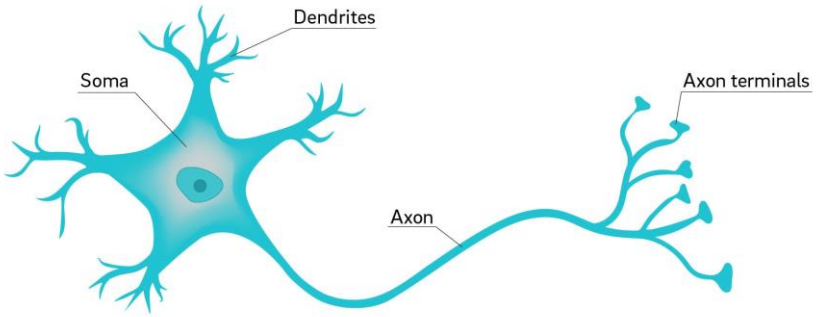


Figure II. 2 Neurone biologique [50]

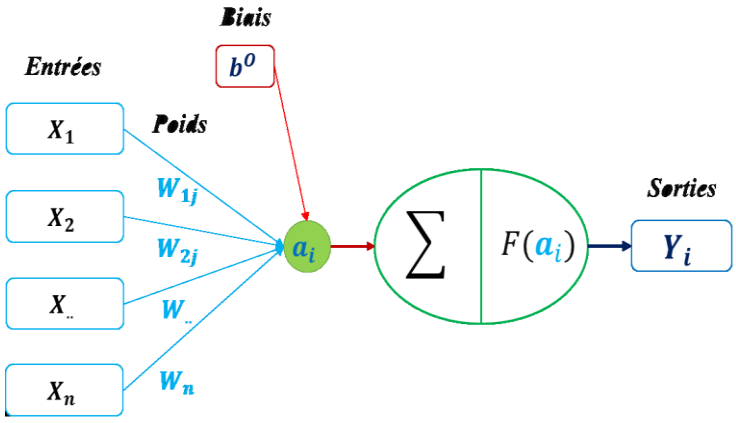


Figure II. 3 Neurone artificiel [46]

Des similitudes dans la conception et les fonctionnalités peuvent être observées dans la Figure II.4.

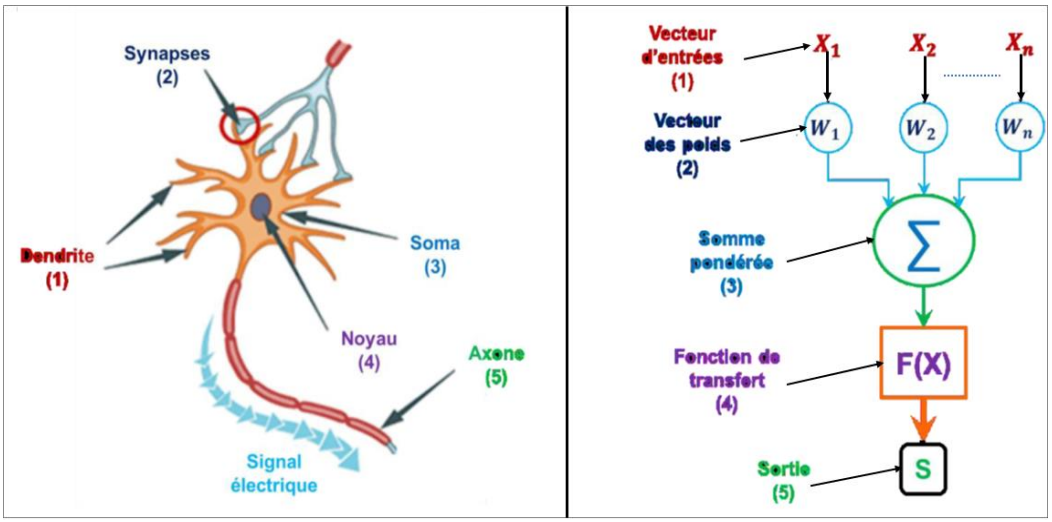


Figure II 4. Comparaison entre un neurone biologique et un neurone artificiel [47, 49,51].

La formulation mathématique de ce neurone artificiel est également clairement détaillée dans **Eq II.1** [52]

$$\boldsymbol{Y_i = F(a_i) = F(b^o + \sum_{i=1}^{N} w_i * X_i)} \qquad \textbf{(II.1)}$$

$\boldsymbol{b_o}$ Est le poids effectué à une entrée constante (biais), $\boldsymbol{w_i}$ est le poids de connexion pour la $\boldsymbol{i^{ème}}$ entrée, et $\boldsymbol{X_i}$ et $\boldsymbol{Y_i}$ représentent respectivement les valeurs de la $\boldsymbol{i^{ème}}$ entrée et sortie. La fonction $\boldsymbol{F}$, également appelée fonction d'activation ou fonction de transfert, est appliquée à l'unité $\boldsymbol{ai}$, qui est le potentiel du neurone. Cette unité effectue le calcul de la somme pondérée de toutes les entrées.

Comme illustré par le modèle de neurone artificiel et son équation (II.1), la fonction d'activation constitue la principale variable inconnue de notre modèle. Cette fonction est un élément constitutif important pour les réseaux de neurone artificiels et pour n'importe quel modèle mathématique, car elle définit leurs propriétés essentielles. Notre choix de la fonction d'activation découle des problèmes spécifiques que le neurone artificiel doit résoudre, comme le suggère la plupart des cas [48]. La plupart des fonctions d'activation dans les réseaux de neurones artificiels sont continués, offrant ainsi une gamme infinie de valeurs comprises dans les intervalles [0, +1] ou [−1, +1] [53]. Il existe diverses formes de fonctions d'activation. Le tableau II.1 donne quelques exemples de fonctions d'activation utilisées pour les réseaux de neurones artificiels dans le logiciel Statistica [54-57].

Tableau. II.1. Fonctions d'activation usuelles existantes dans le logiciel Statistica [54-57]

Nom du Fonction	**Définition mathématique**	**Intervalle**	**Description**	**Désignation Correspondante**
Linéaire	$a = n$	(−∞, +∞)	L'activation du neurone est transmise directement en tant que sortie	***Identity***
Logarithmique Sigmoïde (Logistique)	$a = \frac{1}{1 + e^{-n}}$	(0, +1)	Une courbe en forme de S	***Logistic***
Tangente hyperbolique	$a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$	(−1, +1)	Une courbe sigmoïde (tangente hyperbolique) similaire à la fonction logistique. Fonctionne souvent mieux que la fonction logistique en raison de sa symétrie. Idéal pour les perceptrons multicouches, en particulier les couches cachées	***Tanh***
Exponentielle	$a = e^{-n}$	(0, +∞)	La fonction exponentielle négative	***Exponential***
Sinus	$a = \sin(n)$	(0, +∞)	Peut-être utile si la reconnaissance des données distribuées radialement	***Sine***
Exponentielle Normalisée	$a = \frac{e^{n_i}}{\sum e^{n_i}}$	[0, +1]	Principalement utilisé pour (mais sans s'y limiter) les tâches de classification. Utile pour construire des réseaux de neurones avec des sorties multiples normalisées, ce qui le rend particulièrement adapté à la Création de classificateurs de réseaux de neurones avec des sorties probabilistes	***Softmax***
Gaussienne	$a = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$	/	Ce type de fonction d'activation gaussienne isotrope est uniquement utilisé par les unités cachées d'un réseau neuronal RBF, également appelées fonctions de base radiales. L'emplacement (également connu sous le nom de vecteurs prototypes) et les paramètres de propagation sont équivalents aux poids de couche cachés en entrée d'un réseau neuronal MLP	***Gaussian***

*Avec : $\boldsymbol{a = f(n)}$, **(n)** C'est l'entrée nette d'un neurone (Somme pondérée des entrées du neurone pour les MLP). Pour la fonction gaussienne : **(x)** Représente le vecteur des entrées de neurones, mu $(\boldsymbol{\mu})$ Vecteur des poids d'entrée (centre du RBF), sigma $(\boldsymbol{\sigma})$ Propagation (largeur) du RBF.*

B. Algorithme d’apprentissage

L'apprentissage dans le domaine des réseaux de neurones artificiels se réfère à la manière dont l'architecture et les caractéristiques principales d'un réseau sont définies. L'une de ces propriétés fondamentales est la capacité du réseau à s'adapter et à améliorer sa performance en ajustant les connexions neuronales par le biais de la procédure d'apprentissage, qui est réalisée à l'aide d'algorithmes d'apprentissage [58, 59].

La plupart des algorithmes actuels impliquent la modification des poids de connexion pour que la réponse du réseau corresponde avec les exemples de l'expérience. Après une initialisation aléatoire des poids, les algorithmes d'apprentissage ajustent progressivement ces poids afin de réduire l'écart entre les sorties calculées par le réseau et les sorties expérimentales [53, 58, 59].

Divers algorithmes itératifs de rétropropagation peuvent être mis en œuvre, parmi lesquels on trouve les algorithmes de Descente du gradient, de Levenberg–Marquardt, de Régularisation Bayésienne, de BFGS, etc [53]. Toutefois, dans cette section, je me concentre spécifiquement sur l'algorithme de BFGS, qui sera utilisé comme algorithme d'apprentissage dans cette étude.

L'acronyme BFGS provient de l'association des noms de ses fondateurs : Broyden, Fletcher,

Goldfarb et Shanno [60, 61]. Cet algorithme appartient à la catégorie des méthodes d'optimisation dites "Quasi-Newtoniennes", ces méthodes sont une généralisation de la méthode de Newton [58, 61-65].

C. Réseaux de neurones perceptron multicouche

Le Perceptron Multi-Couches (PMC), appartient à la famille des réseaux de neurones non-bouclés, également appelés Feed-forward Networks. Habituellement, à la suite d'un apprentissage supervisé, ces neurones sont arrangés en couches successives, où les informations circulent dans un seul sens, de la couche d'entrée vers la couche de sortie. Chaque neurone dans une couche est connecté à tous les neurones de la couche précédente et de la couche suivante. Cependant, les neurones au sein d'une même couche ne sont pas interconnectés [51, 52, 66]. Le réseau de ce type est constitué de différentes couches:

La couche d'entrée d'un réseau neuronal renferme des données inhérentes à son point d'origine. Ces données présentent une diversité de qualités, qu'elles soient de nature structurelle, qualitative, quantitative, ou associées aux résultats souhaités.

Les couches cachées, qu'elles soient une ou multiples, englobent plusieurs neurones interconnectés aux données de la couche d'entrée en relation avec les résultats attendus de la couche de sortie.

La couche de sortie, composée d'un ou de plusieurs neurones en fonction du nombre de réponses visées, génère les valeurs de sortie résultant de la combinaison linéaire des fonctions de transfert des couches cachées. La figure II.5 présente le schéma de l'architecture d'un réseau neuronal perceptron multicouches.

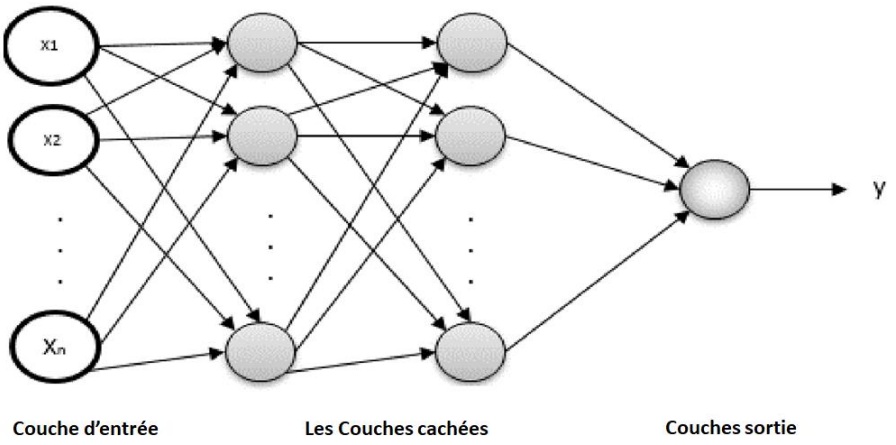


Figure II.5 Architecture d'un réseau neuronal perceptron multicouches

II.4.1.1.1.3 Algorithme de k-plus proches voisins

Le classificateur de K-Plus Proches Voisins (en anglais : K-Nearest Neighbors (K-NN)) est une méthode classique largement utilisée pour résoudre les problèmes de classification et de la régression [67]. K-NN est le plus simple de tous les algorithmes d'apprentissage automatique. L'idée clé de la méthode du plus proche voisin est assez simple : les instances sont classés en fonction de la classe de leurs voisins les plus proches. Il est souvent utile de prendre plus d'un excellent voisin (disons k), d'où le terme K-NN. Cette approche est également appelée classification basée sur des exemples ou classification basée sur des cas, car la classification dépend directement des exemples de formation. En résumé, le classificateur K-NN comporte deux étapes : la première est la sélection des excellents voisins, et la seconde est la détermination de la classe sur la base des distances de ces voisins [40]. (Voir Figure II.6)

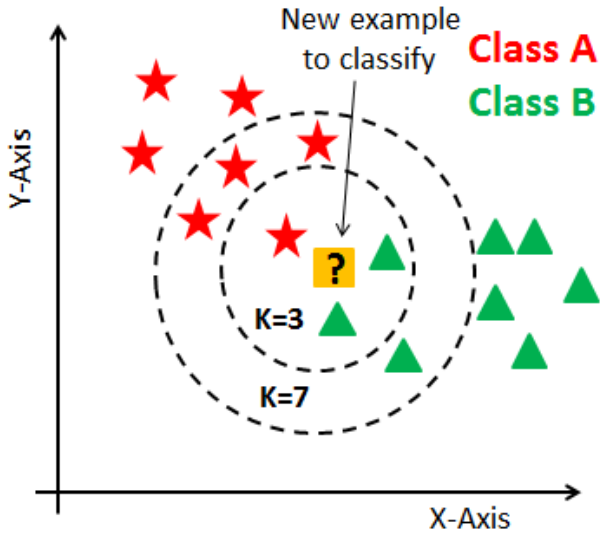


Figure II.6 Algorithme des k-plus proches voisins [5].

II.4.1.1.1.4 Les arbres de décision

Les arbres de décision [68] (en anglais : Decision Trees (DT)) sont parmi les méthodes de classification les plus connues et les plus utilisées. Un arbre de décision est un modèle d'apprentissage automatique qui prend la forme d'une structure arborescente. Il est utilisé pour représenter des décisions et des conditions basées sur des caractéristiques des données. Chaque nœud de l'arbre représente une caractéristique (ou attribut), chaque branche représente une décision basée sur cette caractéristique, et chaque feuille représente le résultat de la décision [69].

L'arbre de décision est construit à partir d'un ensemble de données d'entraînement, où chaque exemple est associé à une étiquette de classe. L'objectif est de diviser récursivement les données en sous-ensembles homogènes en termes de classe, en choisissant les caractéristiques aux nœuds de l'arbre. Cela permet de créer une structure hiérarchique qui peut être utilisée pour prendre des décisions ou faire des prédictions sur de nouvelles données.

II.4.1.1.1.5 Les forêts aléatoires

Les Forêts aléatoires (en anglais : Random Forest (RF)) sont des méthodes d'apprentissage automatique impliquent la construction de plusieurs Arbres de Décision (DT) de manière indépendante [70]. Chaque arbre est construit à partir d'un échantillon Bootstrap de l'ensemble de données d'origine, garantissant ainsi la diversité des arbres. Lorsqu'une nouvelle entrée est présentée, chaque arbre prédit une classe, et la classe finale est déterminée par un vote majoritaire des arbres constitutifs [71, 72]. L'agrégation des prédictions de plusieurs arbres atténue le bruit, rendant le modèle moins sensible aux valeurs aberrantes et plus robuste face aux petites quantités de données [70, 73]. Les Forêts aléatoires intègrent un mécanisme de sélection de fonctionnalités,

permettant de gérer de nombreux paramètres d'entrée sans nécessiter la suppression de certaines variables pour réduire la dimensionnalité [74].

II.4.1.1.1.6 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support (en anglais Support vector machines (SVM)) sont des techniques puissantes de classification et de régression dans le domaine de l'apprentissage automatique [30, 75, 76]. Elles sont particulièrement efficaces pour la classification

- **Principe fondamental** : Le principe de base des SVM repose sur la recherche d'une séparation linéaire optimale entre les différentes classes d'objets dans un espace multidimensionnel.
- **Séparation par un hyperplan** : L'objectif est de trouver un hyperplan (une sorte de plan dans un espace à plusieurs dimensions) qui maximise la marge entre les classes [77]. La marge est la distance entre l'hyperplan et les exemples d'entraînement les plus proches de chaque classe, appelés vecteurs de support.
- **Maximisation de la marge (Optimisation)** : L'optimisation vise à maximiser la marge. C'est un problème d'optimisation convexe qui peut être résolu efficacement.
- **Fonction de décision (Fonction discriminante)** : La fonction de décision d'une SVM est déterminée par l'hyperplan trouvé. Elle attribue une classe à chaque observation en fonction de laquelle de part et d'autre de l'hyperplan elle se trouve.
- **Noyau (Kernel) :** Les SVM peuvent être étendues pour traiter des problèmes de classification non linéaires en utilisant des noyaux [77]. Ces noyaux permettent de transformer les données dans un espace de caractéristiques de dimension supérieure, où une séparation linéaire est possible. Les noyaux couramment utilisés incluent le noyau linéaire, le noyau polynomial et le noyau radial (RBF). Le choix du noyau dépend de la nature des données.
- **Paramètres de pénalité (C)** : Un paramètre important dans SVM est le paramètre de pénalité, qui détermine la tolérance aux erreurs de classification. Un C plus élevé signifie une tolérance plus faible aux erreurs, ce qui conduit à un ajustement plus strict.

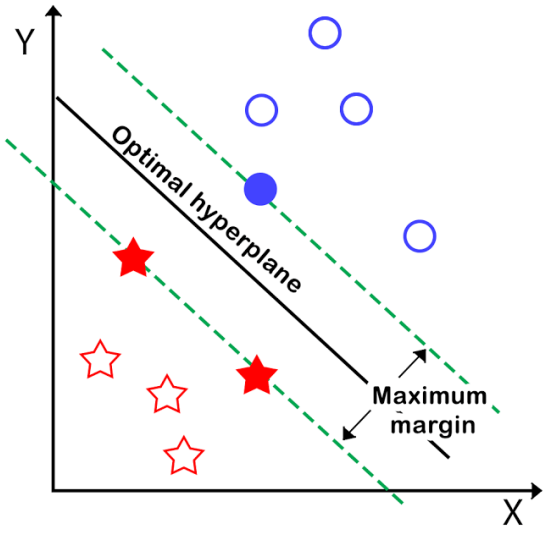


Figure II.7 Algorithme de SVM [5]

II.4.1.2 L'apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'apprentissage automatique où le modèle est exposé à des données non étiquetées, contrairement à l'apprentissage supervisé qui utilise des données étiquetées pour entraîner le modèle à prédire des étiquettes pour de nouvelles données. L'objectif de l'apprentissage non supervisé est d'extraire des structures ou des modèles intrinsèques à partir des données, sans disposer de résultats attendus préalablement définis. Cette approche s'adapte particulièrement bien aux problèmes impliquant de vastes ensembles de données non étiquetées [42]. Dans des contextes où l'annotation de données serait coûteuse ou difficile à obtenir, l'apprentissage non supervisé offre une méthode efficace pour explorer et analyser ces données de manière exploratoire.

Il existe plusieurs techniques d'apprentissage non supervisé, dont voici quelques-unes :

- **Regroupement (Clustering)** : Cette technique consiste à regrouper les données en fonction de similarités intrinsèques [42]. L'idée est de regrouper des observations similaires dans un même cluster tout en maintenant une certaine dissimilarité entre les clusters (voir la Figure II.8).
- **Réduction de la dimensionnalité** : L'objectif ici est de réduire le nombre de variables ou de caractéristiques tout en préservant au mieux l'information contenue dans les données.
- **Association** : est une technique d'apprentissage non supervisé utilisée pour découvrir des relations intéressantes entre les variables dans un ensemble de données. L'objectif principal

est d'identifier des règles d'association indiquant quels éléments apparaissent fréquemment ensemble.

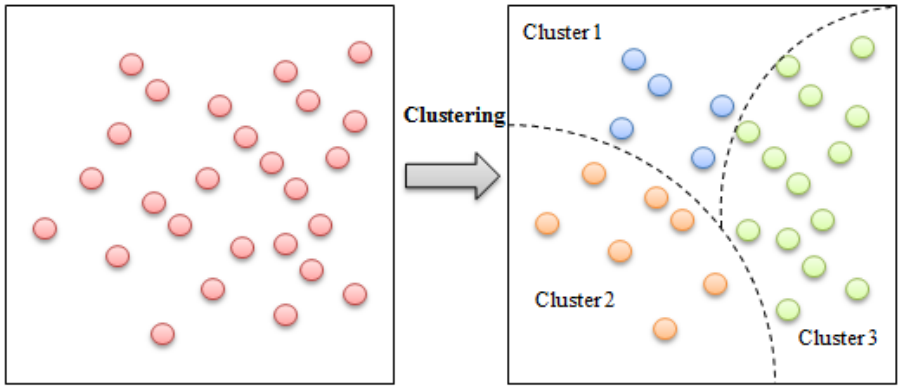


Figure II.8 Apprentissage non supervisé [42]

II.4.1.3 Apprentissage par renforcement

L'apprentissage par renforcement est une branche de l'apprentissage machine où un agent interagit avec un environnement et apprend à prendre des décisions afin de maximiser une récompense cumulative au fil du temps. Contrairement à l'apprentissage supervisé, où le modèle est entraîné sur un ensemble de données étiqueté, et à l'apprentissage non supervisé, qui se concentre sur la découverte de structures dans les données, l'apprentissage par renforcement repose sur l'idée d'apprendre par essais et erreurs dans un environnement dynamique.

II.5 Apprentissage en profondeur

L'apprentissage profond (en anglais : Deep learning (DL), en tant que sous-domaine populaire et en évolution de l'IA, a considérablement progressé et a été appliqué pour résoudre divers problèmes complexes. DL a réalisé de nombreuses percées dans des domaines tels que la reconnaissance, la classification et la segmentation d'images [78, 81]. DL imite le traitement du cerveau humain et est généralement composé de réseaux neuronaux artificiels multicouches. La large application des techniques DL à forte intensité de calcul a été possible grâce aux progrès techniques des unités de traitement graphique, conduisant aux capacités révolutionnaires du DL [82]. Deux algorithmes de réseau neuronal populaires basés sur DL sont le réseau neuronal convolutif et le réseau neuronal récurrent.

II.5.1 Réseau neuronal convolutif

Un réseau neuronal convolutif, (en anglais : Convolutional Neural Network (CNN)), est un type spécifique de réseau neuronal profond conçu principalement pour le traitement des données structurées en réseau, telles que des images. Les CNN ont été particulièrement efficaces dans des

tâches liées à la vision par ordinateur, comme la reconnaissance d'objets, la classification d'images, la détection d'objets, et d'autres applications similaires.

CNN est composé de multicouches artificielles comprenant un processus convolutif qui extrait les caractéristiques hiérarchiques de l'entrée brute et est équipé de couches entièrement connectées en tant que classificateurs [83, 84].

II.5.2 Réseau neuronal récurrent

Un réseau neuronal récurrent (en anglais : Recurrent Neural Network (RNN)) est un type de réseau neuronal qui est particulièrement adapté au traitement de données séquentielles, où l'ordre des éléments a une importance. Les RNN possèdent une mémoire interne qui permet de conserver des informations sur les entrées précédentes pendant un certain laps de temps.

Les RNN sont récurrents car ils présentent une structure de rétroaction, ce qui signifie que l'information peut être transmise d'une étape à la suivante dans la séquence.

II.6 Algorithmes d'optimisation

Les algorithmes d'optimisation sont des procédures ou ensembles d'instructions algorithmiques conçus pour trouver la meilleure solution possible parmi un ensemble de solutions potentielles. Leur objectif est de minimiser ou maximiser une fonction objective tout en respectant éventuellement des contraintes spécifiques. Ces algorithmes sont utilisés dans divers domaines, tels que l'intelligence artificielle, l'apprentissage automatique, l'ingénierie, la planification, la logistique et d'autres disciplines, pour résoudre des problèmes complexes et prendre des décisions optimales.

Il existe plusieurs algorithmes d'optimisation dans le domaine de l'intelligence artificielle, parmi lesquels les plus largement reconnus et utilisés sont :

- **Colonie d'abeilles artificielles** (en anglais : Artificial bee colony (ABC))

L'algorithme ABC est un algorithme d'optimisation inspiré du comportement des abeilles, utilisé dans de nombreux problèmes d'optimisation [85]. Karaboga l'a proposé en 2005 [86]. Les abeilles individuelles sont classées en trois sortes dans l'algorithme (employées, observatrices et éclaireuses), chaque agent ayant une fonction particulière à remplir dans le processus [85, 87].

- **Algorithme de recherche de corbeaux** (en anglais : Crow search algorithm (CSA))

L'algorithme CSA proposé par Askarzadeh en 2016 est un algorithme méta-heuristique [88]. On dit que l'oiseau le plus intelligent est le corbeau [88]. Les corbeaux peuvent reconnaître des

visages et communiquer entre eux pour avertir lorsqu’un inconnu est dans les parages [88]. De plus, ils peuvent communiquer de manière développée, et se souvenir de l'emplacement de leur nourriture plusieurs mois plus tard [88, 89]. Les corbeaux regardent où d'autres oiseaux cachent leur nourriture et la volent [89]. Ils utilisent leur propre expérience du vol pour prédire le comportement des voleurs, afin d'éviter de devenir une victime à l'avenir, et ils peuvent déterminer le chemin sûr pour protéger leurs caches [89]. La principale source de cet algorithme était la façon dont les corbeaux cachent leur nourriture [89, 90].

Les définitions suivantes décrivent les quatre principes majeurs de la CSA [89] :

– Les corbeaux existent sous la forme d’un groupe.

– Les corbeaux se souviennent où cacher leur nourriture.

– Lorsque le vol est commis, chacun des membres traque les autres.

– Les corbeaux gardent leurs caches en étant extrêmement vigilants contre le vol.

- **Algorithme de libellule** (en anglais Dragonfly Algorithm (DA))

C’est un algorithme d’optimisation inspiré des comportements statiques et dynamiques des libellules [91]. Le processus d'optimisation repose sur deux étapes majeures : l'exploration et l'exploitation réalisées en attirant des sources de nourriture et en repoussant les ennemis [92, 93].

- **Optimisation des essaims de particules (en anglais Particle Swarm Optimization (PSO))**

Inspirée par le comportement d’essaims tels que les essaims d’oiseaux et les colonies de fourmis, l’optimisation des essaims de particules est une approche répandue que Kennedy et Eberhart ont d’abord développée. Il résout des problèmes d'optimisation non linéaire complexes avec les avantages d'avoir moins de paramètres à ajuster [94]. Il consiste à trouver la solution en comparant et en actualisant les informations de position et de vitesse d’un individu [95]. PSO est couplé à des algorithmes d’apprentissage automatique pour déterminer leurs paramètres optimaux.

II.7 Applications de l'intelligence artificielle en santé

L’intelligence artificielle révolutionne de nombreux domaines, dont celui de la santé. Grâce à ses capacités d'analyse rapide, de traitement des données massives et d'apprentissage automatique, l'IA offre des opportunités uniques pour améliorer les soins de santé, la recherche médicale et la gestion des données.

II.7.1 Diagnostic médical amélioré

L'IA a connu un développement significatif dans le domaine de l'aide au diagnostic en santé [7]. L'IA peut être utilisée pour analyser des données médicales complexes, telles que des images radiologiques, des scans et des dossiers médicaux électroniques, afin de fournir des diagnostics plus rapides et plus précis. Les algorithmes d'apprentissage automatique peuvent apprendre à détecter des motifs subtils qui pourraient échapper à l'œil humain, contribuant ainsi à une meilleure identification des maladies telles que le cancer, les maladies cardiaques et les troubles neurologiques.

En dermatologie, des algorithmes conçus dans le but de diagnostiquer le mélanome démontrent une efficacité notable [96]. Des chercheurs du Stanford Cancer Institute ont développé un outil d'intelligence artificielle aussi performant que des dermatologues expérimentés pour diagnostiquer les cancers de la peau [96, 97]

En l'anatomopathologie, l'utilisation de techniques d'intelligence artificielle pour analyser des images histologiques a le potentiel d'améliorer la précision des diagnostics, d'affiner les pronostics, et de déterminer le profil génomique de certains types de cancers [98].

En cardiologie, des avancements sont réalisés dans le développement d'outils d'intelligence artificielle dédiés à l'interprétation automatique des électrocardiogrammes [7]. Une IA fondée sur les réseaux de neurones, conçue pour soutenir les professionnels de la santé dans le diagnostic des troubles du rythme cardiaque, a obtenu l'approbation de la Food and Drug Administration (FDA) aux États-Unis [7].

En psychiatrie, les chercheurs Marcel Just et David Brent, affiliés respectivement à l'université Carnegie Mellon et à l'université de Pittsburgh aux États-Unis, ont mis au point un outil d'intelligence artificielle qui utilise un couplage entre un algorithme et l'imagerie cérébrale pour identifier les intentions suicidaires chez les patients. Lors de leur première expérimentation sur un échantillon limité d'individus, cet outil a démontré un taux d’exactitude de 91% [99].

II.7.2 Prédiction des maladies

L'IA peut être utilisée pour analyser les données des patients et identifier les facteurs de risque, permettant ainsi de prédire le développement potentiel de maladies, ce qui facilite la prévention et la gestion précoce.

Des chercheurs de l'institut Gustave Roussy ont mis au point une technologie d'intelligence artificielle capable d'analyser des images de scanner afin de prédire la réponse d'un patient face à son traitement anticancéreux de type immunothérapie [101].

Analyse des images pathologiques : Ces dernières années, l'intelligence artificielle s'est progressivement introduite dans le domaine de la radiologie [102]. Ce qui est utile dans le diagnostic des cancers et d'autres maladies.

Les algorithmes d'apprentissage automatique sont désormais capables d'accomplir des tâches d'analyse d'images avec des performances équivalentes, voire supérieures, à celles réalisées par des professionnels humains [102] ce qui est utile dans le diagnostic des cancers et d'autres maladies.

Des chercheurs de l'université de Stanford ont créé un logiciel d'imagerie fondé sur un réseau de neurones, potentiellement plus efficace que les radiologues pour diagnostiquer la pneumonie. Le logiciel est identifié sous le nom de (CheXnet) [103].

II.7.3 Recherche médicale

L'intelligence artificielle, notamment le "Deep Learning" et d'autres technologies, suscite des attentes prometteuses dans la recherche de nouveaux médicaments [7]. Certains experts prévoient qu'elle facilitera la découverte de traitements novateurs de manière plus efficiente, rapide et économique [7].

Cependant, certains chercheurs ont qualifié ces attentes d’exagérées. Il n’existe actuellement aucun médicament approuvé par l’IA. Les grandes entreprises pharmaceutiques utilisent de systèmes d'intelligence artificielle dans leurs activités de recherche.

Pfizer utilise IBM Watson, un système basé sur l'apprentissage automatique, pour renforcer ses efforts de recherche dans le domaine des médicaments immuno-oncologiques [7]. Sanofi a conclu un accord pour utiliser la plateforme d'intelligence artificielle de la start-up britannique Exscientia, se concentrant sur la recherche de traitements pour les maladies métaboliques [7].

II.7.4 Télémédecine

Le terme télémédecine a vu le jour dans les années 1970 et fait référence à l'utilisation des technologies de l'information et des communications pour améliorer l'accès à l'information sur les soins de santé, contribuant ainsi à l'amélioration de la santé et du bien-être de la communauté [104]. Il permet aux cliniciens de fournir des services médicaux sans contact direct avec les patients, notamment dans les situations nécessitant une distance sociale [101, 104]. Les appels vocaux et vidéo sont les technologies les plus fréquemment utilisées [105].

La médecine numérique et l'intelligence artificielle (IA) ont le potentiel de révolutionner le domaine du diabète en permettant une surveillance continue à distancer des symptômes des

patients, de leurs données physiologiques et de leurs comportements. Cette transformation s'opère grâce à l'utilisation de dispositifs portables, de capteurs et de technologies de smartphone, contribuant ainsi à alléger le fardeau associé à cette surveillance [106].

Bien que ces applications présentent de nombreux avantages, il est important de relever des défis tels que la confidentialité des données, la transparence des algorithmes, et l'acceptation et la confiance des professionnels de la santé et des patients. L'intégration réussie de l'IA en santé nécessite une approche éthique et réglementaire.

II.8 Conclusion

Dans le présent chapitre, nous avons d'abord introduit l'IA et son importance. Ensuite, nous avons expliqué l'apprentissage automatique ainsi que ses différents types. Enfin, nous avons abordé les applications de l'intelligence artificielle dans le domaine de la santé.

Chapitre III

Revue Bibliographique sur L'utilisation de L'intelligence Artificielle dans la Prédiction Médicale

III Revue bibliographique sur l'utilisation de l'intelligence artificielle dans la prédiction médicale

III.1 Introduction

Une transformation majeure s'opère actuellement dans le domaine médical, visant à intégrer la virtualisation dans l'ensemble des pratiques médicales. Cette évolution résulte de la convergence entre la biologie, la médecine, les mathématiques et l'informatique, exploitant notamment la capacité de cette dernière à analyser des volumes considérables de données, communément appelées "Big Data".

Dans ce chapitre, nous allons décrire les travaux réalisés dans la littérature utilisant l'IA pour la prédiction médicale et enfin, conclure.

III.2 Les revues de littérature portant sur l'application de l'intelligence artificielle dans la prédiction médicale.

La prédiction des maladies grâce à l'utilisation de IA a suscité une attention considérable ces dernières années, incitant à des recherches approfondies et à l'exploration de ses capacités dans le domaine médical. L'étude du pouvoir prédictif de l'IA dans le diagnostic et le pronostic des maladies est devenue une priorité.

Dans cette section, nous visons à fournir un aperçu complet des études menées entre 1997 et 2024, comme illustré dans la figure III.1, en approfondissant les méthodologies utilisées pour la modélisation prédictive, l'intégration de divers ensembles de données et l'évolution des algorithmes d'apprentissage automatique. L'exploration englobe les progrès réalisés dans la compréhension de la manière dont l'IA peut contribuer à la détection précoce des maladies, à l'évaluation des risques et à la médecine personnalisée. Grâce à un examen des données expérimentales, au développement de modèles prédictifs et aux approches innovantes exploitant l'intelligence artificielle, ce chapitre cherche à élucider les progrès et les défis liés à l'exploitation de l'IA pour la prédiction des maladies dans le domaine de la recherche biomédicale.

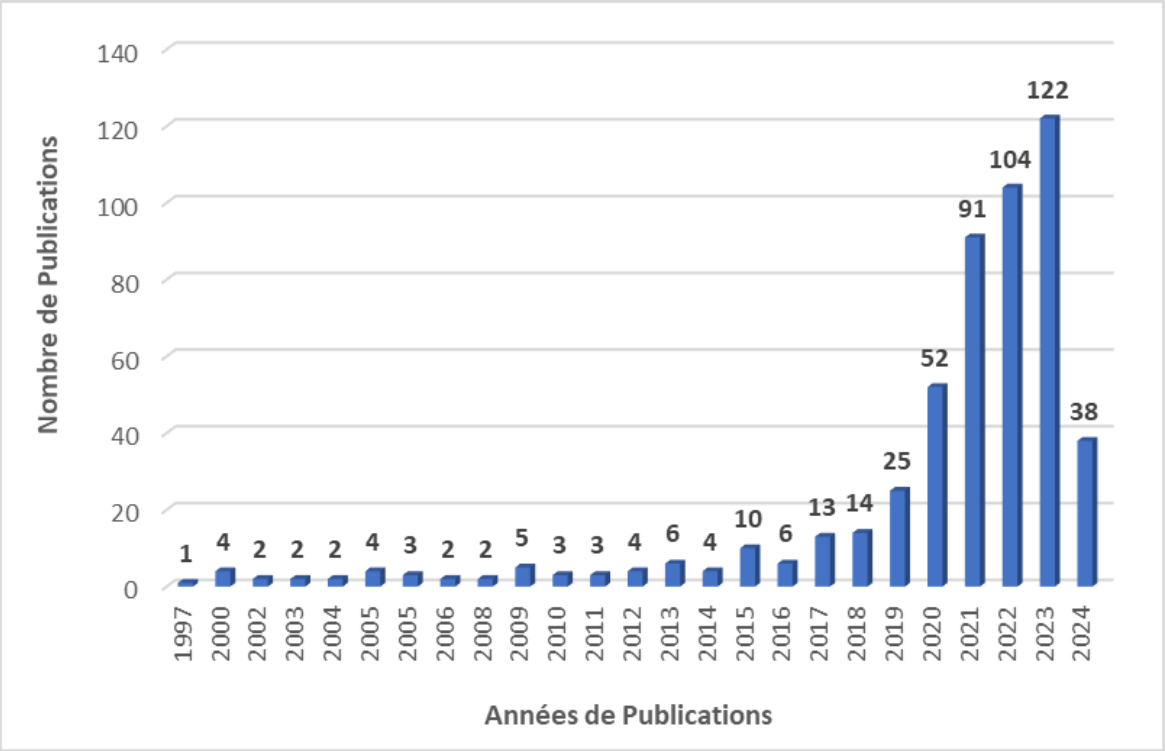


Figure III.1 Nombre de publications sur la prédiction des maladies par l'intelligence artificielle (Science direct 14 / 2 /2024)

III.3 Revue critique de la littérature sur la Prédiction des affection transmissible et non transmissible par l'Intelligence Artificielle

En 2019, Singh et Gupta [107] ont développé une méthodologie visant à détecter et classifier les images tomodensitométriques associées au cancer du poumon en deux catégories distinctes : bénignes et malignes. Leur approche a débuté par l'application de techniques de traitement d'images pour prétraiter les données visuelles. Par la suite, ils ont utilisé des algorithmes d'apprentissage supervisé pour la classification. Les caractéristiques extraites comprenaient des informations texturales ainsi que des caractéristiques statistiques, qui étaient ensuite fournies à plusieurs classificateurs. Ils ont exploité un ensemble de données comprenant 15 750 images cliniques, avec 6 910 images bénignes et 8 840 images de cancer du poumon malin. Pour la classification, ils ont employé sept classificateurs distincts : le classificateur k-voisin le plus proche, un classificateur de machine vectorielle, le classificateur d'arbre de décision, le classificateur bayésien naïf multinomial, le classificateur de descente de gradient aléatoire, le classificateur de forêt aléatoire et le classificateur perceptron multicouche. Les résultats obtenus indiquent que le PMC a démontré la meilleure précision, atteignant un taux impressionnant de

88,55 %. Ce résultat surpasse significativement les performances des autres classificateurs évalués dans cette étude.

L'objectif de la recherche menée par Faisal et al. [108] vise à évaluer le pouvoir discriminant de plusieurs prédicteurs, notamment un SVM, un arbre de décision C4.5, un PMC, un réseau neuronal et NB, et un arbre boosté par gradient pour augmenter l'efficacité de la détection du cancer du poumon par son symptômes. Les chercheurs ont extrait des ensembles de données du référentiel de l'Université de Californie à Irvine (UCI) et effectué un prétraitement pour nettoyer les données. Par la suite, ils ont formé et testé les classificateurs sélectionnés à l'aide d'une validation croisée standard de 10 fois. Les résultats pour chaque classificateur individuel ont été calculés et les trois meilleurs classificateurs, sur la base de mesures de performance telles que la précision, ont été choisis pour former un ensemble en utilisant le vote majoritaire. Cet ensemble, combiné à la forêt aléatoire, a été utilisé pour calculer les résultats, permettant une comparaison complète des performances des classificateurs individuels et d'ensemble. Notamment, l'arbre boosté par gradient a atteint une Exactitude de 90 %, surpassant tous les autres classificateurs individuels et d'ensemble en termes de performances.

Wu et al. [109] ont introduit un modèle RF conçu pour l'identification du cancer du poumon à l'aide d'indices de cancer de routine. La validation du modèle a été réalisée selon une méthode de validation croisée 10 fois. L'ensemble de données utilisé dans cette étude a été collecté auprès de l'Université de Lanzhou et comprenait 277 patients. Les résultats ont démontré une précision de 95,7 %, un rappel de 96,3 % et une aire sous la courbe AUC-ROC de 99,01 %.

En 2022, Mamun et al. [29] ont réalisé une revue approfondie des études antérieures consacrées aux modèles de prédiction du cancer du poumon. Ils ont introduit des techniques d'apprentissage en ensemble, notamment XGBoost, LightGBM, Bagging et AdaBoost, pour prédire le cancer du poumon à partir d'un ensemble de données comprenant 309 participants. Les attributs utilisés dans ces modèles comprenaient des éléments tels que l'âge, le tabagisme, les doigts jaunes, l'anxiété, la pression artérielle, les maladies chroniques, la fatigue, les allergies, la respiration sifflante, la consommation d'alcool, la toux, l'essoufflement et la déglutition. Pour équilibrer l'ensemble de données en vue de l'analyse, la méthode de suréchantillonnage Stands for Synthetic Minority Over-Sampling Technique (SMOTE) a été appliquée. Les modèles développés ont été formés et évalués sur l'ensemble de données en utilisant une approche standard de validation croisée 10 fois. XGBoost a dépassé les autres techniques, affichant une exactitude de 94,42%, une précision de 95,66%, un rappel de 94,46%, et une AUC-ROC de 98,14%.

Dritsas et Trigka [110] ont utilisé l'apprentissage supervisé pour créer des modèles de détection du cancer du poumon chez les individus en fonction de divers symptômes. Plusieurs modèles, dont NB, réseau bayésien (BayesNet), stochastic gradient descent (SGD), SVM, RL, RNA, KNN, arbre de modèle logistique, arbre aléatoire, forêt de rotation (RotF), arbre d'élagage à erreur réduite (RepTree), FA et AdaBoost M1, ont été évalués pour leur exactitude, leur précision, leur rappel, leur

F -mesure et leur AUC-ROC. Après avoir appliqué SMOTE avec une validation croisée 10 fois supérieure, RotF est apparu comme le modèle le plus performant, atteignant une exactitude, une précision, un rappel et une mesure F de 97,1 % et une AUC-ROC de 99,3 % dans l'identification du cancer du poumon.

Maulidina F et al. [111] ont présenté une approche de l'apprentissage automatique pour le diagnostic du cancer du poumon en utilisant la méthode Particle Swarm Optimization-Genetic Algorithm-Support Vector Machine (PSO-GA-SVM). Cette recherche s'est basée sur un ensemble de données de tomodensitométrie (CT Scan) provenant de l'hôpital Cipto Mangunkusumo à Jakarta, en Indonésie. L'algorithme génétique d'optimisation par essaim (PSO-GA) a été appliqué pour optimiser les paramètres du SVM, et les performances de la méthode ont été évaluées en termes d'exactitude, de précision, de rappel et de score F1. Comparativement au SVM sans optimisation des paramètres, le PSO-GA-SVM a atteint une exactitude de 97,69%.

En 2018, Deepti Sisodia et al. [112] ont utilisé trois algorithmes de classification, à savoir DT, SVM et Naive Bayes, pour prédire la probabilité de développer un diabète. Des expériences ont été menées sur la base de données Pima Indians Diabetes (PID) du référentiel UCI. Les performances des trois algorithmes ont été évaluées en termes d'exactitude, de précision, de

F -mesure et de rappel. Naive Bayes surpasse les autres classificateurs avec une précision élevée de 76,30 %. Ces résultats sont confirmés par l'utilisation des courbes ROC

En 2018, Chen et al. [113] ont proposé un modèle prédictif hybride pour aider au diagnostic du diabète de type 2. Ils ont utilisé des K-moyennes (en anglais K-means) pour la réduction des données et ont utilisé l'arbre de décision J48 comme classificateur pour la classification. L'ensemble de données sur le diabète des Indiens Pima du référentiel UCI Machine Learning a été utilisé dans leur étude. Le modèle a atteint une Exactitude de classification de 90,04 %.

En 2018, Olanrewaju R et al. [114] ont proposé un système de prédiction des maladies cardiovasculaires (MCV) utilisant les réseaux neuronaux multicouches à action anticipée et la

rétropropagation. L'ensemble de données MCV du référentiel d'apprentissage automatique UCI a été utilisé. Les performances ont été évaluées en fonction de l'exactitude, de la sensibilité, de la spécificité, de la valeur prédictive positive et de la valeur prédictive négative. Le modèle a été comparé aux algorithmes KNN et NB. Les résultats ont prouvé que les réseaux de neurones atteignaient d'excellentes performances avec une exactitude de 86,7 %.

En 2022, Arroyo JCT et al. [115] ont utilisé l'algorithme génétique pour optimiser les paramètres d'un réseau neuronal artificiel afin de prédire les maladies cardiovasculaires. Un ensemble de données MCV de 70 000 cas avec 12 variables a été utilisé pour la formation et les tests. Les résultats de la simulation ont montré que l'utilisation de AG a amélioré les performances de RNA de 5,08 points de pourcentage, surpassant les autres algorithmes d'apprentissage automatique et atteignant une exactitude de 73,43 %.

En 2023, Talukdar J et al. [116] ont développé un modèle prédictif pour le diagnostic précoce des MCV basé sur un RNA. Ils l'ont appliqué à un ensemble de données comprenant huit variables, examiné et présenté diverses associations entre les facteurs de risque. Le modèle de perceptron multicouche avec algorithme de rétropropagation a atteint une exactitude de 81%.

En 2022 Edeh et al. [117] ont réalisé une étude visant à prédire la maladie du foie liée à l'hépatite "C" en utilisant des approches d'apprentissage automatique. Ils ont construit un modèle d'ensemble basé sur l'intelligence artificielle capable de prédire la fibrose avancée en intégrant des données cliniques et des biomarqueurs sanguins. Leur ensemble modèle a été comparé à trois approches d'apprentissage automatique classiques, démontrant une précision supérieure, notamment avec la technique du perceptron multicouche PML atteignant une précision de 94,1%. Cette prédiction précise est cruciale compte tenu de la prévalence mondiale significative de l'hépatite "C" et de ses conséquences graves sur le foie. Les résultats suggèrent que l'utilisation d'un modèle d'ensemble basé sur l'intelligence artificielle peut améliorer la précision de la prédiction de la maladie, offrant ainsi des avantages potentiels dans la gestion des risques et l'amélioration des résultats thérapeutiques.

Drobo and Smajović 2022 [118] ont proposé l'application des RNA pour le diagnostic de l'hépatite C. Ils ont utilisé une base de données de 1000 répondants répartis en deux groupes : sains (n = 200) et malades (n = 800). Les paramètres de surveillance comprenaient l'albumine, la phosphatase alcaline, l'alanine aminotransférase, l'aspartate aminotransférase, la bilirubine, l'acétylcholinestérase et les anticorps anti-VHC. L'exactitude globale de RNA développé a atteint

97,78%, soulignant ainsi le potentiel énorme de l'intelligence artificielle dans le diagnostic de l'hépatite "C". Les résultats suggèrent que dans l'avenir, une attention particulière devrait être accordée au développement de nouveaux systèmes avec une quantité maximale de données.

Yarasuri et al. 2019 [119] ont mené une étude visant à sélectionner l'outil optimal pour le diagnostic et la détection de l'hépatite. Dans cette recherche, une étude comparative entre divers outils d'apprentissage automatique et réseaux neuronaux a été réalisée. Les critères de performance reposent sur le taux de précision et l'erreur quadratique moyenne. Les algorithmes d'apprentissage automatique tels que les SVM, le KNN et le RNA ont été considérés comme des outils de classification et de prédiction pour le diagnostic de la maladie hépatique. Une étude succincte sur ces algorithmes a été menée, axée sur la précision de prédiction du diagnostic de la maladie. Tous les algorithmes d'apprentissage automatique ont été mis en œuvre et validés.

El-Solh et al. [120] ont implémenté un algorithme de réseau neuronal de régression générale (RNRG) pour prédire la tuberculose pulmonaire active. Le modèle d'entrée a été construit en utilisant 21 paramètres uniques, regroupés en trois catégories principales, à savoir les symptômes, les caractéristiques démographiques, constitutionnelles, et les résultats radiographiques. Le nombre de neurones nécessaires pour former la couche cachée a été déterminé en fonction du nombre de modèles utilisés lors de la formation de l'ensemble de données, étant donné que les RNRG exigent un neurone par motif traité. Pour évaluer le modèle, ils ont appliqué la méthode de validation croisée 10 fois. La précision diagnostique du modèle s'est élevée à 92,30%

Alakus and Turkoglu 2020 [121] ont développé des modèles prédictifs cliniques utilisant l'apprentissage profond et des données de laboratoire pour estimer quels patients sont susceptibles de contracter la maladie COVID-19. Afin d'évaluer les performances prédictives de leurs modèles, ils ont calculé des mesures telles que la précision, le F score, le rappel, l'AUC- ROC et l'exactitude. Ces modèles ont été testés sur 18 résultats de laboratoire issus de 600 patients et validés à l'aide d'une approche de validation croisée à 10 fois ainsi qu'avec une division entraînement-test.

Les résultats expérimentaux indiquent que leurs modèles prédictifs identifient les patients atteints de la maladie COVID-19 avec une exactitude de 86,66%, un F- score de 91,89%, une précision de 86,75%, un rappel de 99,42% et une AUC de 62,50%. Il est observé que les modèles prédictifs formés sur des résultats de laboratoire pourraient être utilisés pour prédire l'infection COVID-19, offrant ainsi une aide précieuse aux experts médicaux pour prioriser correctement les ressources.

Sumit Kumar et al [122] ont proposé une méthode de détection du paludisme utilisant un réseau de neurones convolutionnels profonds. Le jeu de données d'images de cellules de paludisme a été obtenu à partir du site officiel du NIH. L'objectif de la collecte de ce jeu de données était de réduire la charge des microscopistes dans les régions aux ressources limitées et d'améliorer la précision diagnostique grâce à un algorithme basé sur l'intelligence artificielle pour détecter et segmenter les globules rouges. Leur étude démontre qu'une précision de pointe peut être atteinte même en utilisant un réseau convolutionnel à deux couches, établissant ainsi une nouvelle référence dans les efforts de détection du paludisme avec l'IA. Leur méthode fonctionne avec un niveau de précision de pointe sans nécessiter de prétraitement coûteux.

Muhammad Adnan Khan et al [123], ont proposé un système de modélisation de prévision des pandémies de grippe utilisant un réseau neuronal à propagation anticipée (MSDII-FFNN). Ce modèle prédit les épidémies et détermine quel type de grippe pourrait devenir une pandémie, ainsi que les zones géographiques infectées. Les données du modèle ont été collectées à l'aide d'appareils IoT. Le modèle est divisé en deux phases : la phase de formation et la phase de validation, toutes deux connectées via le cloud. Lors de la phase de formation, e modèle a été formé à l'aide de FFNN et mis à jour sur le cloud. Lors de la phase de validation, lorsque les données ont été envoyées via les appareils IoT, le modèle du système a été mis à jour via le cloud et a prédit l'alarme pandémique. Les données sont divisées en un taux de formation de 85 % et un taux de validation de 15 %. En appliquant le modèle proposé à l'ensemble de données, la précision de sortie attendue a atteint 90 %.

III. 9 Conclusion

La prédiction médicale, essentielle pour les cliniciens, a suscité un intérêt croissant dans la littérature. Face à une augmentation notable du taux de mortalité, les chercheurs se tournent vers les méthodes d'intelligence artificielle pour une détection précoce et précise des maladies.

L'application de l'intelligence artificielle dans le domaine médical a connu un succès remarquable, offrant une assistance décisionnelle, un soutien aux professionnels de la santé et des prédictions précises. Toutefois, elle fait face à des défis majeurs tels que la qualité des données, la sécurité, la confidentialité et l'interprétabilité. Des solutions émergentes suggèrent la génération de données de qualité, une optimisation de la phase d'apprentissage, et une exploitation judicieuse des méthodes d'IA, dont l'hybridation des méthodes d’apprentissage automatique avec des algorithmes d’optimisation, promettant une réduction du temps et des coûts.

Ce chapitre a exploré divers travaux adoptant différentes approches pour la prédiction médicale. Le prochain chapitre présentera nos contributions.

Chapitre IV

Matériels et Méthodes

IV Matériels et méthodes

IV.1 Introduction

Cette étude repose sur une fondation méthodologique robuste visant à explorer et à implémenter des contributions majeures liées à la prédiction des affections transmissibles et non transmissibles grâce à IA. Dans ce chapitre crucial, nous exposerons en détail les approches méthodologiques que nous avons adoptées afin d'atteindre nos objectifs de recherche.

La première contribution repose sur l'exploitation de la technologie des RNA pour la détection précoce du COVID-19. Face à l'émergence mondiale de cette pandémie, l'application de l'IA pour la prédiction et le diagnostic précis de cette maladie a acquis une pertinence exceptionnelle. Cette contribution spécifique explore la puissance des RNA dans la détection du COVID-19. Les détails méthodologiques de cette approche novatrice seront scrupuleusement exposés dans cette section, démontrant ainsi son potentiel dans le contexte de la prédiction des maladies infectieuses.

Notre deuxième contribution s'articule autour de l'application de la technologie des RNA pour la détection de l'hépatite virale, en utilisant l'ensemble de données de l'UCI comme base d'analyse. L'hépatite virale, une maladie hépatique inflammatoire, représente un défi majeur pour la santé publique. Cette contribution vise à démontrer l'efficacité des RNA dans la prédiction précise de l'hépatite virale, en exploitant les caractéristiques complexes de l'ensemble de données de l'UCI. Les méthodes spécifiques employées, les paramètres choisis seront exposés en détail dans cette section.

La troisième contribution de cette thèse repose sur une étude dédiée à la prédiction du cancer de poumon à l'aide des techniques avancées de IA. Dans ce cadre, nous avons exploité plusieurs algorithmes, dont les RNA, KNN, SVM, et avons développé un nouvel algorithme hybride : le CSABC-SVM. Le choix de ces méthodes s'est basé sur leur efficacité prouvée à identifier des modèles complexes dans des données médicales, facilitant ainsi la prédiction fiable de cas de cancer du poumon. L'application de ces algorithmes tire profit des caractéristiques distinctives des données cliniques pour une analyse approfondie et une prise de décision autonome. L'objectif de cette contribution est d'évaluer la performance relative de ces approches pour une détection précoce du cancer, pavant la voie vers des progrès notables dans le dépistage préventif des maladies oncologiques.

IV.2 Logiciels utilisés

Divers logiciels ont été utilisés lors de ces études, comme l'illustre le tableau IV. 1 "Microsoft Office Excel 2016" a été utilisé pour organiser et prétraiter la base de données, ainsi que pour effectuer les calculs nécessaires.

L'ensemble de l'entraînement de model RNA et un certain nombre d'étapes ont été réalisés par le logiciel appelé STATISTICA Versions 10 (Stat Soft 2010), compte tenu de la fréquence à laquelle il apparaît dans les travaux d'auteurs impliquant Hamadache et al [124], et celles d'Ammi et al [66].

L'ensemble de l'entraînement des modèles (RNA, SVM, KNN, CSABC-SVM) et un certain nombre d'étapes ont été réalisés à l'aide du logiciel appelé "Python".

Tableau IV. 1 Logiciels utilisés

Nom du logiciel	Logo
Microsoft Office Excel 2016	
STATISTICA	
Python	

IV.3 Mesures d'évaluation des performances des modèles de classification

- **La matrice de confusion**

Une matrice de confusion est une représentation des prédictions basée sur les résultats obtenus lors d'un problème de classification [125]. C'est un tableau qui résume le nombre de prédictions correctes et incorrectes [125]. La matrice décrit la performance d'un modèle classique sur un ensemble de données donné. Elle montre les erreurs commises par un classificateur ainsi que le type d'erreurs qui sont faites [125].

La matrice de confusion est généralement présentée sous la forme d'un tableau 2x2 pour les problèmes de classification binaire (voir la figure IV.1), mais elle peut être étendue pour des problèmes de classification avec plus de classes. Les colonnes de la matrice représentent les classes réelles par le modèle, tandis que les lignes représentent les classes prédites des instances testées.

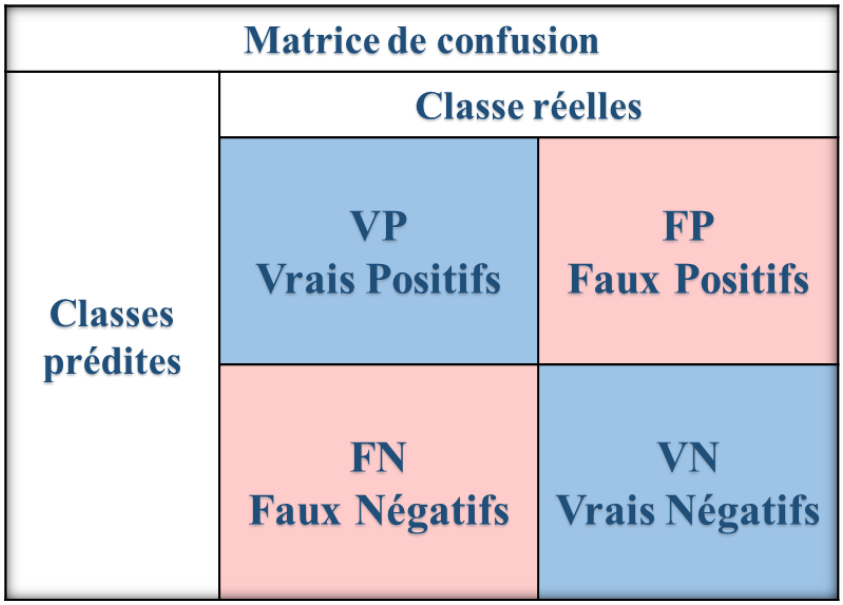


Figure IV.1 Matrice de confusion pour une classification binaire.

VP (vrais positifs) : représente le nombre d'individus malades avec un test positif.

FP (faux positifs) : représente le nombre d'individus non malades avec un test positif.

FN (faux négatifs) : représente le nombre d'individus malades avec un test négatif.

VN (vrais négatifs) : représente le nombre d'individus non malades avec un test négatif.

Exactitude ou Accuracy (Ac) : C'est une mesure de performance utilisée pour évaluer la qualité d'un modèle de classification. Elle représente le nombre de prédictions correctes faites par le modèle par rapport au nombre total de prédictions. En d'autres termes, c'est le rapport entre le nombre de prédictions correctes et le nombre total d'observations. L'exactitude est calculée par Eq IV.1

$$Ac = (VP + VN)/(VP + VN + FP + FN) \quad \text{(IV.1)}$$

- **La sensibilité ou rappel (Se) :** La sensibilité mesure la capacité d'un modèle à identifier correctement les individus malades parmi tous les individus réellement malades présents dans l'ensemble de données. La sensibilité est calculée par Eq IV.2

  $$Se = VP/(VP + FN) \quad \text{(IV.2)}$$

- **La spécificité (Sp) :** La spécificité, dans le contexte du dépistage des maladies, exprime la capacité du modèle à bien détecter les non-malades. En d'autres termes, elle mesure la capacité du test à exclure de manière précise les cas négatifs parmi l'ensemble des individus effectivement sains. La spécificité est calculée par Eq IV. 3.

  $$Sp = VN/(VN + FP) \quad \text{(IV. 3)}$$

La qualité du test repose sur la recherche du meilleur équilibre entre sensibilité et spécificité.

L'évaluation de ce compromis peut être évalué à l'aide de l'indice de la courbe de caractéristique de fonctionnement du récepteur (en anglais : Receiver Operating Characteristic (ROC).

- **La précision (Pr)** : est une mesure de performance utilisée dans le contexte de la classification qui évalue la proportion d'observations positives correctement identifiées parmi toutes les observations identifiées comme positives par le modèle. Elle indique la capacité du modèle à éviter de faire des fausses prédictions positives. La précision est calculée par Eq IV. 4 [121].

  $$Pr = VP/(VP + FP) \quad \text{(IV. 4)}$$

- **Le F-score :** c'est une mesure statistiquement significative de précision et du rappel qui maintient un équilibre entre la précision et le rappel d'un classificateur, il est calculé par Eq IV. 5 [118].

  $$F - Score = (2 * Pr * Recall)/(Pr + Recall) \quad \text{(IV. 5)}$$

- **Le Coefficient de Corrélation de Matthews (MCC)** est une mesure statistique largement utilisée pour évaluer la performance des modèles de prédiction, notamment dans le domaine de la prédiction des maladies. Il mesure la qualité globale des prédictions binaires, évaluant la

capacité d'un modèle à distinguer entre individus malades et sains, en tenant compte à la fois des faux positifs et des faux négatifs. Le MCC génère une valeur entre -1 et +1, où 1 indique une prédiction parfaite, 0 équivaut à une performance aléatoire, et -1 indique une prédiction parfaitement inverse. Cette mesure offre une évaluation concise de l'efficacité d'un modèle dans le contexte spécifique de la prédiction des maladies. Le MCC est calculée par Eq (IV. 6).

$$MCC = (VP * VN) - (FP * FN)/\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)} \quad \text{(IV. 6)}$$

- **La courbes de ROC**

La courbe de Caractéristique de Fonctionnement du Récepteur (en anglais Receiver Operating Characteristic ROC) est un outil graphique utilisé dans l'évaluation des performances des modèles de classification, y compris ceux utilisés dans le domaine médical pour la prédiction des maladies. Cette courbe représente la relation entre le taux de vrais positifs (sensibilité) et le taux de faux positifs (1 - spécificité) pour différents seuils de classification. La figure IV. 2 représente un exemple de courbe ROC.

Axe des x (1 - spécificité) : Il représente le taux de faux positifs, c'est-à-dire le pourcentage d'individus sains incorrectement classés comme malades.

Axe des y (sensibilité) : Il représente le taux de vrais positifs, c'est-à-dire le pourcentage d'individus malades correctement classés comme malades.

Le modèle est plus performant lorsque la courbe ROC se rapproche du coin supérieur gauche du graphique. En d'autres termes, une courbe ROC plus proche du coin supérieur gauche indique une sensibilité et une spécificité plus élevées, ce qui est essentiel dans de nombreuses applications, y compris la prédiction des maladies.

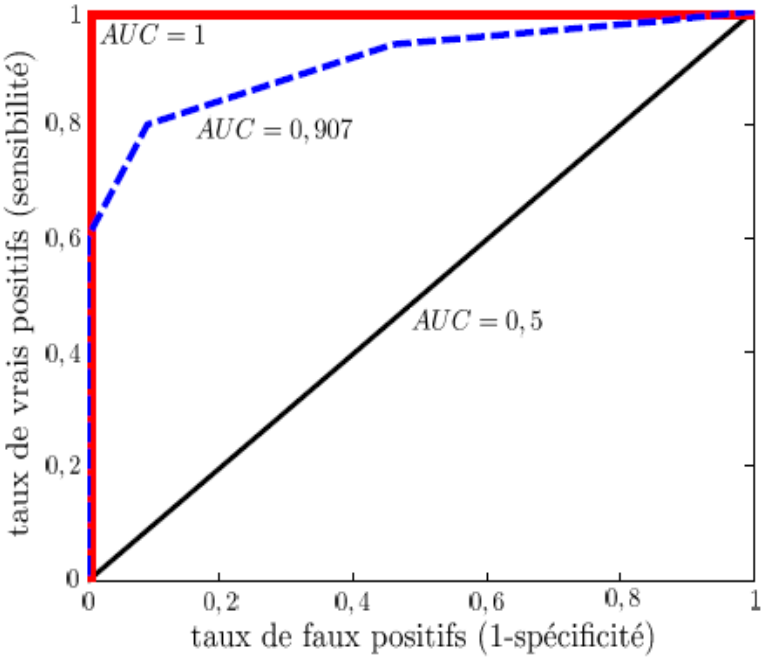


Figure IV. 2 Illustration de la courbe ROC [127].

À partir de la courbe de ROC, l'indice numérique qui permet d'évaluer la performance globale de la courbe est l'aire sous la courbe de ROC (en anglais : Area Under the Curve (AUC)). Cette métrique mesure la capacité du modèle à discriminer entre différentes classes, comme la présence ou l'absence d'une pathologie. En d'autres termes, l'AUC-ROC peut être interprété comme la probabilité que le modèle prédise correctement la maladie chez un patient. Plus l'AUC-ROC est proche de 1, meilleure est la capacité du modèle à effectuer des prédictions précises, renforçant ainsi la confiance dans sa capacité à distinguer les cas positifs des cas négatifs. Cet indice synthétique offre une vue d'ensemble significative des performances du modèle dans le contexte de la prédiction de pathologies.

Deux situations particulières sont identifiées sur la figure IV. 2.

- Une AUC de 0,5 suggère que le test se rapproche du hasard et ne contribue guère au diagnostic. Dans ce cas, la présence du signe indiquant la maladie est indépendante de la pathologie. En d'autres termes, le modèle ne parvient pas à distinguer entre les deux classes.
- Une AUC égale à 1 représente le meilleur résultat possible, avec une sensibilité et une spécificité toutes deux à 100%. Dans cette situation optimale, les deux classes sont complètement séparées, indiquant que le modèle parvient à différencier de manière parfaite les individus atteints de la maladie de ceux qui ne le sont pas.

IV.4 Prédictions des affections transmissibles par IA

IV.4.1 Contribution I : Prédiction de COVID-19 par RNA-PMC

La méthodologie de notre étude était basée sur un ensemble d'étapes, résumées dans l'organigramme représenté sur la figure IV. 3

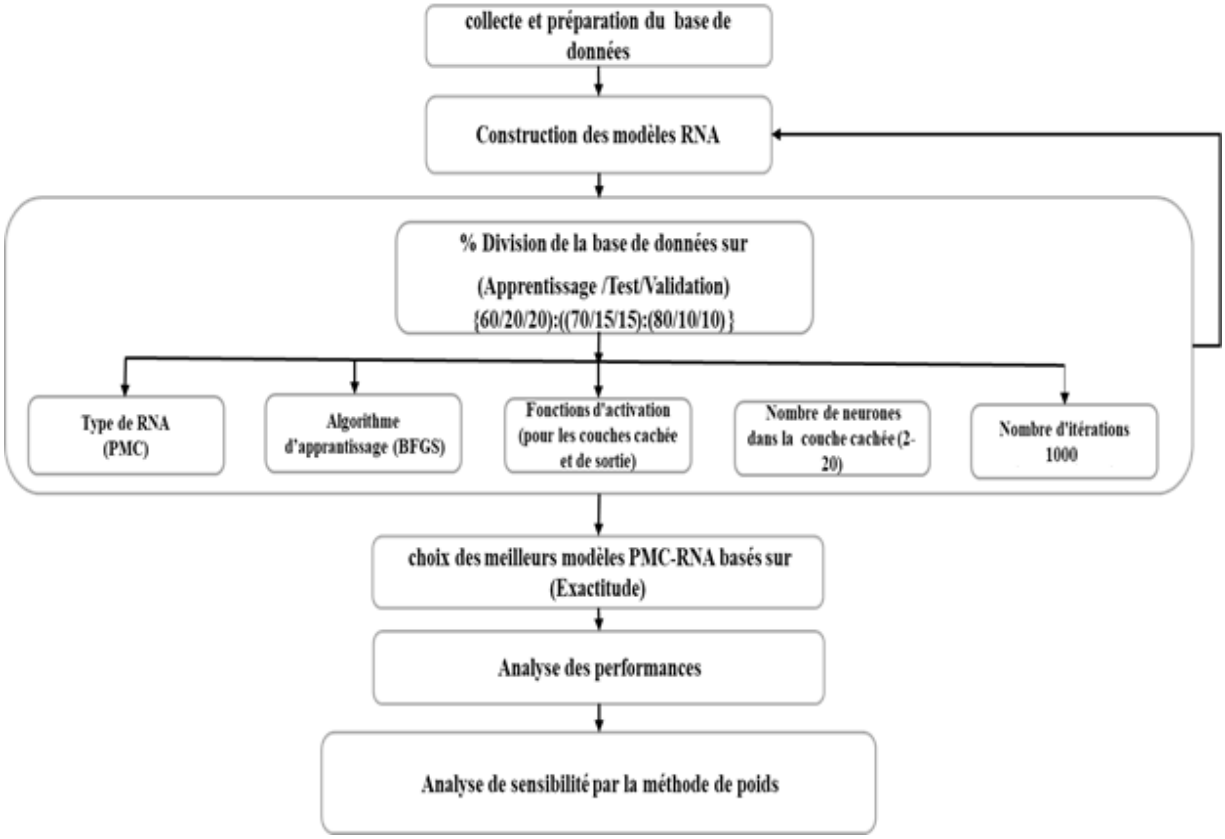


Figure IV. 3 Diagramme de la procédure appliquée pour la prédiction de COVID-19 par RNA-PMC

IV.4.1.1 Collecte de la base de données

Dans cette étude, un ensemble de données sur la COVID-19 a été traité pour concevoir notre modèle. L'ensemble de données utilisé a été collecté à partir de [httpsgithub.comakarshan-githubCovid-19-detector-probability] [128] et se compose de 4 000 instances avec 7 attributs. Le Tableau IV. 2 décrit en détail de tous les attributs.

Tableau IV. 2 Description statistique des attributs de la base de données COVID-19.

Attributs	**Description**
Age	Valeur numérique
Température corporelle	°F (fahrenheit)
Fatigue	Non (0) Oui (1)
Toux	Non (0) Oui (1)
Douleurs corporelles	Non (0) Oui (1)
Mal de gorge	Pas du tout (0) Un peu (1) Trop (2)
Difficulté à respirer	Pas du tout (0) Un peu (1) Trop (2)
Infecté	Non (0) Oui (1)

IV.4.1.2 Prétraitement et analyse des données

Dans cette étude, nous avons utilisé le logiciel "Python" pour notre analyse des données. L'ensemble de données que nous avons utilisé était caractérisé par l'absence de valeurs manquantes ou de valeurs aberrantes, rendant ainsi inutile tout prétraitement supplémentaire des données.

Pour identifier les corrélations entre les attribues, nous avons utilisé une matrice de corrélation. La corrélation entre les variables mesure la relation statistique entre deux variables. Elle indique dans quelle mesure les variations d'une variable sont associées aux variations d'une autre variable. La corrélation est souvent exprimée par le coefficient de corrélation. Où une valeur de coefficient de corrélation est inférieure à zéro indique une corrélation négative et une valeur nulle indique une absence de corrélation entre les variables [129].

De plus, nous avons utilisé des graphiques pour analyser la répartition des classes (infecté, non infecté) de la COVID-19. Nous avons également utilisé des graphiques de l'estimation de la densité

du noyau (kernel Density Estimation (KDE)) pour comparer la répartition des patients par âge, en distinguant ceux atteints de la COVID-19 de ceux qui n'en sont pas affectés. De plus, un graphique montre comment différentes catégories de variables sont associées aux maladies en fonction de la prévalence de la maladie dans l'ensemble de données.

IV.4.1.3 Méthode de classification

Dans ce travail de recherche, un modèle PMC-RNA a été choisi pour la prédiction de COVID-19. Le logiciel STATISTICA version 10 a été utilisé. La construction du modèle PMC-RNA a suivi les étapes suivantes :

- La base de données a été divisée en trois sous-ensembles : une grande partie pour l'apprentissage, et le reste pour le test et la validation, comme suit :
    - Première division : 60 % pour l'apprentissage, 20 % pour les tests et 20 % pour la validation
    - Deuxième division : 70 % pour l'apprentissage, 15 % pour les tests et 15 % pour la validation
    - Troisième division : 80 % pour l'apprentissage, 10 % pour les tests et 10 % pour la validation
- Le modèle est composé d'une couche d'entrée avec 7 neurones, d'une couche de sortie avec deux neurones, et d'une couche cachée avec une variation du nombre de neurones de (2 à 20).
- L'algorithme de rétro-propagation quasi Newton, en particulier l'algorithme de Broyden-Fletcher-Goldfarb-Shanno (BFGS), a été utilisé comme algorithme d'apprentissage [129].
- Les fonctions d'activation sont sélectionnées en utilisant différents types de fonctions dans les couches cachée et logistique dans la couche de sortie, comme illustré dans la figure IV. 4.
- Compte tenu du nombre de couches cachées, nous avons utilisé une seule couche cachée pour tous les modèles PMC-RNA, puisqu'il n'y en a qu'une dans le logiciel 'STATISTICA', version 10 (StatSoft 2010).
- L'exécution implique 1000 itérations pour chaque division, comme illustré dans la figure IV. 3.

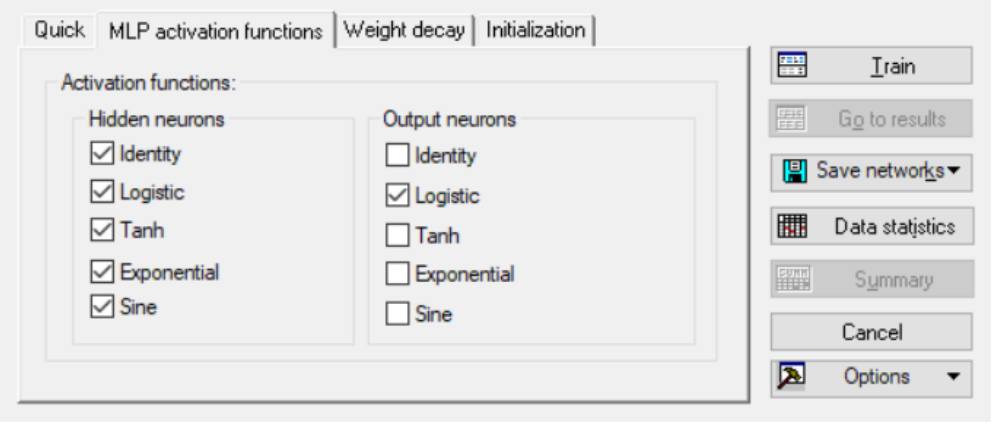


Figure IV. 4. Capture d'écran des fonctions d'activation de STATISTICA utilisées

IV.4.1.4 Mesures d'évaluation des performances et le choix du meilleur modèle

Les performances du modèle sont évaluées à l'aide de la matrice de confusion, ainsi que de paramètres tels que l'exactitude, la sensibilité, la spécificité la précision, le F-score et le MCC. Ces derniers sont calculés en se basant sur les valeurs présentes dans la matrice de confusion, ce qui permet d'obtenir une évaluation détaillée de la performance du modèle dans la classification des données.

Le choix du meilleur modèle repose sur l'exactitude (accuracy). Une valeur d'exactitude optimale est égale à 1, indiquant qu'un modèle a réalisé toutes les prédictions correctes. À l'inverse, une exactitude de 0 signifie que le modèle est incapable de faire des prédictions correctes.

Une exactitude supérieure à 0,9 est considérée excellente, une exactitude dépassant 0,7 est jugée bonne, tandis que toute autre valeur est généralement considérée comme insatisfaisante.

IV.4.1.5 Analyse de sensibilité par la méthode de poids

L'analyse de sensibilité par la méthode des poids est une technique puissante utilisée pour évaluer l'impact de la contribution de chaque attribut d'entrée sur la sortie [130, 131].

Il est utilisé à de nombreuses fins, notamment la validation du modèle, l'évaluation du comportement de diverses variables du modèle et la confirmation du choix des variables du modèle en déterminant la contribution de chaque variable choisie au modèle. Il est donc important d'appliquer cette analyse au modèle PMC-RNA [127].

L'analyse de sensibilité a été réalisée en utilisant la méthode du poids. Cette méthode, initialement proposée par Garson [132] et ensuite reprise par Goh [133], permet de quantifier l'importance relative (IR) des différents entrées sur la sortie du réseau neuronal [134]. Elle repose sur le partitionnement des poids de connexion.

Poids des connexions entrée-cachée $\boldsymbol{W_{ji}^{I}}$

Poids des connexions cachées – sortie $\boldsymbol{W_{1j}^{H}}$

Les procédures détaillées de cette méthode sont décrites dans les études de [46, 51, 130]. Voir La figure IV. 5.

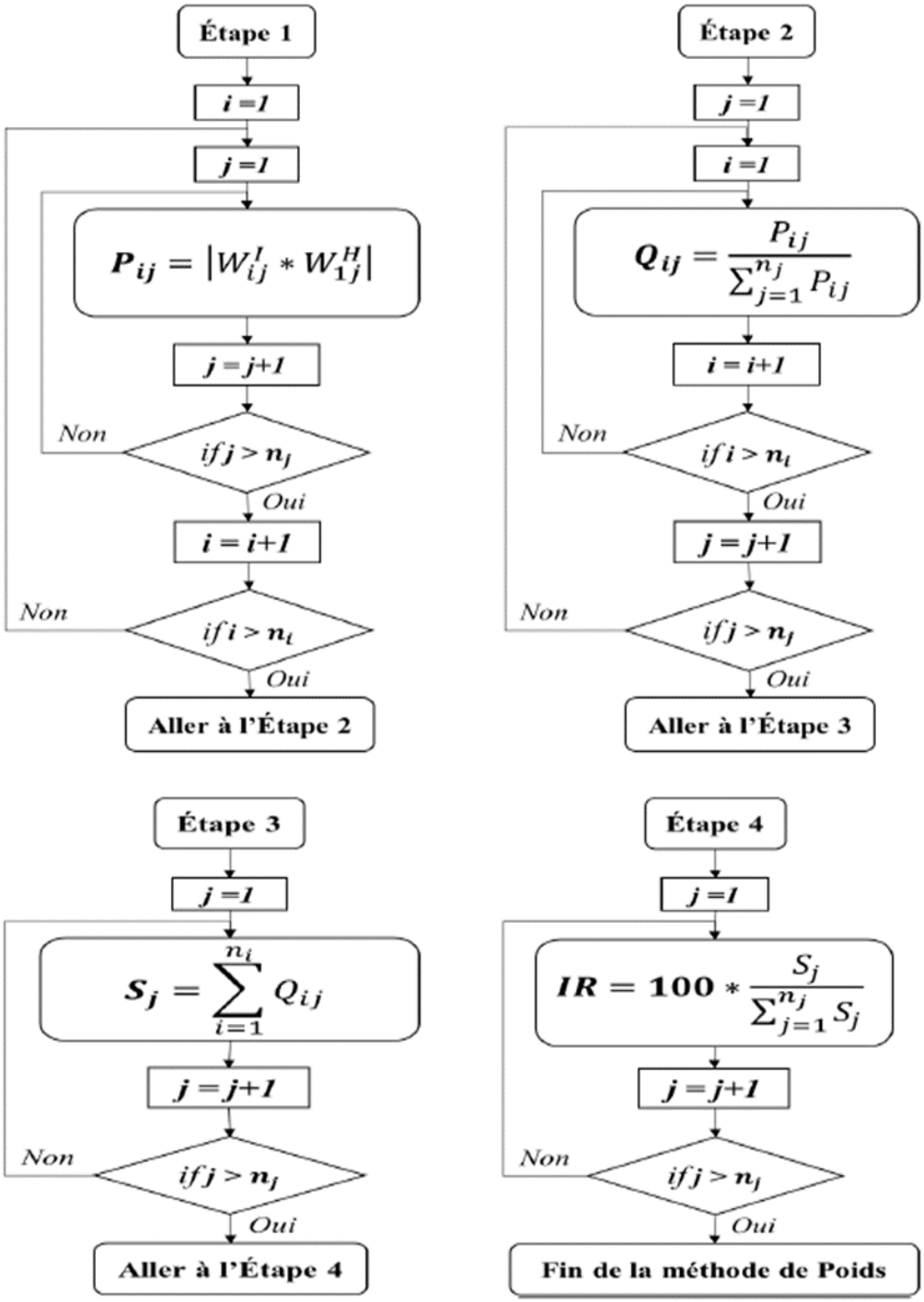


Figure IV. 5 Organigramme via la méthode d'analyse de sensibilité par la méthode de poids [51, 130]

IV.4.2 Contribution II : Prédiction de l'hépatite par PMC-RNA

La méthodologie de notre étude s'articule autour de plusieurs phases essentielles, résumées dans l'organigramme représenté sur la figure IV. 6.

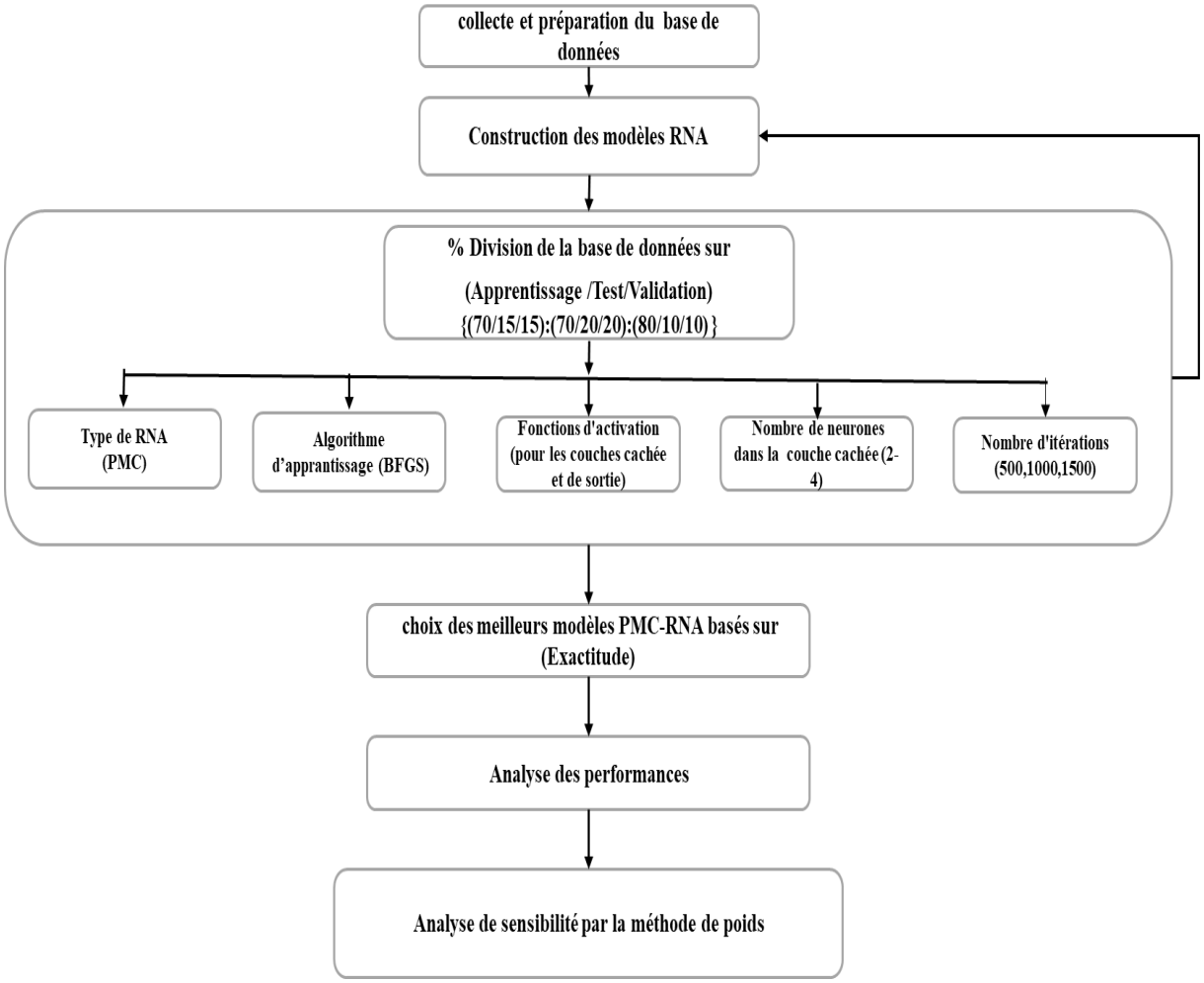


Figure IV. 6 Diagramme de la procédure appliquée pour la prédiction de l'hépatite virale par RNA-PMC

IV.4.2.1 Collecte de la base de données

Dans cette étude, un ensemble de données sur l'hépatite virale a été traité pour concevoir notre modèle. L'ensemble de données utilisé a été collecté à partir du référentiel d'apprentissage automatique de [http://archive.ics.uci.edu/ml/]. L'ensemble de données est composé de 155 instances avec 20 attributs, le tableau IV. 3 décrit les détails de tous les attributs.

Tableau IV. 3 Description statistique des attributs de la base de données d'hépatite.

Attributs	Value
Âge	Valeur Numérique (10, 20, 30, 40, 50, 60, 70, 80)
Sexe	Male (1) Femelle (2)
Stéroïde	Non (1) Oui (2)
Antiviraux	Non (1) Oui (2)
Fatigue	Non (1) Oui (2)
Malaise	Non (1) Oui (2)
Anorexie	Non (1) Oui (2)
Foie gros	Non (1) Oui (2)
Ferme du foie	Non (1) Oui (2)
Rate palpable	Non (1) Oui (2)
Araignées vasculaires	Non (1) Oui (2)
Ascite	Non (1) Oui (2)
Varices	Non (1) Oui (2)
Bilirubine	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alk phosphatase	33, 80, 120, 160, 200, 250
Aspartate Aminotransférase SGOT	13, 100, 200, 300, 400, 500
Albumine	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime (temps de prothrombine ou TP)	10, 20, 30, 40, 50, 60, 70, 80, 90
Histologie	Non (1) Oui (2)
Classe (hépatite viral)	Oui (1) Non (2)

IV.4.2.2 Prétraitement et analyse des données

La base de données sur l'hépatite a été traitée en supprimant les lignes qui contenaient des valeurs manquantes.

Pour détecter les corrélations entre les attributs, nous avons utilisé une matrice de corrélation. De plus, nous avons utilisé des graphiques pour analyser la distribution des classes (oui, non) pour l'hépatite. Nous avons également utilisé un graphique de l'estimation de la densité du noyau KDE pour comparer la répartition des patients par âge et par sexe, en distinguant ceux atteints de l'hépatite de ceux qui n'en sont pas atteints.

IV.4.2.3 Méthode de classification

Dans ce travail de recherche, un modèle PMC-RNA a été sélectionné pour la prédiction de l'hépatite virale. Le logiciel STATISTICA, version 10, a été utilisé. La construction du modèle PMC-RNA a suivi les étapes suivantes :

La base de données a été divisée en trois sous-ensembles : une grande partie pour l'apprentissage, et le reste pour les tests et la validation, comme suit :

- Première division : 70 % pour l'apprentissage, 15 % pour les tests et 15 % pour la validation
- Deuxième division : 70 % pour l'apprentissage, 20 % pour les tests et 10 % pour la validation
- Troisième division : 80 % pour l'apprentissage, 10 % pour les tests et 10 % pour la validation

- Le modèle est composé d'une couche d'entrée avec 19 neurones, d'une couche de sortie avec 2 neurones, et d'une couche cachée avec une variation du nombre de neurones de (2 à 4).
- L'algorithme de BFGS a été utilisé comme algorithme d'apprentissage [129].
- Les fonctions d'activation sont sélectionnées en utilisant différents types de fonctions dans les couches cachée et la logistique dans la couche de sortie, comme illustré dans la figure IV. 4. Ainsi, à la fin de l'exécution, le logiciel fournit la combinaison optimale des deux fonctions pour activer la couche concernée.
- Compte tenu du nombre de couches cachées, nous avons utilisé une seule couche cachée pour tous les modèles PMC-RNA, puisqu'il n'y en a qu'une dans le logiciel STATISTICA, version 10 (StatSoft 2010).
- L'exécution implique un nombre important d'itérations, entre 500 et 1500 itérations pour chaque division, comme illustré dans la figure IV. 6.

Les performances du modèle sont évaluées par la matrice de confusion et par des paramètres tels que l'exactitude, la sensibilité, la spécificité et la précision, qui sont calculés à partir de la matrice de confusion. Le choix du meilleur modèle dépend de la valeur de l'exactitude.

IV.4.2.4 Analyse de sensibilité par la méthode de poids

L'analyse de sensibilité par la méthode des poids a été utilisée pour évaluer l'impact de chaque attribut d'entrée de l'ensemble de données sur la sortie de l'hépatite. Les procédures de cette méthode sont décrites dans les études de [46, 51, 130]. Voir La figure IV. 5.

IV.5 Prédictions des affections non transmissibles par IA

IV.5.1 Contribution III : Prédiction de cancer du poumon par CSABC-SVM

Dans cette étude, la méthodologie a été structurée autour d'une série d'étapes décrites dans l'organigramme représenté à la figure IV. 7

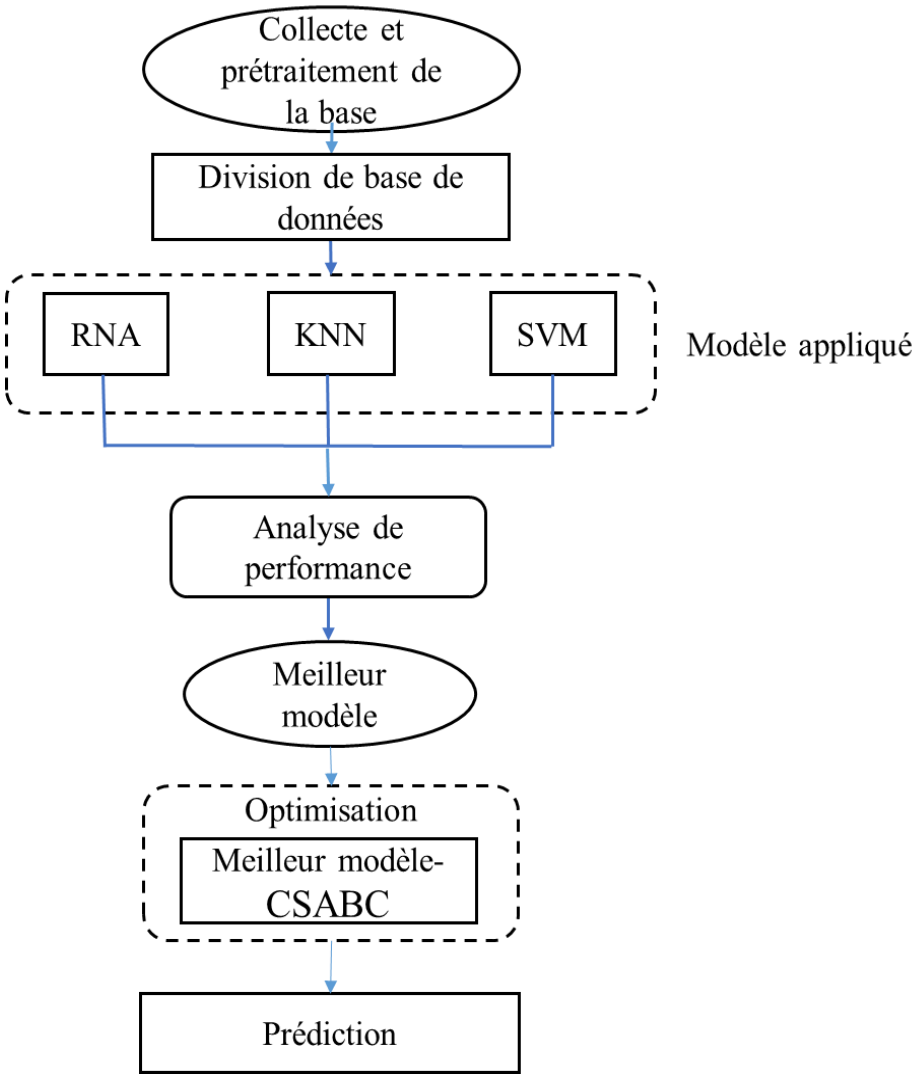


Figure IV. 7 Structure organisationnelle de la procédure appliquée pour la prédiction de cancer du poumon

IV.5.1.1 Collecte de la base de données

Dans cette étude, nous avons traité un ensemble de données sur le cancer du poumon pour développer le modèle proposé comme illustré dans le tableau IV. 4. L'ensemble de données provient de [135], englobant un total de 16 attributs. Dans cet ensemble de données, nous avons trouvé les enregistrements de 309 patients, comprenant 162 hommes et 147 femmes. Parmi ces patients, 39 ont été classés comme normaux, tandis que 270 ont reçu un diagnostic de cancer du poumon.

Tableau IV. 4 Description des attributs de la base de données de cancer du poumon

Attribut	Description
Sexe	Femme = 0 Homme = 1
Âge	Âge en années (21- 87)
Fumeur	Oui = 2 Non = 1
Doigts jaunes	Oui = 2 Non = 1
Anxiété	Oui = 2 Non = 1
Pression des pairs	Oui = 2 Non = 1
Maladie chronique	Oui = 2 Non = 1
Fatigue	Oui = 2 Non = 1
Allergie	Oui = 2 Non = 1
Une respiration sifflante	Oui = 2 Non = 1
Consommation d'alcool	Oui = 2 Non = 1
Tousser	Oui = 2 Non = 1
Essoufflement	Oui = 2 Non = 1
Difficulté à avaler	Oui = 2 Non = 1
Douleur thoracique	Oui = 2 Non = 1
Cancer du poumon	Oui = 2 Non = 1

IV.5.1.2 Prétraitement et analyse des données

Dans cette étude, le langage de programmation python a été utilisé pour notre analyse des données. L'ensemble de données dont nous avons bénéficié était caractérisé par l'absence de valeurs manquantes ou de valeurs aberrantes, rendant inutile un prétraitement des données. Pour identifier les corrélations entre les attributs, nous avons utilisé une matrice de corrélation.

De plus, nous avons utilisé des graphiques pour analyser la répartition des classes (malade, non malade) de cancer du poumon. Nous avons également utilisé des graphiques de l'estimation de la densité du noyau KDE pour comparer la répartition des patients par âge, en distinguant ceux

atteints le cancer du poumon de ceux qui n'en sont pas. De plus, un graphique montre comment différentes catégories de variables sont associées aux maladies en fonction de la prévalence de la maladie dans l'ensemble de données.

IV.5.1.3 Méthode de classification

Dans cette étude, nous avons effectué une analyse exhaustive d'un ensemble de données sur le cancer du poumon, en utilisant plusieurs algorithmes de classification, notamment RNA, KNN et SVM. Pour évaluer leur efficacité, nous avons divisé l'ensemble de données en un ensemble d'apprentissage comprenant 80 %, et un ensemble de test comprenant 20 % des données. Des mesures de performance ont ensuite été calculées pour chaque modèle. Nous avons effectué une comparaison complète des performances des trois modèles pour identifier le modèle le plus prometteur.

IV.5.1.4 Optimisation

Après avoir identifié le modèle le plus prometteur, nous sommes concentrés sur l'optimisation pour améliorer l'exactitude et d'autres mesures de performance. Cette stratégie d'optimisation était distinctive, car nous avons exploité une approche nouvelle combinant les avantages des algorithmes CSA et ABC en conjonction avec le modèle le plus performant. Cette hybridation a été conçue pour améliorer encore les capacités du modèle et fournir des résultats supérieurs.

Suite à l'évaluation des algorithmes à l'aide de mesures de performance, notre analyse a révélé que le SVM surpassait les autres algorithmes, soulignant ses performances supérieures. Néanmoins, il est important de reconnaître que même si le SVM a affiché des performances remarquables par rapport aux autres modèles de notre étude, cette exactitude n'atteint peut-être pas encore le seuil souhaité pour la prédiction du cancer du poumon. Alors que nous aspirons continuellement à atteindre une exactitude encore plus grande dans nos modèles prédictifs, nous restons déterminés à poursuivre l'amélioration continue et à explorer des approches alternatives. En réponse à ce défi, nous avons introduit le modèle hybride CSABC-SVM, conçu pour améliorer encore l'exactitude en capitalisant sur les synergies de ses composants constitutifs. Plus précisément, les algorithmes CSA et ABC sont appliqués pour optimiser les hyperparamètres du modèle SVM (C et γ). Notre engagement indéfectible à atteindre une précision accrue souligne notre engagement à fournir des prévisions fiables et précises dans le domaine de la prédiction du cancer du poumon.

La chose la plus importante que le modèle a faite est d'introduire une méthode d'optimisation hybride, appelée CSABC (La combinaison de l'algorithme de recherche de corbeaux et de la

colonie d'abeilles artificielles), pour trouver les meilleurs hyperparamètres du classificateur SVM. En utilisant une approche itérative sur une population de solutions représentant différentes combinaisons de valeurs pour les paramètres C et γ, le modèle cherche à maximiser l'exactitude du SVM sur les données d'entraînement. Cette approche combine les caractéristiques de l'algorithme de recherche du corbeau et de l'algorithme de colonie d'abeilles artificielles pour améliorer l'efficacité de l'optimisation des hyperparamètres. Ce processus d'optimisation se déroule sur 100 itérations, évoluant la population de solutions tout en maintenant le suivi de la meilleure solution rencontrée.

Le processus d'optimisation impliquant CSA et ABC hybridées avec SVM suit généralement deux étapes principales : l'exploration et l'exploitation. L'exploration pour trouver des solutions potentielles, et l'exploitation pour améliorer leur qualité.

Les étapes d'exécution du CSABC-SVM sont les suivantes :

1. Définition de la classe CSABC :

Nous définissons la classe CSABC responsable de l'optimisation des hyperparamètres SVM à l'aide de l'algorithme hybride CSA et ABC. Le constructeur prend deux paramètres : la taille de la population (nombre de solutions) et le nombre maximal d'itérations.

2. Optimisation des hyperparamètres :

A l'intérieur de la méthode optimisée de la classe CSABC, une population de solutions potentielles (combinaisons d'hyperparamètres C et γ) est générée aléatoirement.

La condition physique de la population est évaluée de manière itérative. L'adéquation de chaque solution est déterminée en entraînant un SVM avec les hyperparamètres donnés et en calculant son exactitude sur les données d'entraînement.

Met à jour la meilleure solution (les hyperparamètres qui donnent la plus grande exactitude) trouvée lors de chaque itération.

La population évolue en sélectionnant des solutions en fonction de leur aptitude et en introduisant des mutations pour explorer de nouvelles combinaisons d'hyperparamètres.

Le processus d'optimisation se poursuit pendant 100 itérations.

3. Sélection de la meilleure solution :

- Une fois le processus d'optimisation terminé, la meilleure solution (meilleurs hyperparamètres) est obtenue à partir de l'algorithme hybride CSA et ABC.

4. Extraction des hyperparamètres :

Les valeurs optimales des hyperparamètres de SVM C et γ sont extraites des meilleures solutions, qui sont utilisées pour construire le classificateur SVM.

IV.5.1.5 Importance des variables

Dans le domaine du ML, l'analyser de l'importance des variables et sa visualiser sont des techniques importantes et populaires [132]. En raison de la facilité d'utilisation et de l'interprétabilité du classement des caractéristiques ou de l'analyse des risques, ces techniques sont particulièrement utilisées dans des domaines comme la biologie et les sciences sociales [126]. L'importance et le classement des caractéristiques sont déterminés en fonction de la valeur du coefficient de chaque caractéristique [126, 136].

IV.6 Conclusion

L'ensemble des méthodes traduit notre engagement envers une approche rigoureuse et novatrice dans la prédiction des affections transmissibles et non transmissibles. Ces contributions s'appuient sur des méthodologies robustes, ancrées dans l'intelligence artificielle, pour relever les défis complexes de la santé mondiale. Cette section établit ainsi le cadre méthodologique nécessaire à la réalisation de nos objectifs de recherche, ouvrant la voie à des avancées significatives dans le domaine de la prédiction médicale.

Chapitre V

Résultats et Discussion

V Résultats et discussion

V.1 Introduction

Dans le domaine de la prédiction des maladies, nos efforts ont porté leurs fruits, produisant des résultats prometteurs. À travers une étude exhaustive, nous avons évalué les performances de différents modèles dans cette tâche cruciale. Certains de ces résultats ont été préalablement publiés dans " International Journal of Information Technology". Ce chapitre marque un moment décisif où nous examinons en détail et discutons des résultats que nous avons obtenus, dans le but de contribuer de manière significative à ce domaine de recherche.

Pour garantir la clarté et la cohérence de notre analyse, nous avons organisé nos résultats en différentes sections réfléchies. Dans la première partie, nous présenterons les performances des modèles RNA pour la prédiction du COVID-19. Par la suite, nous discuterons des résultats obtenus à partir de ces modèles RNA pour la prédiction de l'hépatite virale, et Nous avons mené une étude comparative avec des modèles existants dans la littérature.

Enfin, nous examinerons les résultats des modèles RNA, KNN, SVM, et CSABC-SVM pour la prédiction de cancer du poumon. Nous consacrerons une section à une comparaison des performances des modèles CSABC-SVM avec celles des SVM, ainsi qu'avec les modèles existants dans la littérature. Cette analyse comparative vise à fournir des connaissances précieuses pour l'amélioration des méthodes de prédiction des maladies, et à guider les futures recherches dans ce domaine crucial de la médecine préventive.

V.2 Prédiction des affections transmissibles

V.2.1 Contribution I : Prédiction de Covid-19 par RNA-PMC

V.2.1.1 Résultat de l'analyse exploratoire des données

Corrélation entre les variables

À partir de la matrice de corrélation illustrée dans la figure V.1, on constate une très faible corrélation entre les attributs. Les valeurs des coefficients de corrélation se situent dans l'intervalle [- 0.029, + 0.028]. Nous pouvons conclure que les attributs de notre ensemble de données ne sont pas fortement corrélés.

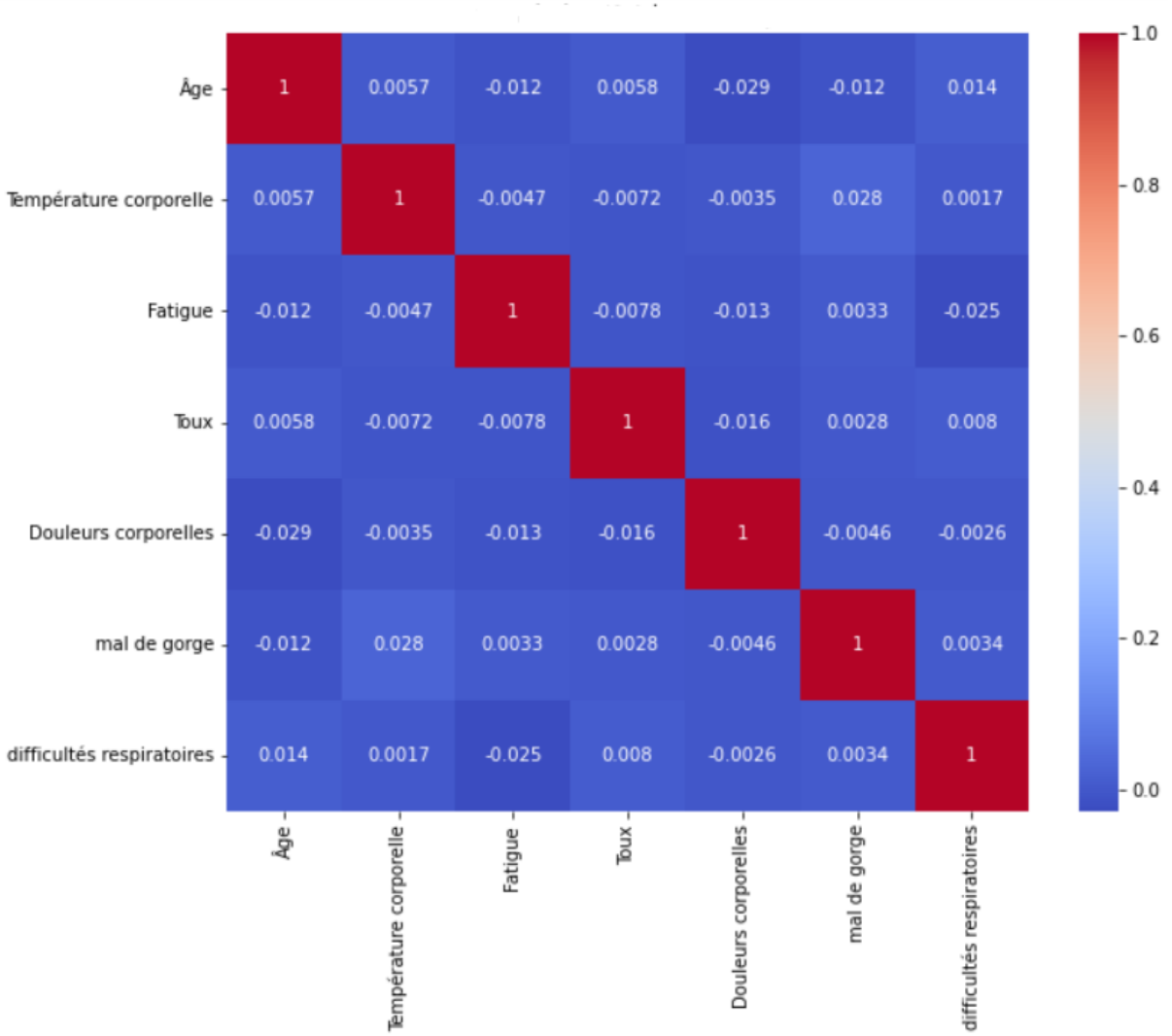


Figure V. 1 Matrice de corrélation de la prédiction de COVID-19 par RNA-PMC

Dans l'analyse de cet ensemble de données, il est essentiel de comprendre la répartition des patients dans chaque catégorie. Comme le montre la figure V. 2, sur les 4 000 patients, 2681 appartiennent à la classe « 0 », ce qui indique qu'ils sont en bonne santé, tandis que 1349 développent le COVID-19.

Figure V. 3 montre la distribution de densité d'un ensemble de données de patients infectés et non infectés. On peut observer que les patients âgés de 55 à 90 ans constituent le groupe le plus touché selon l'ensemble de données appliqué. Le graphique suggère ainsi que l'âge est un facteur important pour le COVID-19 et que la probabilité de développer la maladie augmente avec l'âge.

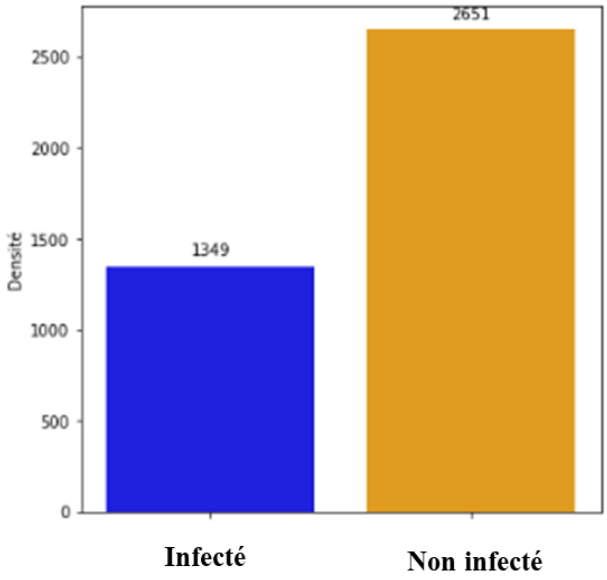


Figure V. 2 Analyse de la distribution des classes (infecté, non infecté) de COVID-19

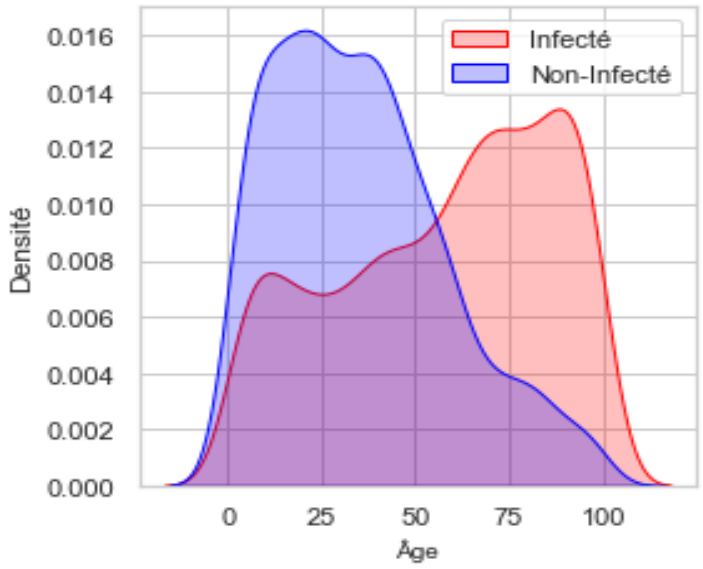


Figure V. 3 Répartition par âge des patients avec et sans COVID-19

La figure V. 4 montre comment les classes de différentes variables catégorielles sont liées aux maladies en fonction de la prévalence de la maladie dans l'ensemble de données. La plupart des personnes malades souffrent de maux de gorge, de fatigue, etc.

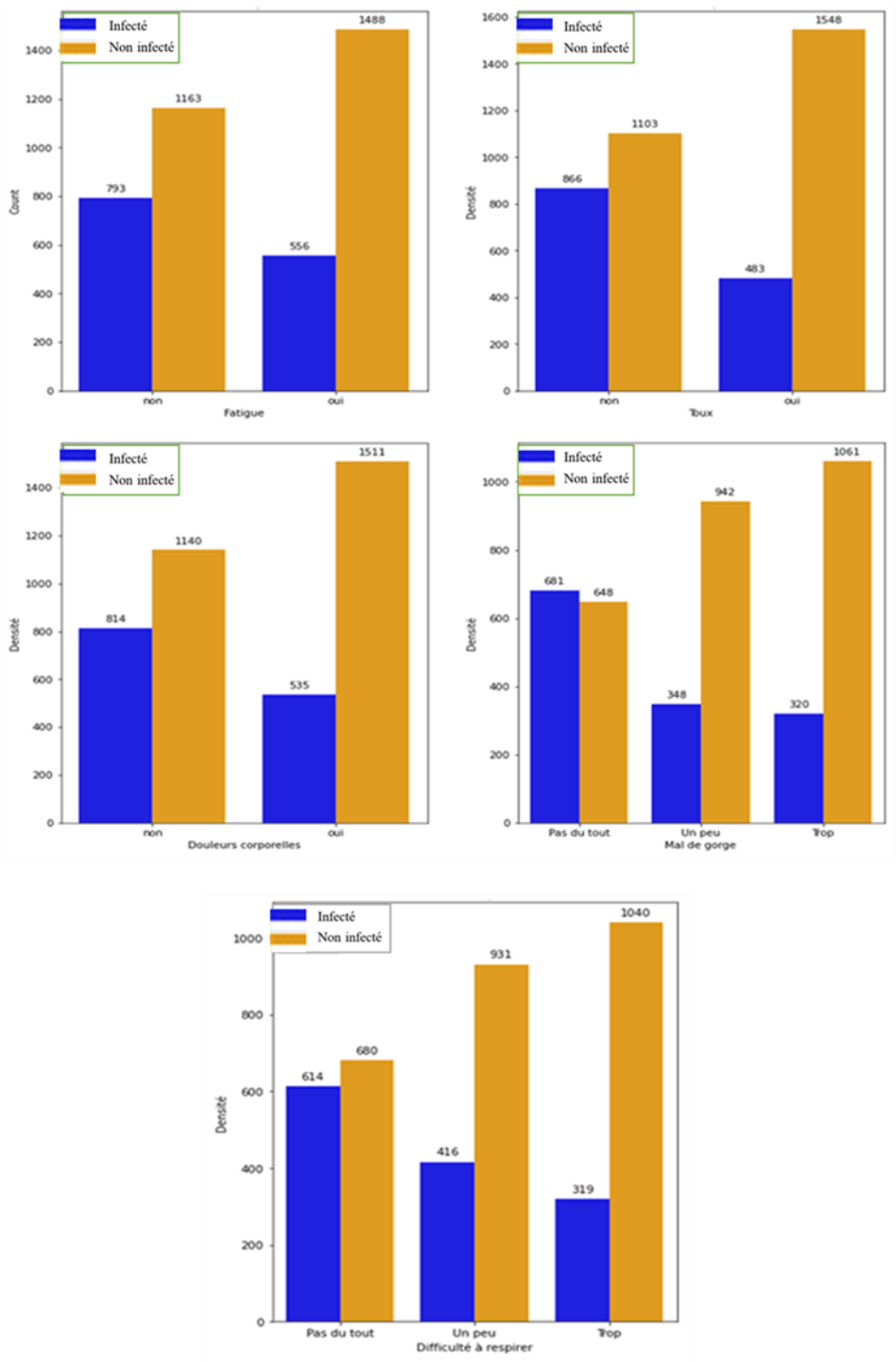


Figure V. 4 Relation de différentes variables catégorielles avec le COVID-19

V.2.1.2 Résultats du modèle proposé RNA- PMC

Les résultats de cette étude ont été obtenus en appliquant l'algorithme de classification RNA- PMC sur la base de données, visant à atteindre une exactitude élevée dans la prédiction de COVID19.

Choix de meilleur modèle

Diverses divisions et itérations ont été effectuées afin d'obtenir les meilleurs modèles RNA- PMC. Le choix du meilleur modèle repose sur la valeur maximale d'exactitude. Les résultats sont disponibles dans la figure V. 5. La meilleure exactitude obtenue est de 83.45 %. Cette performance a été réalisée en divisant la base de données comme suit :80% pour l'apprentissage, 10% pour le test et 10 % pour la validation (Voir la figure V. 6). Ce processus a été exécuté sur 1000 itérations.

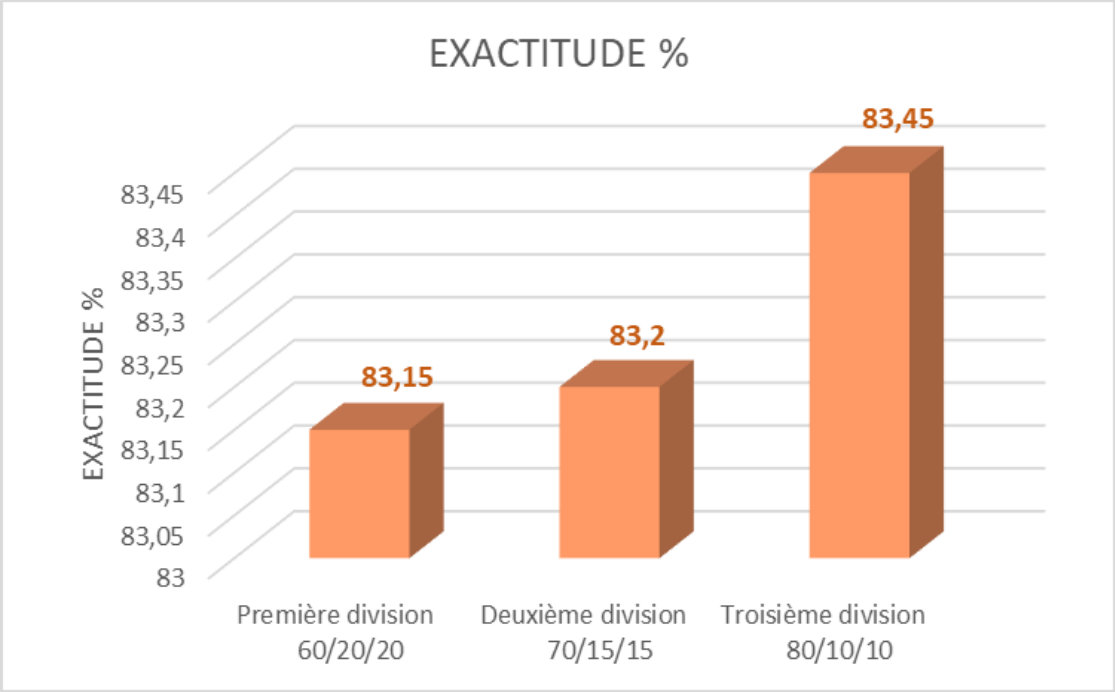


Figure V. 5 Présentation de l'exactitude pour chaque division de la base de données

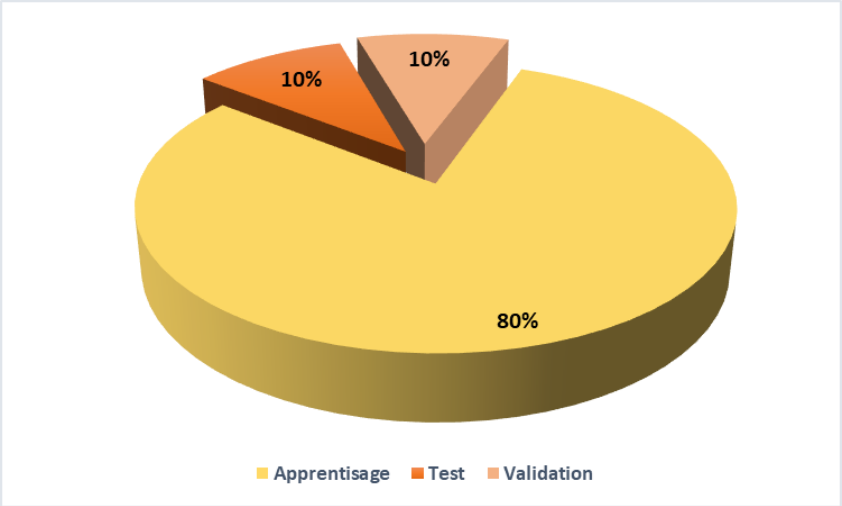


Figure V. 6 Représentation des meilleurs pourcentages de division de la base de données de COVID-19.

La structure des modèles RNA- PMC est présentée dans le tableau V.1, comprenant une couche d'entrée avec 7 neurones, une seule couche cachée avec 4 neurones, et une couche de sortie avec 2 neurones.

Tableau V.1 Architecture du meilleur modèle RNA-PMC pour la prédiction de COVID-19

Les Couches	**Couche d'entrée**	**Couche cachée**
Nombre de neurones	7	4
Fonction d'activation	/	Logistique
Formule mathématique	/	$a = \frac{1}{1+e^{-n}}$ (V.1)
Algorithme D'apprentissage	BFGS	

Matrice de confusion

À partir de la matrice de confusion présentée dans la figure V. 7, nous notons que notre modèle a correctement prédit 3 338 cas sur 4 000, incluant 925 cas de COVID-19 parmi les individus réellement infectés et 2 413 cas véritablement négatifs. Cependant, 662 cas ont été prédits de manière incorrecte. Parmi ceux-ci, 238 ont été classés comme positifs (FP), et 424 (FN) ont été classés comme des cas négative. Cette analyse souligne l'importance d'évaluer d'autres métriques telles que la sensibilité, la spécificité, la précision etc. pour obtenir une perspective complète de la performance du modèle dans la prédiction du COVID-19.



Figure V. 7 Matrice de confusion de la prédiction de Covid-19 par RNA- PMC

Evaluation des performances de modèle de classification

Le tableau V.2 présente les paramètres de performances de modèle appliquer. Tout d'abord, l'exactitude (Accuracy) de 83.45% indique que le modèle a réussi à classer correctement les instances dans une proportion significative. Cela suggère une performance globalement solide de modèle.

Une sensibilité de 68.56% indique que le modèle a réussi à détecter 68.56% des cas positifs de COVID-19, démontrant ainsi sa capacité à identifier avec succès les patients réellement infectés. Cela indique une sensibilité modérée du modèle

La spécificité élevée de 91.02% indique la capacité du modèle à exclure correctement les cas négatifs, ce qui est crucial pour la détection précise des cas positifs.

La précision de 79.53% indique la proportion de prédictions positives qui étaient réellement correctes, renforçant ainsi la confiance dans la capacité du modèle à fournir des résultats précis.

Le F-score de 73,64% représente une mesure équilibrée entre la précision et la sensibilité, illustrant ainsi la performance globale du modèle de manière équilibrée.

Le MCC de 62,04 % mesure la qualité globale des prédictions, prenant en compte les quatre éléments de la matrice de confusion. Un MCC élevé indique une forte corrélation entre les prédictions du modèle et les observations réelles, soulignant ainsi la fiabilité de la performance globale du modèle.

Tableau V.2 Performances du meilleur modèle de la prédiction de Covid-19 par RNA-PMC.

Les performances	Valeur
Ac %	83.45
Se %	68.56
Sp %	91.02
Pr %	79,53
F-score %	73,64
MCC %	62,04

La figure V. 8 illustre la courbe ROC pour le modèle RNA-PMC appliquée. L'AUC-ROC de 88.24% est une indication de la capacité du modèle à discriminer entre les classes positives et négatives, avec une valeur élevée indiquant une bonne performance.

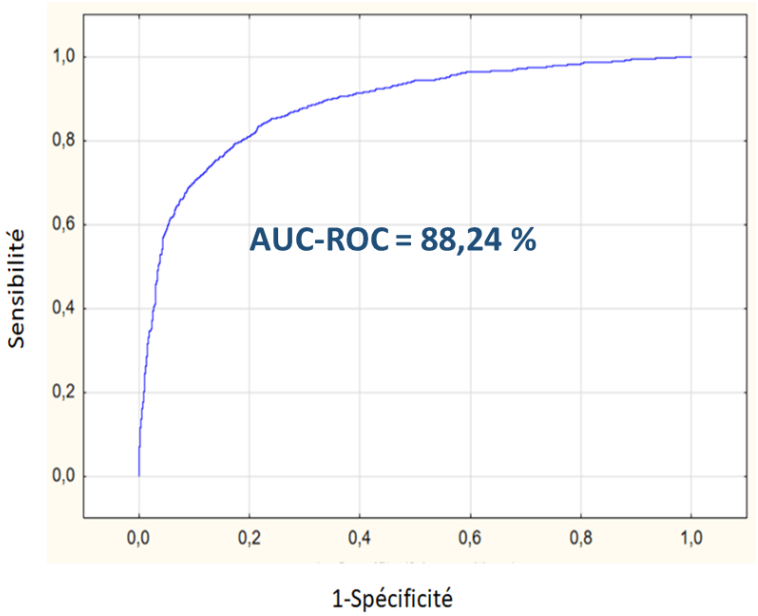


Figure V. 8 Courbe ROC pour la prédiction de COVID-19 par RNA-PMC

V.2.1.3 Analyse de sensibilité par méthode de poids

L'analyse de sensibilité de chaque entrée sur la sortie dans notre étude donne les résultats qui sont représentés sur la figure V. 9. Les poids du modèle apparaissent dans le tableau B.1 (annexe B).

Les résultats de l'analyse de sensibilité mettent en évidence que la difficulté à respirer est le facteur le plus influent, avec une importance relative de 32,02 %, suivi de près par l'âge et le mal de gorge, qui présentent une importance relative de 21,22 % et 18,22 %, respectivement. Ces observations suggèrent que l'âge exerce un impact significatif sur les prédictions du modèle, ce qui est cohérent avec les connaissances épidémiologiques indiquant que l'âge peut jouer un rôle déterminant dans la sévérité des symptômes de COVID-19.

D'un autre côté, certaines caractéristiques ont des contributions presque similaires, telles que les douleurs corporelles, la toux, la fatigue et la température corporelle. Cela signifie que ces facteurs ont une influence comparable sur le modèle, mais leur importance relative est moins marquée par rapport aux principaux facteurs identifiés. Cette similitude dans leurs contributions n'indique qu'aucun de ces facteurs n'émerge comme particulièrement dominant, mais qu'ils restent néanmoins pertinents pour l'analyse globale.

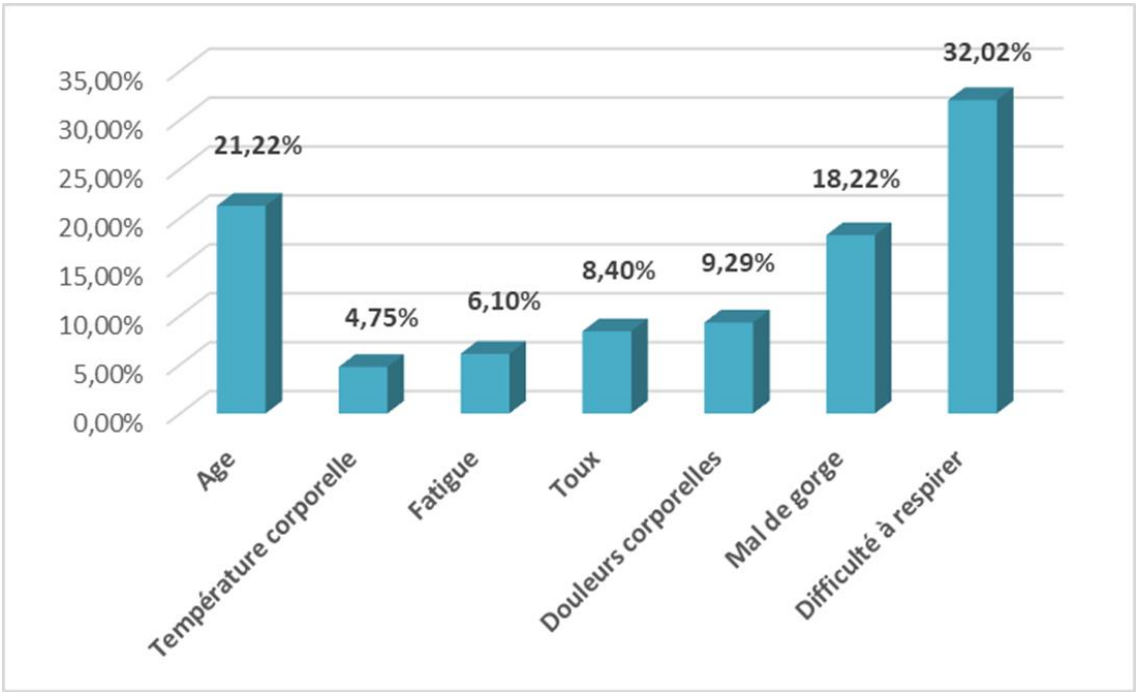


Figure V. 9 Histogramme d'importance relative pour le COVID-19

V.2.2 Contribution II : prédiction de l'hépatite virale par RNA- PMC

V.2.2.1 Résultats de l'analyse exploratoire des données

- **Corrélation entre les variables**

À partir de la matrice de corrélation illustrée dans la figure V.10, on constate une très faible corrélation entre les attributs. Les valeurs des coefficients de corrélation se situent dans l'intervalle [−0.55, 0.65]. Nous pouvons conclure que les attributs de notre ensemble de données ne sont pas corrélés.

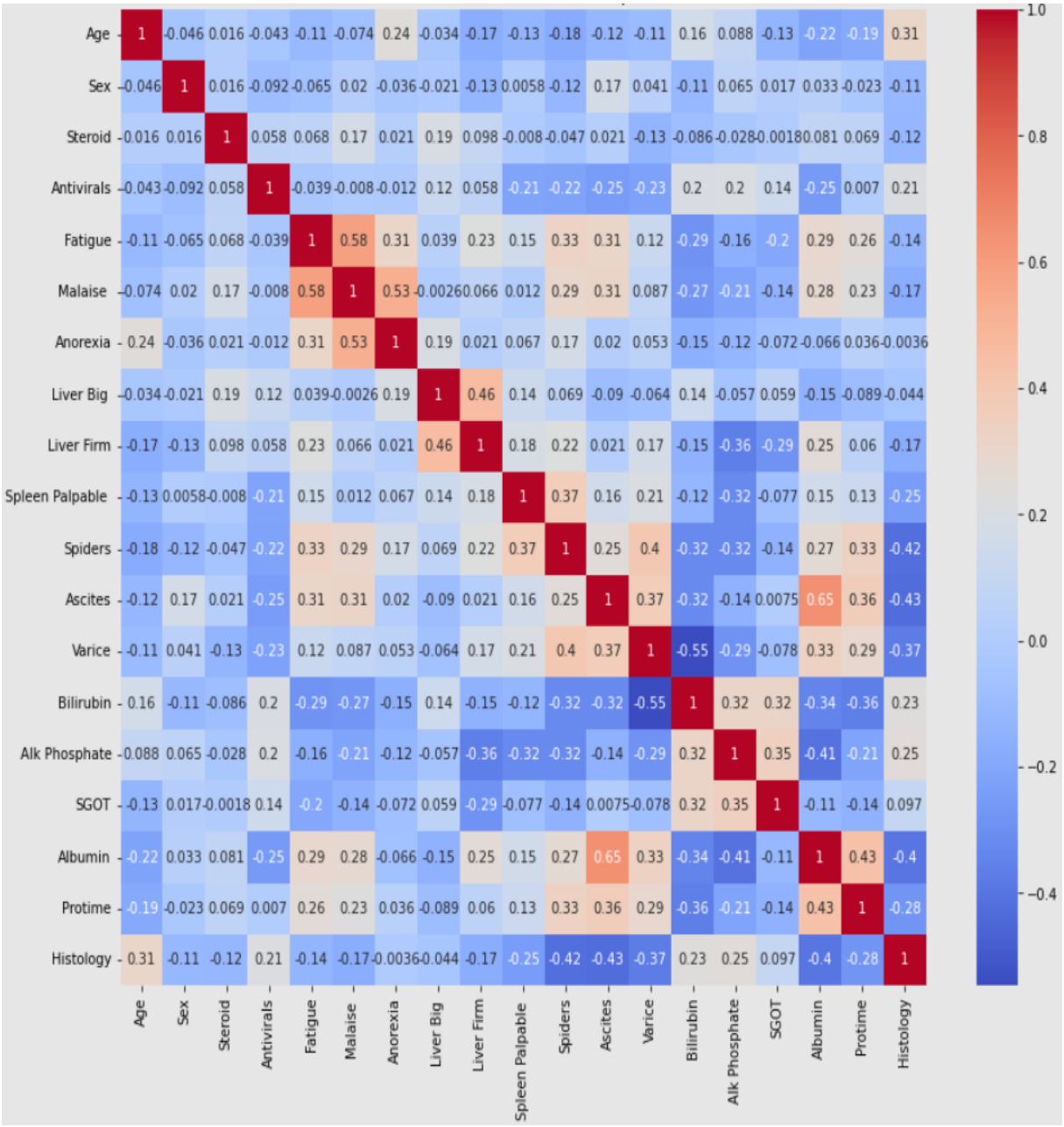


Figure V.10 Matrice de corrélation de la prédiction de l'hépatite virale par RNA-PMC

Dans l'analyse de cet ensemble de données, il est essentiel de comprendre la répartition des patients dans chaque classe. Comme le montre la figure V. 11, sur les 80 patients, 67 appartiennent à la classe « Non (2) », ce qui indique qu'ils sont en bonne santé, tandis que 13 appartiennent à la classe « Oui (1) », ce qui indique qu'ils ont l'hépatite.

La figure V.12 montre la distribution de densité d'un ensemble de données de patients malades et non malades. On peut observer que les patients âgés de 40 à 55 ans constituent le groupe le plus touché selon l'ensemble de données appliqué. Le graphique suggère ainsi que l'âge est un facteur important pour l'hépatite et que la probabilité de développer la maladie augmente avec l'âge.

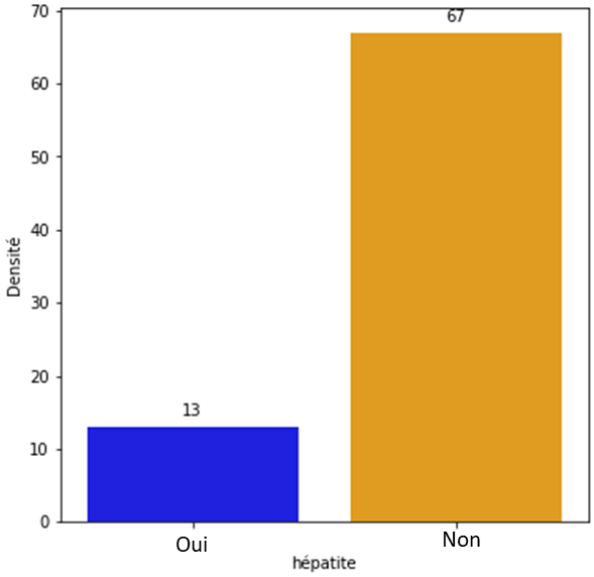


Figure V. 11 Analyse de la distribution des classes (oui, non) d'hépatite virale

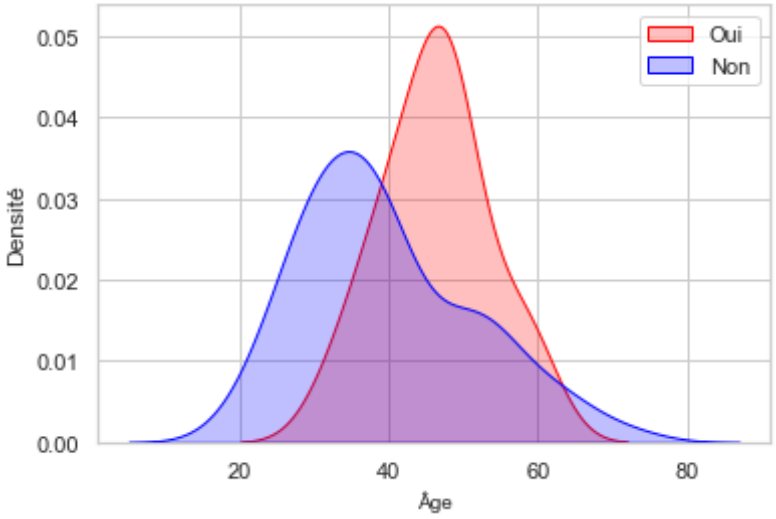


Figure V.12 Répartition par âge des deux classes de patients atteints d'hépatite virale

La figure V.13 montre comment les différentes classes de sexe sont liées à l'hépatite virale en fonction de la prévalence de cette maladie dans l'ensemble de données. Parmi les individus en mauvaise santé, 13 sont des hommes, tandis qu'aucun n'est une femme. Cela suggère une disparité potentielle entre les sexes dans la prévalence de l'hépatite au sein de l'ensemble de données.

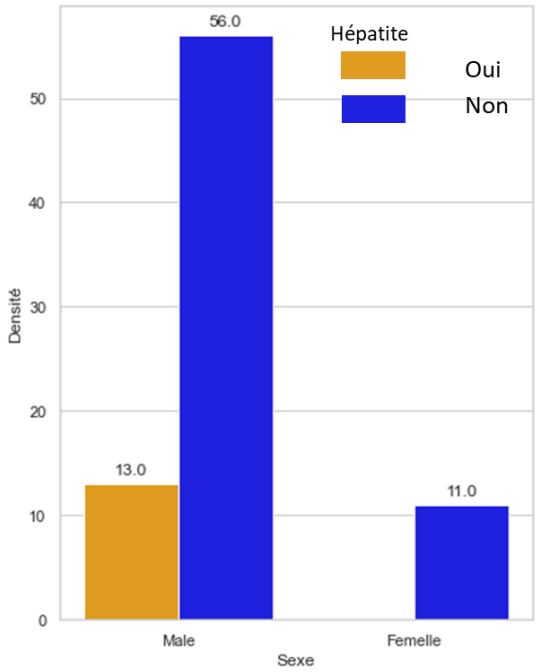


Figure V.13 Répartition par sexe des deux classes de patients atteints d'hépatite virale

V.2.2.2 Résultats du modèle proposé RNA -PMC

Les résultats de cette étude ont été obtenus en appliquant l'algorithme de classification RNA -PMC sur la base de données d'hépatite virale, visant à atteindre une précision élevée dans la prédiction d'hépatite virale.

Le choix de meilleur modèle

Diverses divisions et itérations ont été effectuées afin d'obtenir les meilleurs modèles RNA -PMC, les résultats sont disponibles dans l'annexe A, et la figure V.14. La meilleure exactitude obtenue est de 97,5 %. Cette performance a été réalisée en divisant la base de données comme suit :80% pour l'apprentissage, 10% pour le test, et 10% pour la validation (voir la figure V. 15). Ce processus a été exécuté sur 1000 itérations.

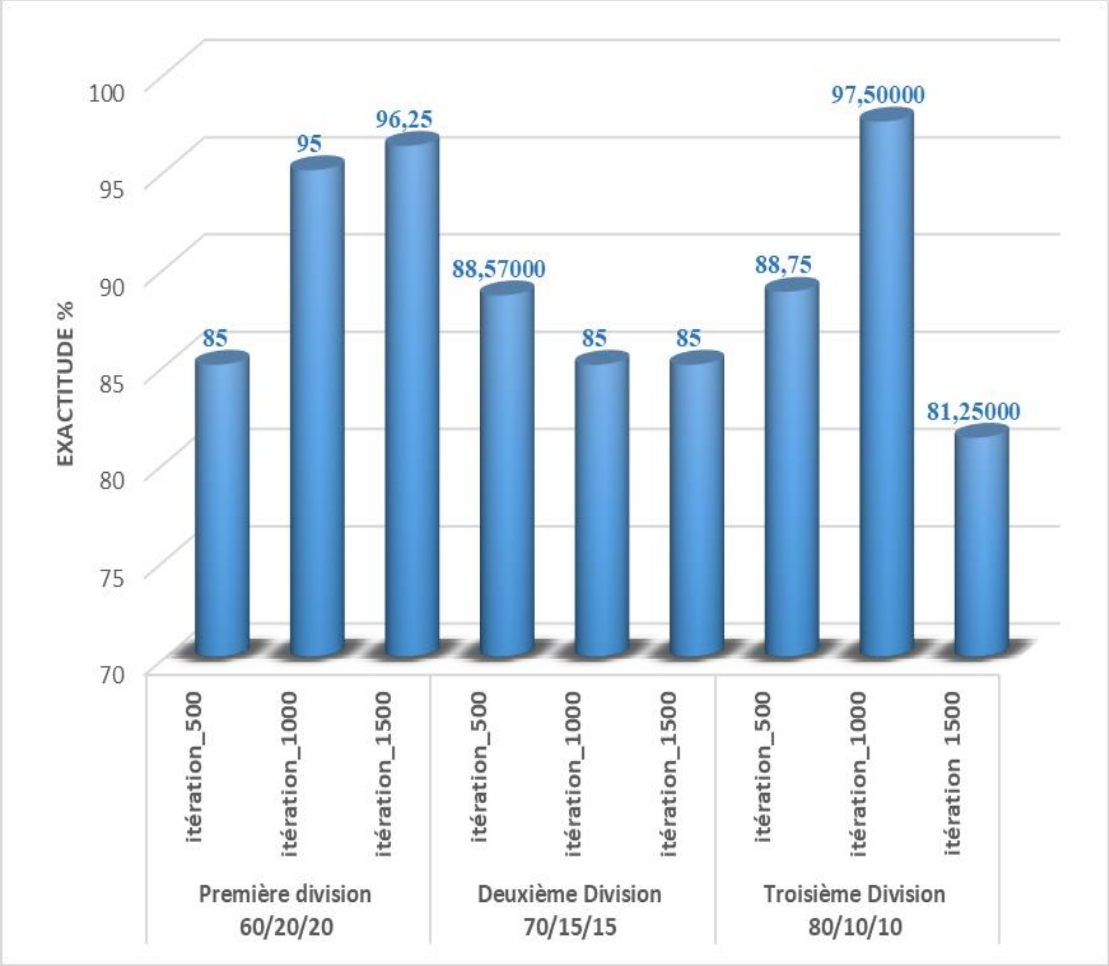


Figure V. 14 Présentation de l'exactitude pour chaque division de la base de données

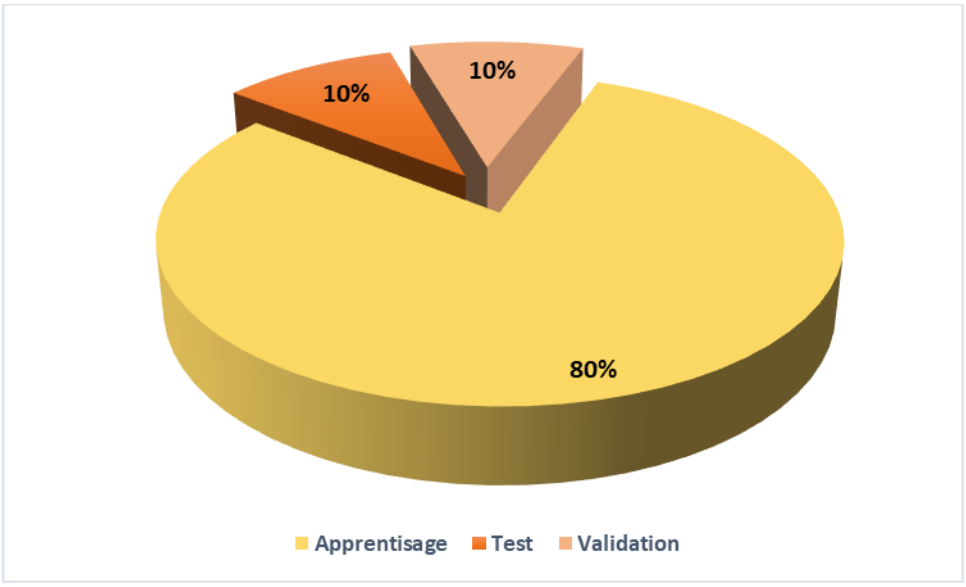


Figure V. 15 Présentation des meilleurs pourcentages de division de la base de données.

La structure des modèles RNA-PMC est présentée dans le tableau V.3, comprenant une couche d'entrée avec 19 neurones, une seule couche cachée avec 3 neurones, et une couche de sortie avec 2 neurones. L'algorithme BFGS a été sélectionné pour l'apprentissage. La fonction d'activation utilisée est Identity pour la couche cachée.

Tableau V.3 Architecture du meilleur modèle RNA-PMC pour la prédiction de l'hépatite virale

Les Couches	**Couche d'entrée**	**Couche cachée**
Nombre de neurones	19	3
Fonction d'activation	/	Identity
Formule mathématique	/	a = n V.1
Algorithme de formation	BFGS	

Matrice de confusion

À partir de la matrice de confusion présentée dans la figure V.16, nous observons que le modèle RNA-PMC a démontré une performance globalement excellente. Sur un total de 80 instances, le modèle a réussi à prédire correctement 78 d'entre elles. Plus précisément, il a correctement identifié les 12 cas d'hépatite (VP) parmi les individus réellement infectés, illustrant ainsi une forte sensibilité. De plus, le modèle a correctement classé 66 instances (VN) comme n'étant pas infectées par l'hépatite, montrant une bonne spécificité. Cependant, une instance a été mal prédite (FN=1), signalant un cas d'hépatite non détecté, et un faux positif (FP) a également été observé, ce qui signifie que le modèle a prédit incorrectement la présence d'hépatite dans un cas où la maladie n'était pas présente.



Figure V.16 Matrice de confusion pour la prédiction de l'hépatite virale par RNA-PMC

Evaluation des performances de modèle de classification

À partir des résultats du tableau V. 4, il est clair que les performances du modèle PMC-RNA dans la prédiction de l'hépatite sont extrêmement prometteuses. Avec une exactitude de 97.5%, le modèle a réussi à classer correctement la grande majorité des instances, démontrant ainsi son efficacité globale. Une sensibilité de 92.30% indique la capacité du modèle à identifier correctement 92.30% des cas réels d'hépatite parmi les individus infectés. De plus, une spécificité de 98.5% confirme la capacité du modèle à minimiser les faux positifs, assurant ainsi une fiabilité élevée dans la détection des cas négatifs d'hépatite. La précision de 92.3% souligne la capacité du modèle à classifier correctement environ 92.3% de toutes les instances, que ce soient les cas positifs ou négatifs. Un F-score de 92.3% offre un équilibre entre la précision et la sensibilité, démontrant une performance globale solide du modèle. Le MCC de 90.81% indique une forte corrélation entre les prédictions du modèle et les observations réelles, une mesure robuste tenant compte des quatre éléments de la matrice de confusion. Enfin, une aire sous la courbe AUC-ROC de 95.12 suggère une excellente capacité du modèle à discriminer entre les classes positives et négatives (voir la figure V.17).

Ces résultats globalement élevés indiquent que le modèle RNA-PMC est prometteur pour la prédiction de l'hépatite virales.

Tableau V. 4 Performances du meilleur modèle de la prédiction de l'hépatite virale par RNA-PMC

Mesures de performances	Valeur
Ac %	97.5
Se %	92.30
Sp %	98.5
Pr %	92.3
F-score %	92.3
MCC %	90.81

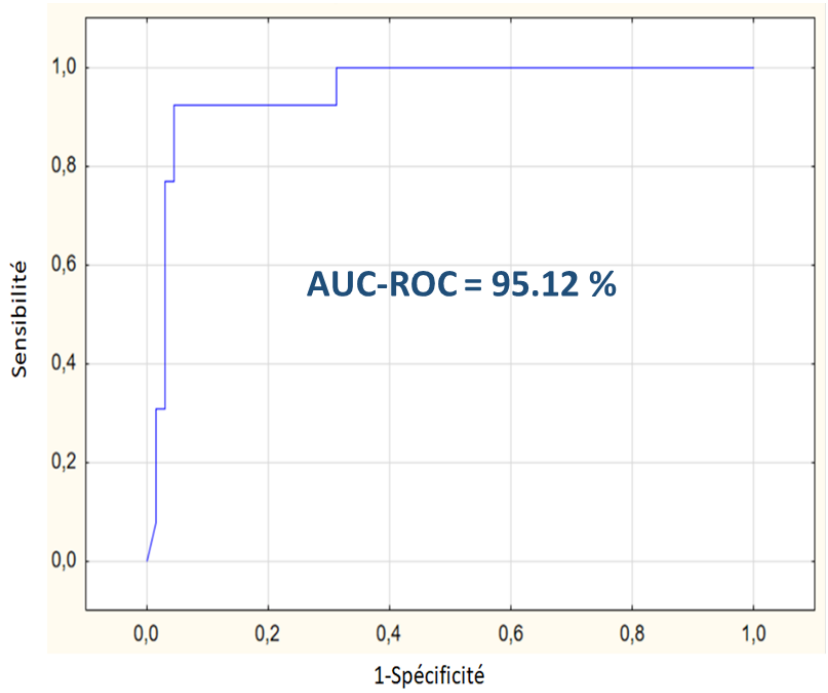


Figure V.17 Courbe ROC pour la prédiction de l'hépatite virale par modèle RNA-PMC

V.2.2.3 Analyse de sensibilité par méthode de poids

Nous avons réalisé une analyse de sensibilité en utilisant la méthode du « poids » afin d'évaluer l'impact de chaque attribut d'entrée sur la sortie du modèle. Les résultats, présentés de manière graphique dans la figure V.18, offrent des perspectives essentielles concernant l'influence

relative de chaque paramètre dans la prédiction de l'hépatite virale. Les poids du modèle apparaissent dans l'annexe B.2.

Parmi les attributs étudiés, le SGOT émerge comme le facteur le plus significatif avec une importance relative de 12,83%, suivi par anorexie à 9,34% et la Bilirubine à 8,12%. Ces résultats indiquent que ces trois variables jouent un rôle crucial dans la prédiction de l'hépatite, suggérant une forte corrélation entre ces indicateurs et la maladie hépatique.

D'autres facteurs tels que le malaise, l'ascite, la Rate palpable, la prise d'antiviraux, le phosphate alcalin, la fatigue, le foie gras, les stéroïdes, le Protime, l'histologie, l'âge, les varices, l'albumine, la Ferme du foie, la présence d'araignées vasculaires, et le sexe, montrent des contributions variables, chacun ayant un impact distinct sur la prédiction de l'hépatite virale.

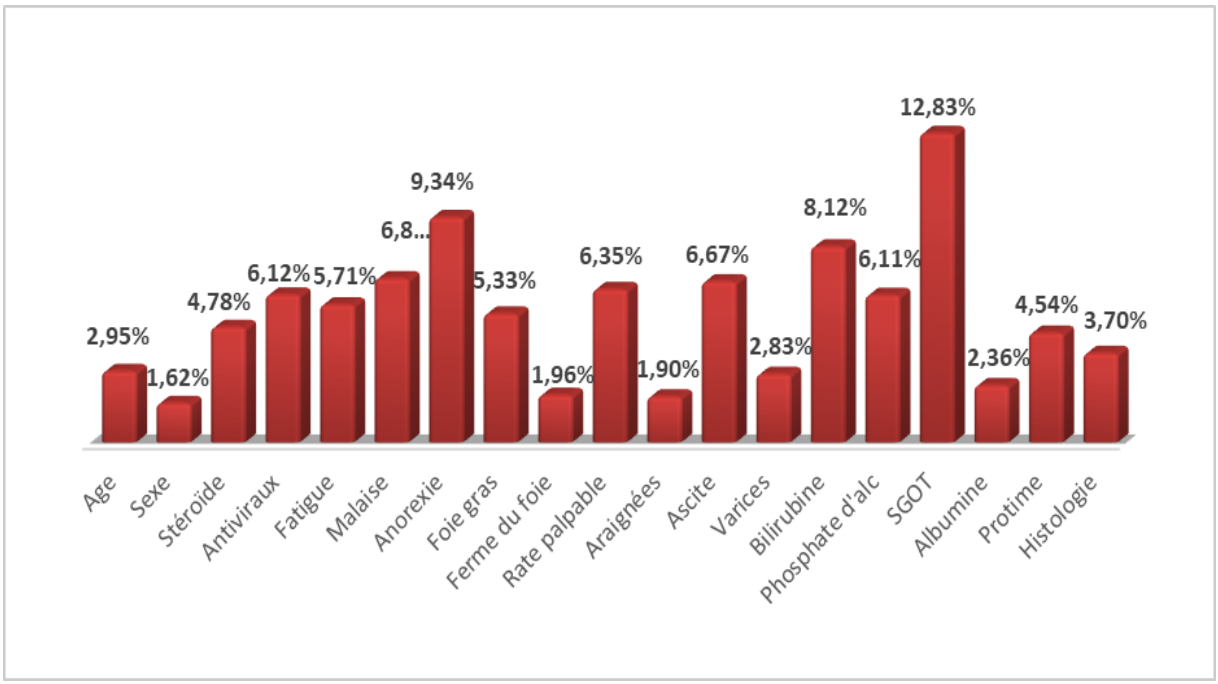


Figure V.18 Histogramme d'importance relative pour la prédiction de l'hépatite virale par RNA-PMC

V.2.2.4 Analyse comparative entre notre modèle et les modèles existants dans la littérature.

Plusieurs études de haute qualité ont été menées pour la détection et la prédiction de l'hépatite. Nous avons comparé notre travail avec d'autres recherches dans le domaine de prédiction de l'hépatite utilisant l'IA afin de confirmer l'efficacité de notre approche. Les résultats de cette comparaison sont clairement exposés dans le tableau V. 5.

Tableau V. 5. Comparaison de la prédiction de l'hépatite virale avec d'autres études de littérature

Auteur	Nombre de données dans la base de données	Modèle	Ac %
Yarasuri et al. 2019 [119]	155	RNA	96
Roslina and Noraziah 2010 [137]	155	SVM	74.55
Polat and Güneş 2007 [138]	155	PCA-AIRS	94.12
Çalişir and Dogantekin 2011 [139]	155	PCA–LSSVM	96.12
Modèle proposé	155	**PMC-RNA**	**97.5**

Les résultats de comparaison entre notre modèle PMC-RNA et les modèles existants dans la littérature pour la prédiction de l'hépatite virale révèlent des performances prometteuses. Notre modèle PMC-RNA a atteint un taux d'exactitude remarquable de 98.75%, dépassant significativement les autres approches examinées. En particulier, PCA–LSSVM et RNA ont montré des performances honorables avec des taux d'exactitude de 96.12% et 96%, respectivement.

La robustesse de notre modèle peut être attribuée à son architecture avancée basée sur les réseaux neuronaux artificiels, ce qui lui permet d'extraire des motifs complexes à partir des données. En comparaison, les approches telles que SVM et ACP - AIRS ont affiché des taux d'exactitude inférieurs à notre modèle, suggérant que notre méthode pourrait être plus efficace dans la capture de la variabilité des données liées aux l'hépatite virale.

V.3 Prédiction des affections non transmissibles

V.3.1 Contribution III : Prédiction de cancer du poumon par CSABC-SVM

V.3.1.1 Résultat de l'analyse exploratoire des données

À partir de la matrice de corrélation illustrée dans la figure. V.19, on constate une corrélation très faible entre les attributs. Les valeurs des coefficients de corrélation se situent dans l'intervalle [-0,29, 0,57]. Nous pouvons conclure que les attributs de notre ensemble de données ne sont pas fortement corrélés.

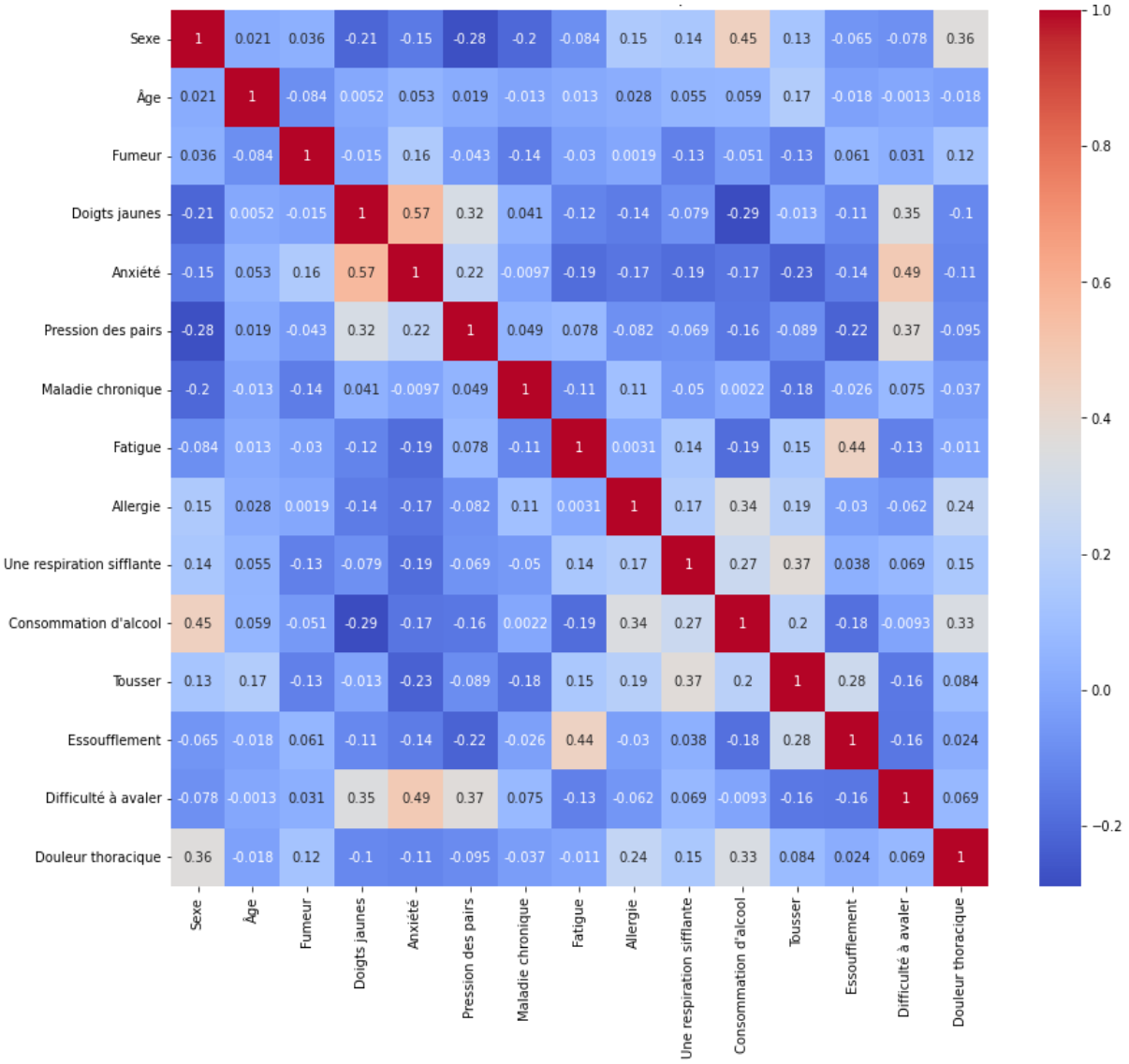


Figure. V.19 Matrice de corrélation montrant les corrélations entre tous les attributs de l'ensemble de données.

Dans l'analyse de cet ensemble de données, il est essentiel de comprendre la répartition des patients dans chaque classe. La figure V.20 montre que sur les 309 patients, 39 sont en bonne santé, tandis que 270 présentent le cancer du poumon. Cette répartition permet d'obtenir un aperçu significatif de la distribution des classes au sein de l'ensemble de données

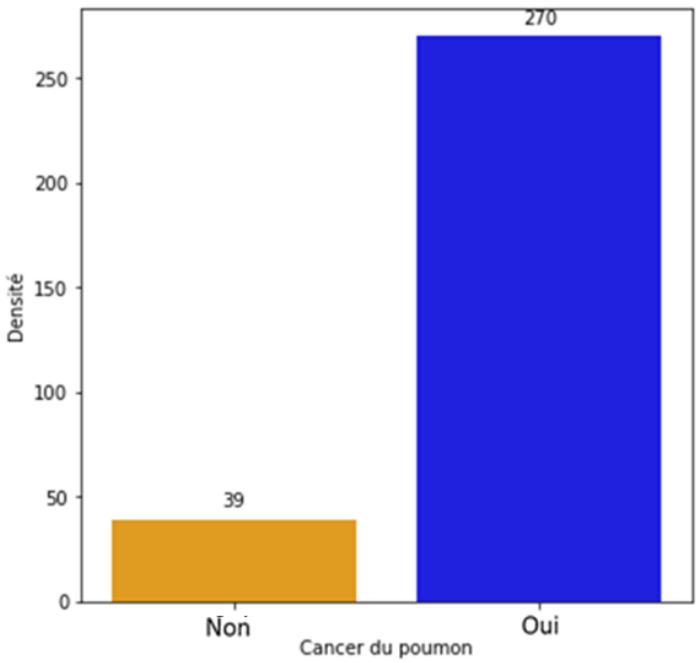


Figure. V. 20 Analyse de la distribution des classes (oui, non) de cancer du poumon

La figure. V. 21 illustre la distribution de densité d'un ensemble de données sur le cancer du poumon, en distinguant les patients atteints de la maladie de ceux qui ne le sont pas. Notamment, la tranche d'âge de 50 à 70 ans ressort comme la plus affectée, avec un pic d'incidence observé à l'âge de 60 ans.

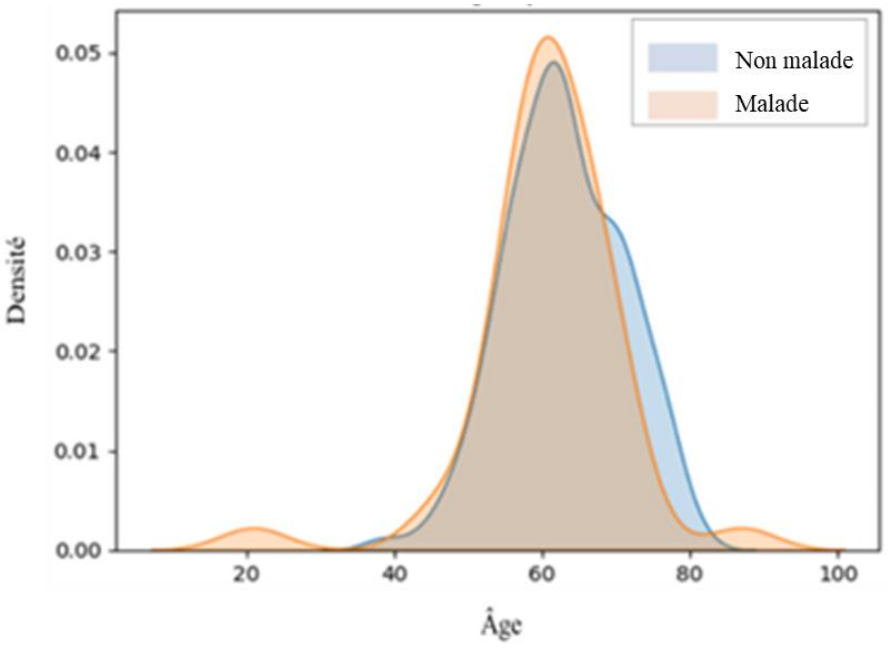


Figure. V. 21 Graphiques KDE pour les patients avec et sans cancer du poumon en fonction de la répartition d'âge

La figure V. 22 présente une analyse détaillée des classes de différentes variables catégorielles en relation avec le cancer du poumon, en se basant sur la prévalence de la maladie dans l'ensemble de données. Les données révèlent que parmi les personnes en mauvaise santé, 145 appartiennent à la population masculine tandis que 125 sont de sexe féminin. En ce qui concerne le tabagisme, 155 individus sont fumeurs tandis que 115 ne fument pas. De même, 165 personnes sont consommatrices d'alcool, tandis que 105 ne consomment pas d'alcool.

Il est intéressant de noter que la majorité des individus en mauvaise santé présentent des symptômes tels que des douleurs thoraciques, soulignant ainsi des corrélations potentielles entre ces symptômes et le cancer du poumon. Ces informations fournissent un aperçu précieux pour mieux comprendre la distribution des caractéristiques liées à la maladie dans la population étudiée. Il est crucial de prendre en compte ces associations dans l'analyse globale des facteurs de risque du cancer du poumon et d'approfondir la compréhension des relations entre les variables catégorielles examinées.

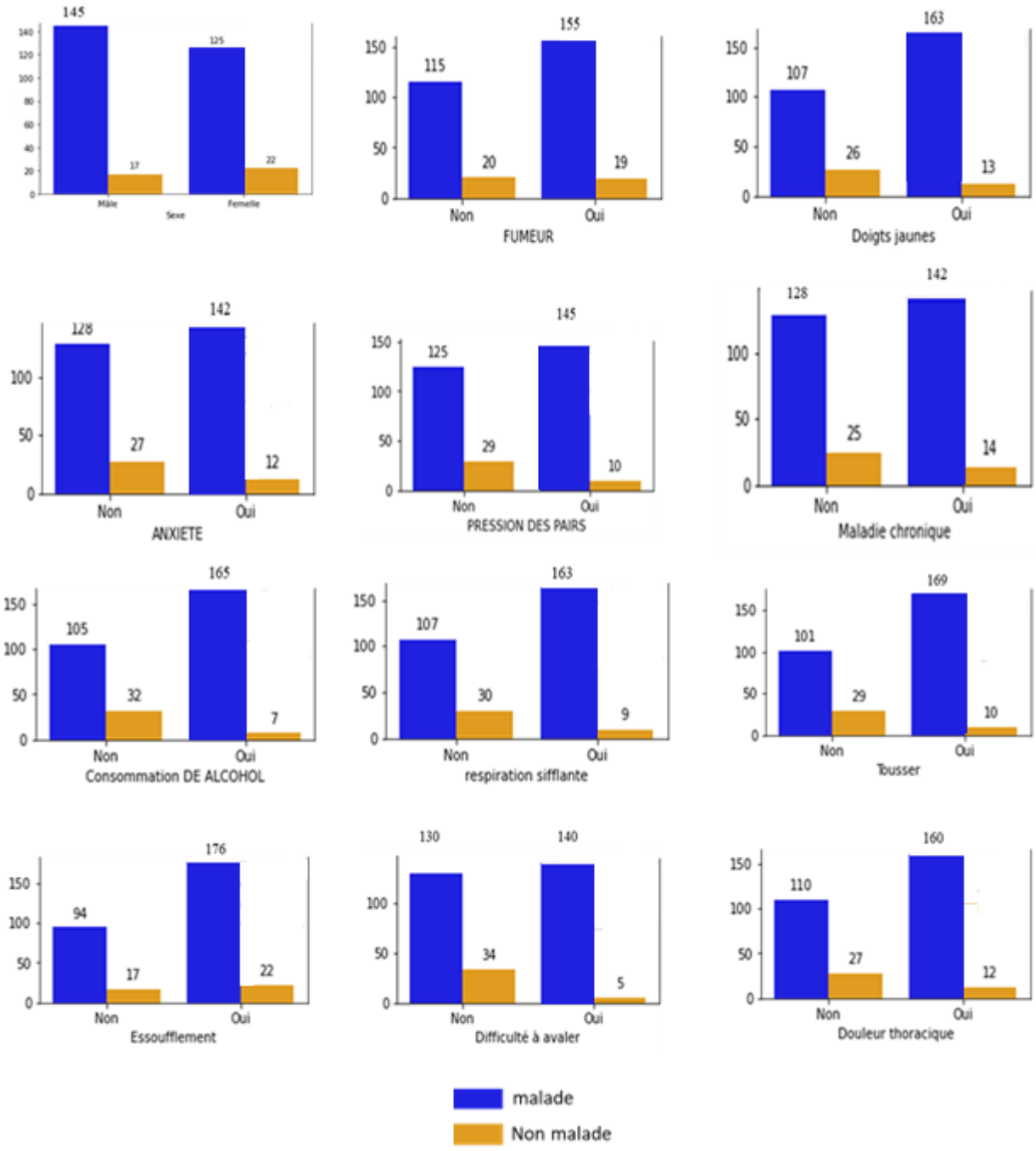


Figure. V. 22 Analyse de la relation entre différentes variables catégorielles et la variable cancer de poumon.

V.3.1.2 Résultats des modèles proposés

Nous avons évalué l'efficacité des différents algorithmes appliqués par diverses mesures statistiques telles que l'exactitude, la sensibilité, la spécificité, la précision, le F-score, le MCC et la courbe ROC. Notre étude a comporté une analyse comparative approfondie de ces algorithmes en se basant sur les mesures d'évaluation des performances. La figure V. 23 présente l'exactitude

des algorithmes de classification utilisés et le tableau V. 6 présente les paramètres de performance des algorithmes de classification utilisés.

Parmi les algorithmes testés, le SVM s'est distingué comme l'algorithme le plus performant, démontrant un équilibre exceptionnel entre précision, sensibilité et spécificité. Il a atteint la plus haute exactitude, s'élevant jusqu'à 96,80 %, surpassant tous les autres algorithmes évalués. En examinant différentes mesures de performance, notamment une sensibilité de 98,33 %, et une F-score de 98,33 %. Cette performance globale positionne le SVM comme le choix optimal pour la prédiction de cancer du poumon.

Tableau V. 6 Performances des modèles de classification pour la prédiction de cancer du poumon.

Mesures de performances	**RNA**	**KNN**	**SVM**
Ac %	95.18	95.21	**96.80**
Se %	93.33	96.66	**98.33**
Sp %	100	50	50
Rappel %	93.33	96.66	**98.33**
Pr %	100	98.30	98.33
F-score %	96.55	97.47	**98,33**
MCC %	55.77	38.42	48.33

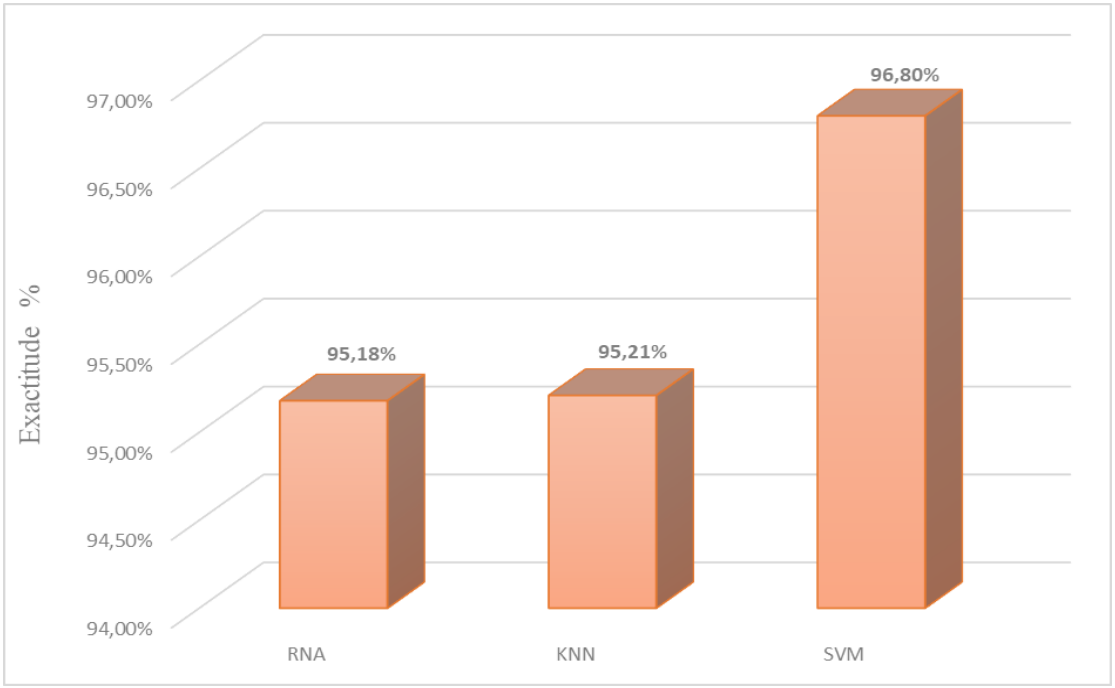


Figure V. 23 Comparaison de l'exactitude entre les trois modèles (RNA, KNN et SVM)

La figure 24, la figure 25, et la figure 26 présentent les courbes ROC pour les modèles RNA, KNN, et SVM. Ces courbes fournissent une visualisation graphique de la performance de chaque modèle. En examinant ces courbes, nous pouvons confirmer les résultats mentionnés précédemment. L'aire sous la courbe AUC-ROC de 0,97, 0,67, et 0,93 pour les modèles RNA, KNN, et SVM, respectivement, indique une excellente capacité de discrimination entre les classes positives et négatives. En général, une AUC-ROC proche de 1 indique une grande capacité du modèle à discriminer entre les deux classes, tandis qu'une valeur proche de 0,5 indique des performances de discrimination limitées. Par conséquent, les valeurs que nous avons obtenues indiquent de fortes performances discriminantes de nos modèles, renforçant la confiance dans leur capacité à fournir des prédictions précises. Il s'agit d'une mesure importante lors de l'évaluation des modèles de classification dans des contextes médicaux où la distinction entre les catégories est cruciale.

- Cette surperformance peut s'expliquer algorithmiquement comme suit :

Les machines à vecteurs de support et les réseaux neuronaux artificiels ont leurs points forts et leurs faiblesses, les SVM peuvent être considérées comme meilleures que les RNA lorsqu'elles sont appliquées à de petits ensembles de données [76].

Le modèle SVM peut gérer des ensembles de données avec un grand nombre d'attributs, ce qui est souvent le cas dans des domaines tels que l'apprentissage automatique appliqué aux données

réelles. En revanche, KNN attribue une importance égale à tous les attributs, ce qui peut ne pas être optimal dans des espaces de grande dimension.

Le modèle SVM utilise un noyau pour mesurer la similarité entre les données, ce qui lui permet d'accorder plus d'importance aux attributs importantes et de réduire l'influence des attributs moins pertinentes. Cela permet une meilleure séparation des classes

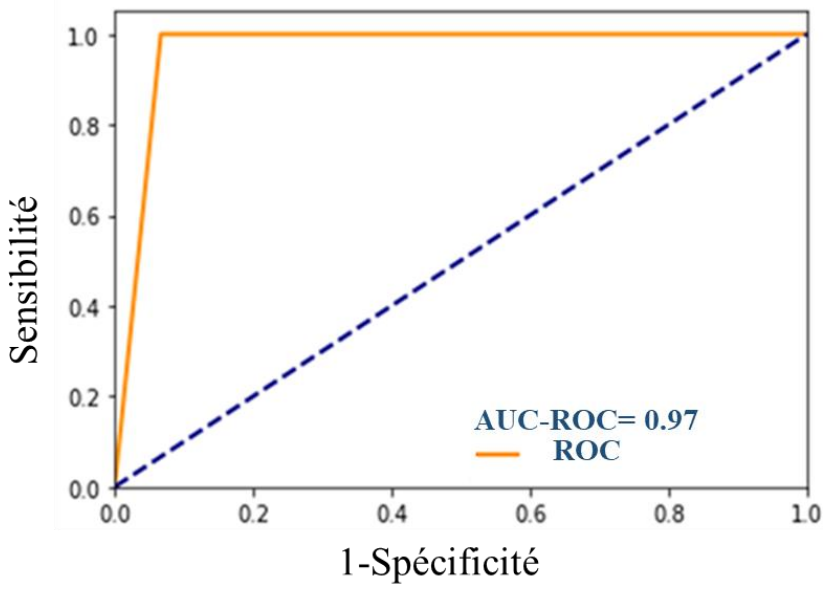


Figure V. 24 Courbe ROC via RNA

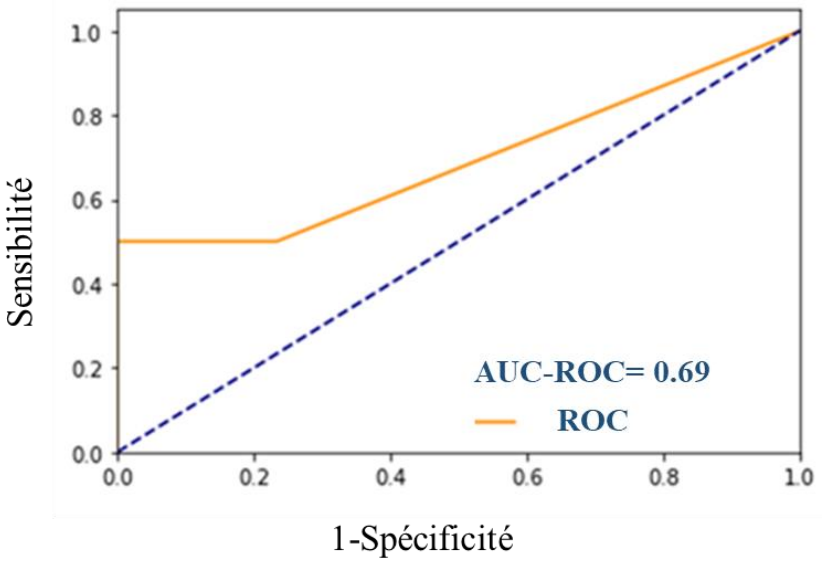


Figure V. 25 Courbe ROC via KNN

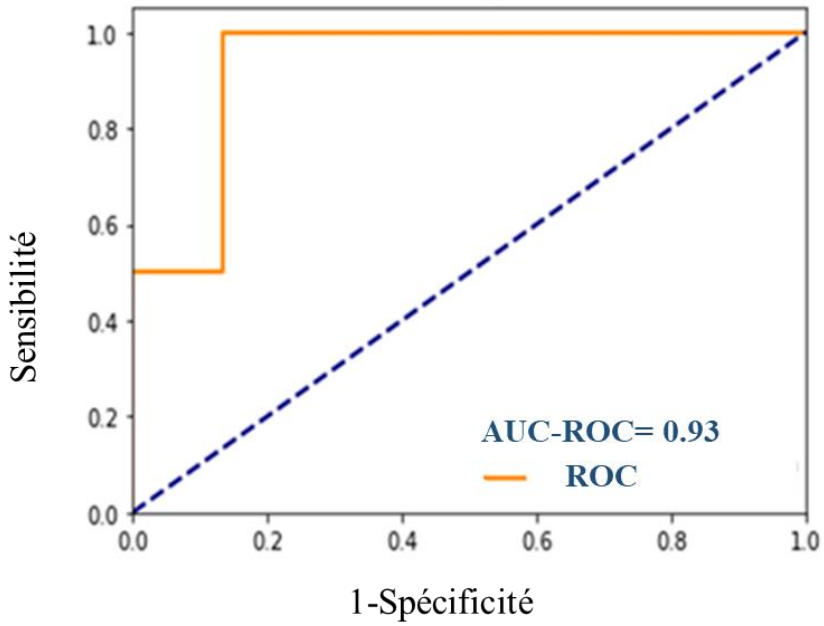


Figure V. 26 Courbe ROC via SVM

V.3.1.3 Résultats d'optimisation

Les performances du modèle hybride CSABC-SVM ont été évaluées à l'aide de mesures telles que l'exactitude, la sensibilité, la spécificité, la précision, le rappel, le F-score et le MCC, comme le montre le tableau 7. La figure V.27 présente une comparaison des performances des modèle SVM et CSABC-SVM.

Tableau V. 7. Performances de prédiction de cancer du poumon par CSABC-SVM

Mesures de performances	CSABC-SVM
Ac %	98,38
Se %	100
Sp %	50
Pr %	98.36
F-score %	99.17
MCC %	70.12

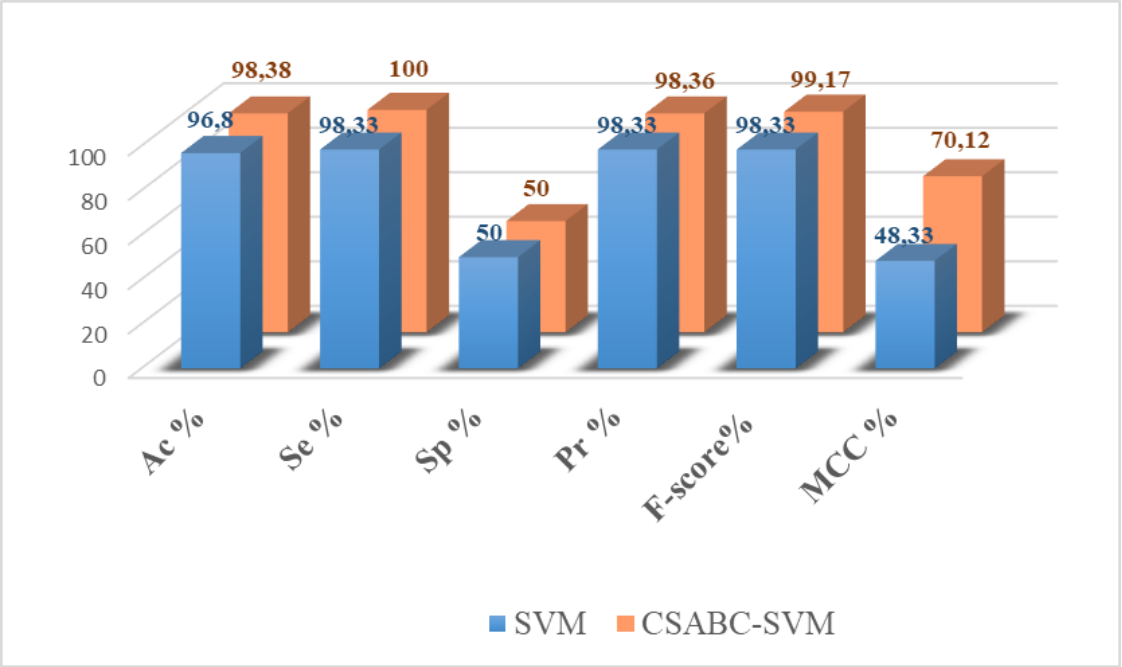


Figure V. 27 Comparaison des performances de SVM et CSABC-SVM pour la prédiction de cancer du poumon

Le modèle CSABC-SVM a affiché un taux d'exactitude remarquable de 98,38 %, représentant une amélioration substantielle de 1,58 % par rapport au modèle SVM, comme illustré dans la figure V. 27. Le modèle a réussi à classer correctement la grande majorité des instances. Cette amélioration significative des performances de prédiction souligne l'efficacité du modèle obtenue grâce à l'hybridation des algorithmes CSA et ABC.

La sensibilité atteint 100 %, ce qui indique que le modèle CSABC-SVM identifie de manière exhaustive tous les cas positifs de cancer du poumon.

Le F-score de 99.17% offre un équilibre entre la précision et la sensibilité, démontrant une performance globale solide du modèle. Le MCC de 70.12% indique une forte corrélation entre les prédictions du modèle et les observations réelles.

Grâce à la courbe ROC dans la figure V. 28, nous pouvons confirmer les résultats mentionnés précédemment, car l'aire sous la courbe AUC-ROC du modèle CSABC-SVM indique une bonne capacité de discrimination entre les classes.

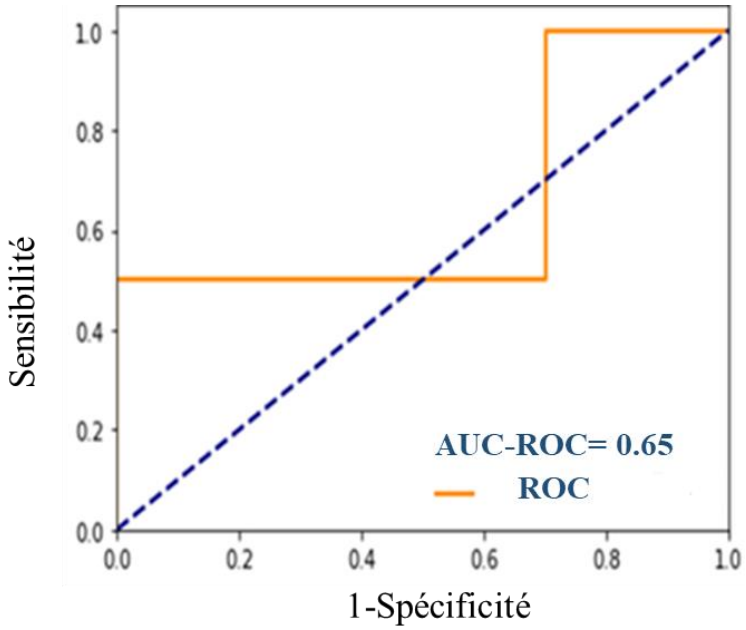


Figure V. 28 Courbe ROC via CSABC-SVM

La structure finale du modèle CSABC- SVM qui a donné les meilleures performances est présentée dans le tableau V. 8

Tableau V. 8 Paramètres de modèle CSABC- SVM via la prédiction de cancer du poumon.

Paramètres	**Valeur**
C	0.9540
Kernel	RBF
γ	0.8609

V.3.1.4 Importance des variables

Les contributions de chaque attribut sont illustrées dans la figure V. 29. L'importance relative des attributs sélectionnés a été calculée à l'aide de la méthode l'importance des fonctionnalités de permutation [141]. La figure présente efficacement les attributs les plus influentes responsables du cancer du poumon. À partir de cette figure, nous pouvons identifier les sept attributs les plus significatives en fonction de leur importance relative. Il est évident que l'âge est une caractéristique ou un facteur important pour l'identification et la prédiction du cancer du poumon, suivi de la toux, de la consommation d'alcool, de l'anxiété, tandis que les autres attributs présentent des degrés d'importance presque similaires.

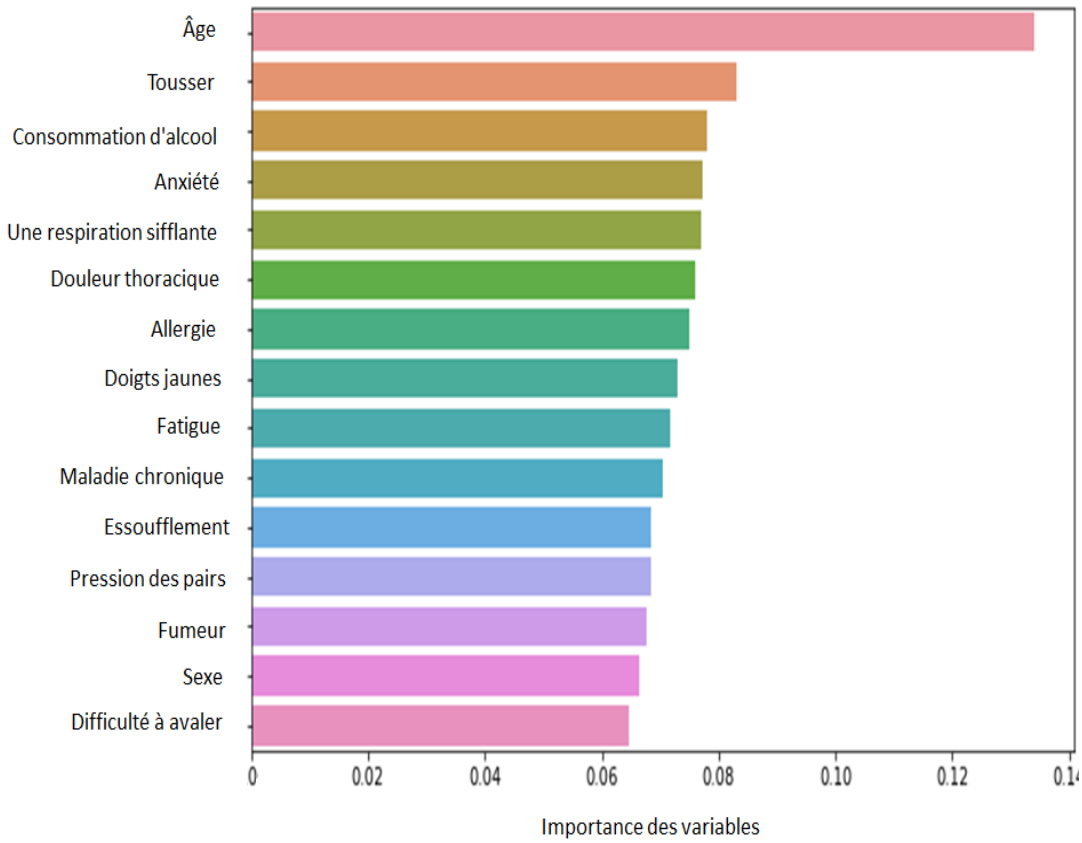


Figure V. 29 Illustration de l'importance des variables via CSABC-SVM

V.3.1.5 Interface graphique pour la prédiction de cancer du poumon

L'interface graphique développée pour la prédiction du cancer de poumon vise à offrir une expérience conviviale et intuitive aux utilisateurs, tout en permettant l'accès à des fonctionnalités avancées de prédiction basées sur des modèles d'apprentissage automatique. Elle est réalisée en utilisant le Framework Tkinter de Python pour assurer une portabilité et une facilité d'intégration avec divers systèmes d'exploitation.

Le but de cette interface est de prédire si une personne est atteinte d'un cancer de poumon ou non. Pour cela, l'utilisateur doit remplir le formulaire ci-dessous (voir la figure V. 30), qui contient les informations nécessaires à la prédiction.

Après avoir rempli le formulaire et cliqué sur prédiction, l'interface affiche les résultats de la prédiction comme illustré dans la figure V. 31.

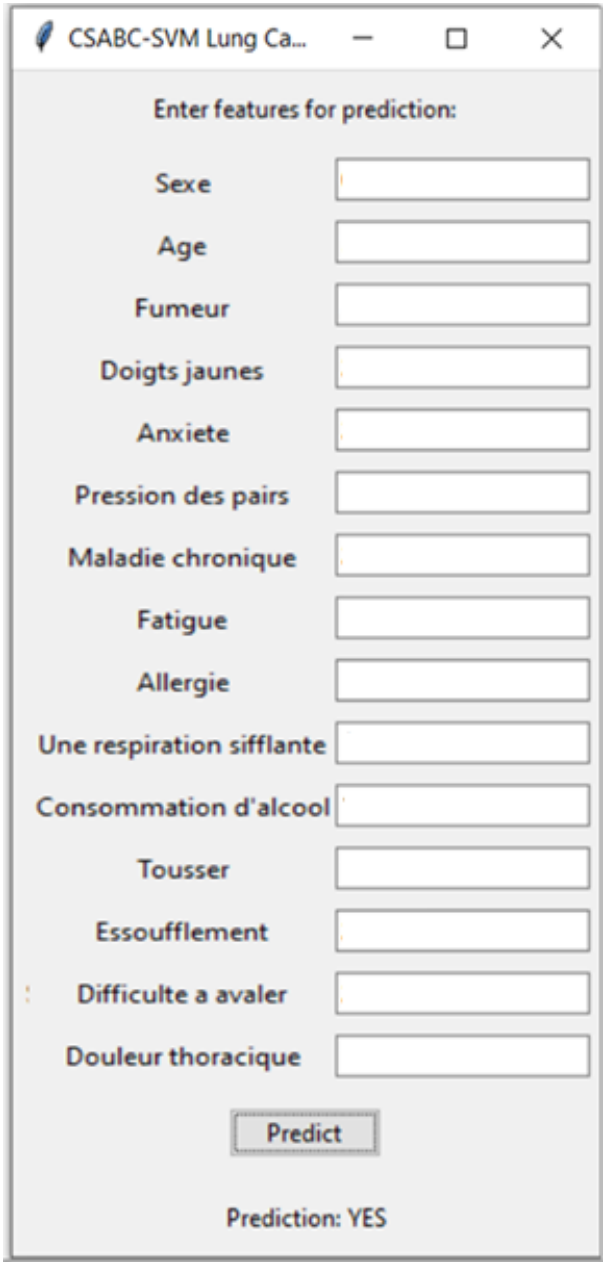


Figure V. 30 Interface graphique de cancer du poumon via CSABC-SVM

Dans le cas d'une prédiction positive l'interface affiche que le patient est atteint du cancer du poumon (voir la figure V.31). Dans le cas d'une prédiction négative, l'interface affiche que le patient n'est pas atteint du cancer du poumon.

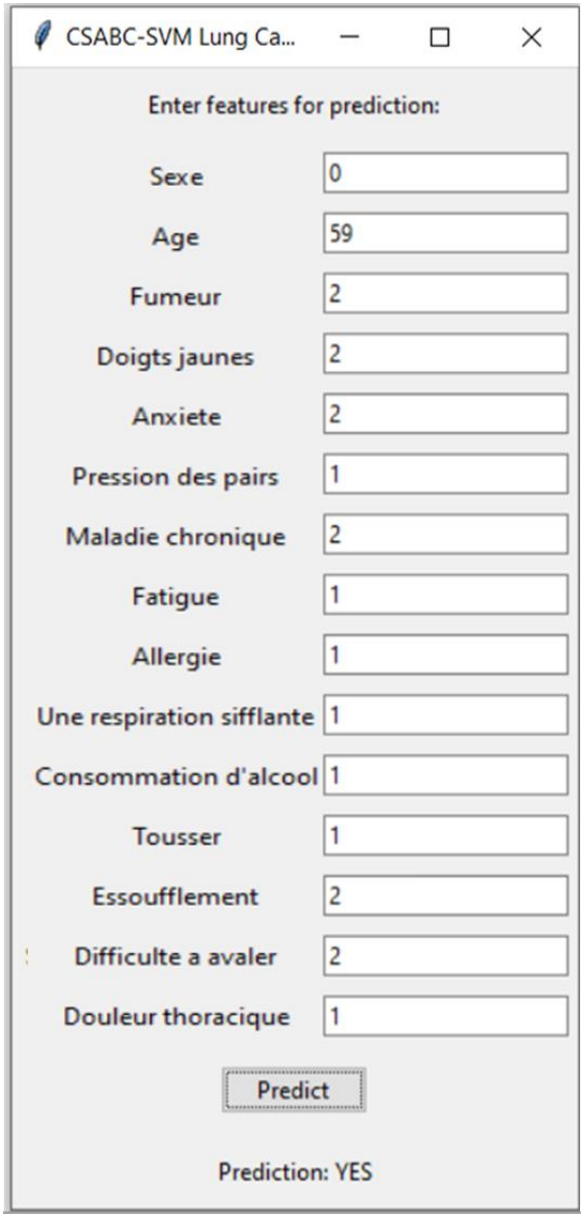


Figure V.31 Interface graphique de prédiction positive de cancer du poumon via CSABC-SVM

L'utilisation de modèles d'apprentissage automatique réduit les erreurs humaines dans l'interprétation des données médicales, améliorant ainsi la précision des diagnostics. Cette interface graphique facilite un diagnostic précoce et précis du cancer, offrant ainsi une solution pratique et accessible pour améliorer les soins de santé.

V.3.1.6 Comparaison avec d'autres études portant sur la prédiction du cancer du poumon

De nombreuses études de haute qualité ont été menées pour détecter et prédire le cancer du poumon. En comparaison avec d'autres recherches dans le domaine de la prédiction du cancer du poumon à l'aide de IA, notre étude se distingue par son approche globale et sa nouvelle technique d'hybridation, comme illustré dans le tableau V. 9.

Dans le cadre de notre étude comparative des performances des modèles de prédiction du cancer du poumon, plusieurs algorithmes ont été évalués en termes d'exactitude. RotF a affiché une exactitude de 97.1%, indiquant une performance remarquable dans la détection du cancer pulmonaire. En revanche, le modèle basé sur Gradient Boosted Tree a montré une exactitude de 90%, suggérant une performance légèrement inférieure. XGBoost a présenté une exactitude de 94.42%, tandis que le modèle PSO-GA-SVM a atteint une exactitude de 97.69%, dépassant celle du Gradient Boosted Tree et se rapprochant du RotF. Les approches basées sur l'hybridation, notamment KNN-SVM, WOA_SVM ont démontré des performances robustes avec des exactitude respectives de 97.6% et 95%. Notablement, notre modèle, CSABC-SVM, a surpassé tous les autres avec une exactitude de 98.38%, soulignant l'efficacité de notre approche hybride dans la prédiction du cancer du poumon.

Tableau V. 9. Comparaison du présent travail via la littérature

Auteur	**Nombre de données dans la base de données**	**Modèle**	**Ac (%)**
Faisal MI et al. 2019 [108]	32	Gradient Boosted Tree	90
Maulidina F et al. 2021 [107]	252	PSO-GA-SVM	97.69
Mamun M et al. 2022 [29]	309	XGBoost	94.42
Dritsas E et al. 2022 [110]	309	RotF	97.1
Vijila Rani K et al. 2022 [141]	700	KNN-SVM	97.6
Nanglia P et al. 2021 [142]	500	WOA-SVM	95
Présent travail	309	**CSABC-SVM**	**98.38**

V.4 Conclusion

En conclusion de ce travail de recherche, les contributions significatives dans les domaines de la prédiction du COVID-19, de l'hépatite viral, ainsi que du cancer du poumon, démontrent l'efficacité et la polyvalence des techniques avancées d'intelligence artificielle appliquées à la médecine. Les résultats obtenus mettent en évidence la capacité des modèles basés sur l'IA, tels que les RNA, les SVM et les algorithmes d'optimisation, à extraire des informations pertinentes à partir de données médicales complexes. Ces avancées offrent un potentiel considérable pour améliorer l'exactitude des diagnostics préventifs et pour soutenir les professionnels de la santé dans leurs efforts de prise en charge précoce des maladies. En somme, cette recherche contribue non seulement à l'évolution des méthodologies prédictives, mais également à l'amélioration de la gestion des maladies, ouvrant ainsi la voie à des perspectives prometteuses pour l'avenir de la médecine personnalisée et de la santé publique.

Conclusion Générale

Conclusion générale

Dans le domaine de la recherche scientifique, cette thèse représente une avancée significative dans l'utilisation de IA pour la prédiction et la compréhension des affections transmissibles et non transmissibles. À travers une exploration des progrès, des défis et des opportunités dans ce domaine en constante évolution, il devient évident que l'IA joue un rôle pivot dans le remodelage du paysage médical.

Chaque contribution au sein de ce travail sert de témoignage à l'efficacité des méthodes informatiques pour analyser des modèles complexes et guider des interventions de santé prédictive. Les résultats obtenus démontrent l'efficacité et la précision croissantes des modèles prédictifs basés sur l'IA, permettant une identification plus rapide et précise des maladies. Particulièrement dans le contexte des épidémies et des pandémies, telles que le COVID-19, la capacité à prédire les épidémies infectieuses avec précision est d'une importance primordiale. L'IA contribue de manière significative à la surveillance épidémiologique, à l'analyse de données massives et à la réponse rapide aux menaces microbiennes émergentes.

La prédiction du COVID-19 à l'aide des RNA a démontré une précision remarquable, offrant un aperçu du potentiel de l'IA dans la prévision des maladies infectieuses. De même, la prédiction réussie de l'hépatite à l'aide d'RNA souligne davantage l'efficacité de ces techniques dans l'identification et la réduction des menaces pour la santé publique. Cette réalisation a des implications significatives pour les stratégies de santé publique et le contrôle des épidémies

De plus, le modèle hybride CSABC-SVM utilisé pour prédire le cancer de poumon a non seulement atteint une exactitude élevée, mais a également approfondi notre compréhension des mécanismes de progression de la maladie. En élucidant les facteurs de risque, notamment l'impact du tabagisme et de la consommation d'alcool, cette recherche aide non seulement aux efforts de prédiction, mais offre également des perspectives précieuses pour des approches de médecine personnalisée adaptées aux profils de susceptibilité individuelle. Elle fournit des perspectives significatives pour la prévention précoce, l'amélioration globale des résultats médicaux, et potentiellement la réduction des coûts associés à une prise en charge tardive des maladies.

La prédiction aide le médecin à prendre des décisions concernant le diagnostic en fournissant des informations probables sur la présence ou l'évolution d'une maladie. Cela permet d'orienter les examens, de prioriser les traitements, et d'améliorer la précision et la rapidité des soins, entraînant ainsi des économies de coûts significatives

La prédiction des affections permet de prendre des mesures de prévention et de surveillance à différents niveaux :

- À l'échelle personnelle, elle aide à éviter certains facteurs de risque et à adopter un mode de vie préventif.
- À l'échelle nationale et mondiale, elle permet aux pays de fournir les ressources humaines et matérielles nécessaires pour gérer et contrôler les maladies, de développer un système de santé approprié, et de mettre en œuvre des mesures de sensibilisation et de prévention.

En regardant vers l'avenir, les implications de ces avancées sont profondes. Avec le perfectionnement continu des modèles prédictifs et l'intégration de l'analyse de big data, nous sommes à l'aube d'une nouvelle ère dans le domaine de la santé. Envisagez un avenir où les algorithmes d'IA surveillent sans effort les tendances de santé de la population, prédisent les épidémies de maladies et identifient des stratégies de traitement personnalisées basées sur les prédispositions génétiques individuelles et les expositions environnementales.

Cependant, alors que nous avançons vers cet avenir, il est impératif de reconnaître les implications éthiques et sociétales des soins de santé pilotés par l'IA. Garantir un accès équitable à ces technologies, protéger la confidentialité des patients et sont parmi les défis qui nous attendent.

Cette thèse marque une étape importante dans l'application de l'IA aux soins de santé, nous guidant vers un avenir où les maladies peuvent être prédites, prévenues et traitées avec une précision sans précédent, tout en potentiellement réduisant les coûts associés à la gestion des maladies.

Comme perspectives, nous envisageons dans le futur les points suivants :

- Utilisation d'autres techniques d'IA pour la prédiction d’autre maladies comme les accident vasculaire cérébral.
- Utilisation d'autres techniques d'IA telles que le CNN pour l'analyse d'images médicales.
- Utilisation d'autres algorithmes d'optimisation tels que PSO, GWO.
- La création de jumeaux numériques des patients, qui digitalise leurs caractéristiques telles que le génome et les antécédents médicaux, permet des traitements personnalisés facilitant l'évaluation de la réactivité d'une personne à une molécule en fonction de ses données génétiques.

Références bibliographiques

1. Nixon SA, Hunt M (2015) Global health: Where do physiotherapy and rehabilitation research fit? Physiother Canada 67:219–220. https://doi.org/10.3138/ptc.67.3.GEF
2. World Health Organization. Comité régional de la Méditerranée orientale (2015).

   Sécurité sanitaire mondiale, avec une attention particulière pour le MERS-CoV et le virus A(H5N1).
3. Rees EE, Ng V, Gachon P, et al (2019) Stratégie d ' évaluation des risques servant à la détection précoce et à la prédiction des éclosions de maladies infectieuses associées aux changements climatiques. 45:132–140
4. https://www.ohchr.org/fr/health/non-communicable-diseases. Consulté le 03 juin 2024
5. Villain Edouard (2021) Utilisation de l'intelligence artificielle pour l'aide au diagnostic des patients atteints de pathologies neuro dégénératives. 3–127
6. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56
7. Moukrim B (2019) Intelligence artificielle en santé: espoirs et craintes des médecins généralistes
8. Arduin H (2018) Modélisation mathématique des interactions entre pathogènes chez l'hôte humain : Application aux virus de la grippe et au pneumocoque
9. Traboulsi H (2016) Étude des rôles de la voie antioxydante Nrf2 et la voie anti-inflammatoire PPARγ dans le contrôle de l'inflammation lors d'infections sévères par l'influenza
10. Bouscambert-Duchamp MD (2010) Étude du réassortiment génétique des virus influenza d'origines et de sous-types différents
11. Health Organization. (2021). Tuberculosis (TB). [https://www.who.int/news-room/fact-sheets/detail/tuberculosis]
12. Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, et al. (2021) Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics 8:1–17. https://doi.org/10.3390/informatics8020027

13. Nnoaham KE, Clarke A (2008) Low serum vitamin D levels and tuberculosis: A systematic review and meta-analysis. Int J Epidemiol 37:113–119. https://doi.org/10.1093/ije/dym247

14. Parisi L, RaviChandran N, Manaog ML (2020) A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis. Neural Comput Appl 32:3839–3852

15. Lanini S, Ustianowski A, Pisapia R, et al. (2019) Viral hepatitis: etiology, epidemiology, transmission, diagnostics, treatment, and prevention. Infect Dis Clin 33:1045–1062

16. Sallard E, Halloy J, Casane D, et al (2020) Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus. médecine/sciences 36:783–796

17. Fadhila B (2022) Histoire de l'évolution des coronavirus et l'apparition de nouveaux variants Covid-19 et Covid-20

18. Swapna G, Vinayakumar R, Soman KP (2018) Diabetes detection using deep learning algorithms. ICT express 4:243–246

19. OMS: Global Report on c. Isbn. 978, 6–86 (2016).

20. World Health Organization. Fact sheet: Diabetes. Published 2021. Accessed October 11, 2021. https://www.who.int/news-room/fact-sheets/detail/diabetes

21. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2004;27(suppl 1): s5-s10. doi:10.2337/diacare.27.2007.S5

22. https://intermed-undh.weebly.com/home/le-diabete-tout-ce-quil-faut-savoir1380664. Consulté le 10 février 2024

23. https://www.nhs.uk/conditions/cardiovascular-disease/. Consulté le 03 juin 2024

24. Yan T, Zhu S, Xie C, et al. (2022) Coronary artery disease and atrial fibrillation: a bidirectional Mendelian randomization study. J Cardiovasc Dev Dis 9:69

25. Liu J, Dong X, Zhao H, Tian Y (2022) Predictive classifier for cardiovascular disease based on stacking model fusion. Processes 10:749

26. Sun L, Zhou Y, Zhang M, et al. (2022) Association of major chronic noncommunicable diseases and life expectancy in China, 2019. In: Healthcare. MDPI, p 296

27. World Health Organization Maladies cardiovasculaires

28. https://www.ameli.fr/assure/sante/themes/risque-cardiovasculaire/definition-facteurs-

favorisants. Consulté le 07 mars 2024

29. Mamun M, Farjana A, Al Mamun M, Ahammed MS (2022) Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. 2022 IEEE World AI IoT Congr AIIoT 2022 187–193. https://doi.org/10.1109/AIIoT54504.2022.9817326

30. Raoof SS, Jabbar MA, Fathima SA (2020) Lung Cancer Prediction using Machine Learning: A Comprehensive Approach. 2nd Int Conf Innov Mech Ind Appl ICIMIA 2020 - Conf Proc 108–115. https://doi.org/10.1109/ICIMIA48430.2020.9074947

31. https://fr.freepik.com/vecteurs-premium/diagramme-montrant-stades-du-cancer-du-poumon_1909898.htm

32. Rajesh S, Choudhury NA, Moulik S (2020) Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms. In: 2020 IEEE 17th India council international conference (INDICON). IEEE, pp 1–5

33. Kleeff J, Korc M, Apte M, et al (2016) Pancreatic cancer. Nat Rev Dis Prim 2:1–23. https://doi.org/10.1038/nrdp.2016.22

34. Vincent A, Herman J, Schulick R, et al (2011) Pancreatic cancer. Lancet 378:607–620

35. Goodfellow I, Bengio Y, Courville A. Deep Learning Book. Disponible en ligne: https://www.deeplearningbook.org

36. Russel S, Norvig P. (2010) Artificial Intelligence: A Modern Approach. Third Edition. Pearson Education France.

37. Zemihi M (2022) L ’ utilisation de l ’ intelligence de artificielle dans la découverte de nouveaux candidats-médicaments et dans l ’ imagerie médicale Mazari Zemihi To cite this version : HAL Id : hal-03298153 soutenance et mis à disposition de l ’ ensemble de la Contact

38. Alliot J, Schiex T, Brisset P, Garcia F (2002) Intelligence artificielle et informatique théorique.

39. Cornuéjols A, Miclet L (2011) Apprentissage artificiel: concepts et algorithmes. Editions Eyrolles

40. Kumar A, Sarkar BK (2017) A case study on machine learning classification. Int J Inf Decis Sci 9:179–208. https://doi.org/10.1504/IJIDS.2017.084885

41. Géron A (2017) Hands-on machine learning with Scikit-Learn and TensorFlow. Concepts, tools, Tech to build Intell Syst

42. Houfani D (2023) Une approche basée sur les SMA et méta-heuristique pour la prédiction (PHM) médicale

43. Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. John Wiley & Sons

44. Dai W, Brisimi TS, Adams WG, et al. (2015) Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform 84:189–197

45. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19:1–16

46. Roubehie Fissa M (2021) Estimation de quelques propriétés physico-chimiques des hydrocarbures en fonction de leurs Propriétés (RQSP) en utilisant la méthode de Réseaux de Neurone Artificiel (RNA)

47. Hagan MT, Demuth HB, Beale MH (1996) Neural network design, PWS Pub. Co., Boston, 3632. Haykin, S.(1994). Neural networks a Compr Found Prentice Hall PTR

48. Suzuki K (2011) Artificial neural networks: methodological advances and biomedical applications. BoD–Books on Demand

49. Haykin S (1998) Neural networks: a comprehensive foundation. Prentice Hall PTR

50. Baillot, D. (2018). Why are neuron axons long and spindly? Medical Xpress, July, 11–13.https://ucsdnews.ucsd.edu/pressrelease/why_are_neuron_axons_long_and_spindly%0Ahttps://medicalxpress.com/news/2018-07-neuron-axons-spindly-theyre-optimizing.html%0Ahttps://medicalxpress.com/news/2018-07-neuron-axons-spindly-theyre-optimizing.html%0A

51. Khaouane L (2013) Etude et Modélisation de la Biosynthèse des Antibiotiques à partir de Différentes Souches Productrices– Cas de Pleuromutiline". Thèse de Doctorat en génie des procédés, Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

52. Bitam S (2018) Modélisation de l'activité thérapeutique à partir de la structure moléculaire". Thèse de Doctorat en génie des procédés, Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

53. Si Moussa C (2010) Modélisation des équilibres de phases à pressions élevées par les réseaux de neurones artificiels

54. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Math Control signals Syst 2:303–314

55. Bishop CM (1995) Neural Networks for Pattern Recognition. Clarendon Press google Sch 2:223–228

56. Statsoft. (2020a). Statsoft.documentation Statistica Help SANN Overviews. Disponible en ligne:https://documentation.statsoft.com/STATISTICAHelp.aspx?path=SANN/Overview/SANNO verviewsActivationFunctions.

57. Statsoft. (2020b). Concepts Fondamentaux en Statistique- Réseaux de Neurones- Fonctions d'Activation. Disponible en ligne : http://www.statsoft.fr/concepts-statistiques/reseaux-deneurones-automatises/reseaux-de-neurones-automatises.htm#fonctions.

58. Ammar MY (2007) Mise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition batch/continu

59. Chtita S (2017) Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments

60. Minoux M (1983) Programmation mathématique. Théorie et algorithmes dunod. Minoux– Paris : Bordas et Cnetenst.

61. Oussar Y (1998) Réseaux d'ondelettes et réseaux de neurones pour la modélisation statique et dynamique de processus". Thèse de doctorat en Robotique, Université Pierre et Marie Curie- Paris VI, France. Disponible en ligne : https://hal.archives-ouvertes.fr/pastel-00000677.

62. Dreyfus G, Martinez J M, Samuelides M, Gordon M B, Badran F, & Thiria S (2008). Apprentissage statistique : Réseaux de neurones : Cartes topologiques Machines à

vecteurs supports, eyrolles.

63. Urbani, D. (1995). "Méthodes statistiques de sélection d'architectures neuronales : application à la conception de modèles de processus dynamiques". Thèse de doctorat en Microélectronique et Microinformatique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://www.neurones.espci.fr/Theses_PS/Urbani_D/these.pdf.

64. Lucea M (2006) Modélisation dynamique par réseaux de neurones et machines à vecteurs supports: contribution à la maîtrise des émissions polluantes de véhicules automobiles. Thèse de doctorat en Electronique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://pastel.archives-ouvertes.fr/pastel-00001943.

65. Sigma. (2020). Laboratoire SIGnaux, Modèles, Apprentissage statistique. 10, Rue Vauquelin 75231 PARIS Cedex 05, France. Site Officiel : https://www.neurones.espci.fr/.

66. Ammi Y (2016). "Modélisation des phénomènes de transport lors d′osmose inverse". Thèse de Doctorat en génie des procédés, Option : Modélisation des phénomènes de transport en sciences de l'eau et d'environnement, Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

67. Laidi M, Abdallah el HA, Si-Moussa C, et al. (2021) CMC of diverse Gemini surfactants modelling using a hybrid approach combining SVR-DA. Chem Ind Chem Eng Q 27:299–312

68. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106. https://doi.org/10.1007/bf00116251

69. Ben Amor N, Benferhat S, Elouedi Z (2006) Réseaux bayésiens naïfs et arbres de décision dans les systèmes de détection d'intrusions. Tech Sci informatiques 25:167–196. https://doi.org/10.3166/tsi.25.167-196

70. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

71. Marshall RJ (2001) The use of classification and regression trees in clinical epidemiology. J Clin Epidemiol 54:603–609

72. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. J Med Syst 26:445–463

73. Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. J Artif Intell Res 11:169–198

74. Shaikhina T, Lowe D, Daga S, et al (2019) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 52:456–462. https://doi.org/10.1016/j.bspc.2017.01.012

75. Hu X (2017) Support Vector Machine and Its Application to Regression and Classification TO REGRESSION AND CLASSIFICATION

76. Madani A, Benkortbi O, Laidi M (2023) In silico prediction of the inhibition of new molecules on SARS-CoV-2 3CL protease by using QSAR: PSOSVR approach. Brazilian J Chem Eng. https://doi.org/10.1007/s43153-023-00332-z

77. Rashidi HH, Tran NK, Betts EV, et al. (2019) Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. Acad Pathol 6:2374289519873088. https://doi.org/10.1177/2374289519873088

78. Tompson JJ, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. Adv Neural Inf Process Syst 27:

79. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

80. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp 234–241

81. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90

82. Raina R, Madhavan A, Ng AY (2009) Large-scale deep unsupervised learning using graphics processors. In: Proceedings of the 26th annual international conference on machine learning. pp 873–880

83. Bizopoulos P, Koutsouris D (2018) Deep learning in cardiology. IEEE Rev Biomed Eng 12:168–193

84. Xia Y, Wulan N, Wang K, Zhang H (2018) Detecting atrial fibrillation by deep convolutional neural networks. Comput Biol Med 93:84–92

85. Yasen M, Al-Madi N, Obeid N (2018) Optimizing Neural Networks using Dragonfly Algorithm for Medical Prediction. 2018 8th Int Conf Comput Sci Inf Technol CSIT 2018 71–76. https://doi.org/10.1109/CSIT.2018.8486178

86. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes university, engineering faculty, computer

87. Euldji I, Belghait A, Si-Moussa C, et al. (2023) A new hybrid quantitative structure property relationships-support vector regression (QSPR-SVR) approach for predicting the solubility of drug compounds in supercritical carbon dioxide. AIChE J. https://doi.org/10.1002/aic.18115

88. Askarzadeh A (2016) A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Comput Struct 169:1–12. https://doi.org/10.1016/j.compstruc.2016.03.001

89. Sayed GI, Hassanien AE, Azar AT (2019) Feature selection via a novel chaotic crow search algorithm. Neural Comput Appl 31:171–188. https://doi.org/10.1007/s00521-017-2988-6

90. Tared S, Khaouane L, Hanini S, et al (2024) Enhancing lung cancer prediction through crow search, artificial bee colony algorithms, and support vector machine. Int J Inf Technol 1–11

91. Mirjalili S (2016) Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Comput Appl 27:1053–1073

92. Laidi M, Abdallah el HA, Si-Moussa C, et al (2021) CMC of diverse Gemini surfactants modelling using a hybrid approach combining SVR-DA. Chem Ind Chem Eng Q 27:299–312

93. Madani A (2023) Study and Development of QSAR Techniques for the Discovery of

Therapeutically Relevant Molecules: Prediction of Potential Inhibitors of SARS-CoV-2, Doctoral thesis

94. de Almeida BSG, Leite VC (2019) Particle swarm optimization: A powerful technique for solving engineering problems. Swarm Intell Adv new Perspect Appl 31–51

95. Mitikiri P, Jana G, Sural S, Chattaraj PK (2018) A machine learning technique toward generating minimum energy structures of small boron clusters

96. Esteva A, Kuprel B, Novoa RA, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118

97. Diebolt V, Azancot I, Boissel FH, et al. (2019) « Intelligence artificielle » : quels services, quelles applications, quels résultats et quelle valorisation aujourd'hui en recherche clinique ? Quel impact sur la qualité des soins ? Quelles recommandations ? Therapie 74:141–154. https://doi.org/10.1016/j.therap.2018.11.006

98. Heeke S, Delingette H, Fanjat Y, et al. (2019) La pathologie cancéreuse pulmonaire à l'heure de l'intelligence artificielle: entre espoir, désespoir et perspectives. In: Annales de Pathologie. Elsevier, pp 130–136

99. Monnier E. L'intelligence artificielle relève aussi le défi du suicide [Internet]. Science et vie. 2018 [cité 6 sept 2019]. Disponible sur: https://www.science-et-vie.com/technos-et-futur/l-intelligence-artificielle-releve-aussi-le-defi-du-suicide-17226

100. Sun R, Limkin EJ, Vakalopoulou M, et al. (2018) A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy : an imaging biomarker , retrospective multicohort study. 19:

101. Anichini G, Geffroy B, John É, et al. (2021) L'intelligence artificielle à l'épreuve des savoirs tacites. Analyse des pratiques d'utilisation d'un outil d'aide à la détection en radiologie. 39:43–69. https://doi.org/10.1684/sss.2021.0200

102. Lavdas I, Glocker B, Rueckert D, et al. (2019) Machine learning in whole-body MRI: experiences and challenges from an applied study using multicentre data. Clin Radiol 74:346–356

103. Rajpurkar P, Irvin J, Zhu K, et al. (2017) CheXNet : Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 3–9

104. Huang JA, Hartanti IR, Colin MN, Pitaloka DAE (2022) Telemedicine and artificial intelligence to support self-isolation of COVID-19 patients: Recent updates and challenges. Digit Heal 8:20552076221100630

105. Tozour JN, Bandremer S, Patberg E, et al. (2021) Application of telemedicine video visits in a maternal-fetal medicine practice at the epicenter of the COVID-19 pandemic. Am J Obstet Gynecol MFM 3:100469

106. Fagherazzi G, Ravaud P (2018) Digital diabetes : perspectives for diabetes prevention ,

management and research

107. Singh GAP, Gupta PK (2019) Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Comput Appl 31:6863–6877. https://doi.org/10.1007/s00521-018-3518-x

108. Faisal MI, Bashir S, Khan ZS, Hassan Khan F (2019) An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer. 2018 3rd Int Conf Emerg Trends Eng Sci Technol ICEEST 2018 1–4. https://doi.org/10.1109/ICEEST.2018.8643311

109. Wu J, Zan X, Gao L, et al. (2019) A machine learning method for identifying lung cancer based on routine blood indices: Qualitative feasibility study. JMIR Med Informatics 7:1–12. https://doi.org/10.2196/13476

110. Dritsas E, Trigka M (2022) Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn Comput 6:. https://doi.org/10.3390/bdcc6040139

111. Maulidina F, Rustam Z, Pandelaki J (2021) Lung Cancer Classification using Support Vector Machine and Hybrid Particle Swarm Optimization-Genetic Algorithm. 2021 Int Conf Decis Aid Sci Appl DASA 2021 751–755. https://doi.org/10.1109/DASA53625.2021.9682259

112. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585

113. Chen W, Chen S, Zhang H, Wu T (2017) A hybrid prediction model for type 2 diabetes using K-means and decision tree. In: 2017 8th IEEE international conference on software engineering and service science (ICSESS). IEEE, pp 386–390

114. Olanrewaju R, Musa N, Mohd Paudzi NFA (2018) Early Prediction of Cardiovascular Disease Level Based using Artificial Neural Networks

115. Arroyo JCT, Delima AJP (2022) An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction. J Adv Inf Technol 13:95–99. https://doi.org/10.12720/jait.13.1.95-99

116. Talukdar J, Singh TP (2023) Early prediction of cardiovascular disease using artificial neural network. Paladyn 14:. https://doi.org/10.1515/pjbr-2022-0107

117. Edeh MO, Dalal S, Dhaou I Ben, Agubosim CC (2022) Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease. 10:. https://doi.org/10.3389/fpubh.2022.892371

118. Drobo A, Smajović A (2022) Application of artificial neural networks in diagnosis of Hepatitis C. 1–5

119. Yarasuri VK, Indukuri GK, Nair AK (2019) Prediction of hepatitis disease using machine learning technique. In: 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, pp 265–269

120. El-Solh AA, Hsiao C-B, Goodnough S, et al. (1999) Predicting active pulmonary tuberculosis using an artificial neural network. Chest 116:968–973

121. Alakus TB, Turkoglu I (2020) Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals 140:110120

122. Kumar S, Vardhan H, Priya S, Kumar A (2023) Malaria detection using deep convolution neural network. arXiv Prepr arXiv230303397

123. Khan MA, Abidi WUH, Al Ghamdi MA, et al (2021) Forecast the influenza pandemic using machine learning. Comput Mater Contin 66:331–340. https://doi.org/10.32604/cmc.2020.012148

124. Hamadache M, Hanini S, Benkortbi O, et al. (2016) Artificial neural network-based equation to predict the toxicity of herbicides on rats. Chemom Intell Lab Syst 154:7–15. https://doi.org/https://doi.org/10.1016/j.chemolab.2016.03.007

125. Bhadula S, Sharma S, Juyal P, Kulshrestha C (2019) Machine Learning Algorithms based Skin Disease Detection. Int J Innov Technol Explor Eng 9:4044–4049. https://doi.org/10.35940/ijitee.b7686.129219

126. Ali MM, Paul BK, Ahmed K, et al. (2021) Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Comput Biol Med 136:104672. https://doi.org/10.1016/j.compbiomed.2021.104672

127. Feuilloy M (2009) Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme. Introduction Prédiction de la syncope Analyses du signal d'impédancemétrie thoracique.

128. httpsgithub.comakarshan-githubCovid-19-detector-probability. Accessed 18 Des 2023

129. Wessel MD, Jurs PC (1994) Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks. Anal Chem 66:2480–2487

130. Ammi Y, Khaouane L, Hanini S (2015) Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. Korean J Chem Eng 32:2300–2310. https://doi.org/10.1007/s11814-015-0086-y

131. Roubehie Fissa M, Lahiouel Y, Khaouane L, Hanini S (2019) QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. J Mol Graph Model 87:109–120. https://doi.org/https://doi.org/10.1016/j.jmgm.2018.11.013

132. Garson GD (1991) Interpreting neural-network connection weights. AI Expert 6:46–51

133. Goh ATC (1995) Back-propagation neural networks for modeling complex systems. Artif Intell Eng 9:143–151

134. Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Modell 160:249–264. https://doi.org/10.1016/S0304-3800(02)00257-0

135. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Accessed 22 Jan 2023

136. Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P (2012) Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics 28:1766–1774. https://doi.org/10.1093/bioinformatics/bts238

137. Roslina AH, Noraziah A (2010) Prediction of hepatitis prognosis using support vector machines and wrapper method. Proc - 2010 7th Int Conf Fuzzy Syst Knowl Discov FSKD 2010 5:2209–2211. https://doi.org/10.1109/FSKD.2010.5569542

138. Polat K, Güneş S (2007) Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. Appl Math Comput 189:1282–1291. https://doi.org/10.1016/j.amc.2006.12.010

139. Çalişir D, Dogantekin E (2011) A new intelligent hepatitis diagnosis system: PCA-LSSVM. Expert Syst Appl 38:10705–10708. https://doi.org/10.1016/j.eswa.2011.01.014

140. Gürsoy Mi, Alkan A (2022) Investigation of diabetes data with permutation feature importance based deep learning methods. Karadeniz Fen Bilim Derg 12:916–930. https://doi.org/10.31466/kfbd.1174591

141. Vijila Rani K, Joseph Jawhar S (2022) Lung lesion classification scheme using optimization techniques and hybrid (KNN-SVM) Classifier. IETE J Res 68:1485–1499. https://doi.org/10.1080/03772063.2019.1654935

142. Nanglia P, Kumar S, Mahajan AN, et al. (2021) A hybrid algorithm for lung cancer classification using SVM and Neural Networks. ICT Express 7:335–341. https://doi.org/10.1016/j.icte.2020.06.007

Annexes

Tableau A.1 Exactitude avec nombre d'itérations

Division		Nombre d'itérations	Nombre de neurones	Ac %
Apprentissage	**Première division**	500	MLP 19-2-2	85
Test		1000	MLP 19-2-2	95
Validation		1500	MLP 19-2-2	96.25
Apprentissage	**Deuxième division**	500	MLP 19-2-2	88.75
Test		1000	MLP 19-2-2	85
Validation		1500	MLP 19-2-2	85
Apprentissage	**Troisième division**	500	MLP 19-2-2	88.75
Test		**1000**	**MLP 19-3-2**	**97.5**
Validation		2000	MLP 19-2-2	81.25

Tableau B.1 Poids et biais du meilleur modèle RNA -PMC (7-4-2).

Entrée – Connexions des couches cachées								**Caché – Connexion des couches de sortie**			
Poids							Biais	Poids		Bias	
$\boldsymbol{w_{j1}^{I}}$	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	b_{j}^{H}	w_{0j}^{H}	w_{1j}^{H}	b_{0}^{O}	b_{1}^{O}
-7,9082	0,2426	0,8064	0,8391	0,8467	1,3481	1,5873	2,6894	2,7161	-2,7130	1,4857	-1,6180
-0,2384	0,2674	1,1877	1,4547	1,6071	1,0823	1,5770	-3,0605	3,3300	3,3701		
-13,9786	6,8792	-0,0439	-3,1437	-5,7352	32,2298	-7,2238	3,3143	-0,7615	0,7594		
0,8235	0,9355	0,6087	0,8339	-0,2480	0,4567	21,5793	-0,3658	-1,6026	1,6449		

Tableau B.2. Poids et biais du meilleur modèle RNA-PMC (19-3-2).

Entrée – Connexions des couches cachées																			**Caché – Connexion des couches de sortie**			
Poids																		Bias	Poids		Bias	
w_{j1}^{I}	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	w_{j8}^{I}	w_{j9}^{I}	w_{j10}^{I}	w_{j11}^{I}	w_{j12}^{I}	w_{j13}^{I}	w_{j14}^{I}	w_{j15}^{I}	w_{j6}^{I}	w_{j17}^{I}	w_{j18}^{I}	w_{j19}^{I}	b_j^H	w_{0j}^{H}	w_{1j}^{H}	b_0^O	b_1^O
-0,46080	0,11684	-0,22269	-0,70907	0,65643	0,08052	0,42548	-0,25708	-0,16828	-0,76646	-0,08083	0,05756	-0,24681	-0,87328	0,64427	-1,27013	0,10139	-0,30246	-0,12262		0,11103	-2,35992		
0,05115	0,02371	0,54031	0,68331	-0,63788	-1,45887	0,80055	0,32468	-0,33978	0,54838	0,28531	-1,23065	0,51456	1,30768	0,06551	-1,13638	0,30974	0,61407	0,09449		0,18051	-0,33261		
0,36195	-0,49103	-1,02526	-0,43872	0,41931	-0,95727	2,39792	1,53112	-0,08677	-0,62311	-0,32043	-1,27576	0,08383	0,14058	1,46135	-1,80020	-0,46261	-0,63954	1,36749		0,11780	-2,12627		

Valorisations Scientifiques : articles et séminaires

Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-024-01770-9

ORIGINAL RESEARCH

Enhancing lung cancer prediction through crow search, artificial bee colony algorithms, and support vector machine

Samira Tared[1] · Latifa Khaouane[1] · Salah Hanini[1] · Affaf Khaouane[1] · Mohamed Roubehie Fissa[1]

Received: 28 October 2023 / Accepted: 30 January 2024
© The Author(s), under exclusive licence to Bharati Vidyapeeth's Institute of Computer Applications and Management 2024

Abstract Lung cancer is one of the leading causes of death worldwide. According to statistics from the World Health Organization (WHO) for the year 2020, lung cancer has caused the death of about 1.80 million people. It is the most common form of cancer, and it is often diagnosed in later stages when metastatic spread to other areas of the body has occurred. Early detection and treatment can help reduce the high death rate. In this research, we present a novel approach for predicting lung cancer using artificial intelligence (AI) techniques. We evaluated and compared several machine learning (ML) models, including artificial neural networks (ANN), K-nearest neighbors (KNN), and support vector machine (SVM), to identify the best-performing model. However, to enhance accuracy and efficiency, we introduced a hybridization optimization approach known as Crow Search Algorithm -Artificial Bee Colony-Support Vector Machine (CSABC-SVM). The CSABC-SVM algorithm combines the strengths of the CSABC Optimizer and SVM to create a powerful predictive model. Our study yielded a remarkable accuracy of 98.36%, demonstrating the effectiveness of the CSABC-SVM hybrid model in lung cancer prediction. These results underscore the potential of AI in transforming medical diagnostics, helping to improve healthcare outcomes.

Keywords Lung cancer · Prediction · Artificial intelligence · Optimization · CSABC-SVM

✉ Samira Tared
tared.samira@gmail.com; samirstard@gmail.com

Latifa Khaouane
latifa_khaouane@yahoo.fr

Salah Hanini
s_hanini2002@yahoo.fr

Affaf Khaouane
affoufa80@gmail.com

Mohamed Roubehie Fissa
roubehiefissa.mohamed2021@gmail.com

[1] Biomaterials and Transport Phenomena Laboratory (LBMPT), Department of Process and Environmental Engineering, Faculty of Technology, University of Yahia Fares, 26000 Medea, Algeria

1 Introduction

Lung cancer is one of the leading causes of death worldwide [1–3]. According to the WHO study, lung cancer was the deadliest cancer, with about 1.80 million people dying from it in 2020 [4]. It is the second most prevalent kind of cancer and is often diagnosed at later stages when metastatic dissemination to other regions of the body has occurred [2]. As it originates in the lining of the bronchi and bronchioles, lung cancer obstructs gas exchange in the lungs and causes breathing difficulties [4]. It begins as an abnormal growth of cells in the lungs and can spread to other parts of the body [4]. Difficult to evaluate and discover in the early stages since it does not create pain or symptoms in certain situations [5]. A patient diagnosed with Lung cancer may experience symptoms such as chest pain, cough, shortness of breath, wheezing, hemoptysis i.e. coughing up blood, Pancoast syndrome (shoulder pain), hoarseness (paralysis of vocal cords), weakness, weight loss, fatigue [5]. Symptoms are classified according to the location and size of the tumor [5]. Early detection and treatment can help to lower such a high death rate [6].

In recent years, there has been continuous development in cancer research [7]. Multitudes of procedures are used to diagnose cancer and predict the outcomes of cancer treatment [4]. Researchers used a variety of methods,

Published online: 04 March 2024



CERTIFICATE Of PARTICIPATION

Samira Tared

has participated in 4th International Conference on Applied Engineering and Natural Sciences on 10-13 November in 2022 at Konya/Turkey.

PAPER TITLE *Prediction of Hepatitis Prognosis Using Multilayer Perceptron*

PRESENTATION TYPE *Oral*

ICAENS 2022

ICAENS 2022 CONFERENCE CHAIRMAN

Asst. Prof. Dr. Umut ÖZKAYA



CERTIFICATE Of PARTICIPATION

Samira Tared

has participated in 1st International Conference on Scientific and Academic Research on 10-13 December in 2022 at Konya/Turkey.

PAPER TITLE *Diabetes disease prediction using artificial neural network*

PRESENTATION TYPE *Oral*

ICSAR

ICSAR 2022 CONFERENCE CHAIRMAN
Asst. Prof. Dr. Umut ÖZKAYA