République Algérienne Démocratique et Populaire
Ministère de L'enseignement Supérieur et de la Recherche Scientifique

Université A/Mira de Béjaia
Faculté des Sciences Exactes
Département d'Informatique

Projet de fin de cycle

Spécialité : *Master 2 Systèmes d'Information Avancés (SIA)*

Thème

Intégration des techniques d'apprentissage automatique dans le suivi d'une maladie chronique
Cas d'étude : Prédiction d'une maladie cardiovasculaire

Réalisé par : - BENACHOUR Imene
- KATTI Melissa

Encadré par : Dr K. AKILAL

Soutenu le 30 juin 2025 devant les membres du jury

Dr A. ACHROUFENE	Président	U.A/Mira Béjaia
Dr D. BOULAHROUZ	Examinatrice	U.A/Mira Béjaia
Dr F. BOUCHEBBAH	Examinateur	U.A/Mira Béjaia
Dr N. SAAD	Examinatrice	U.A/Mira Béjaia

2024/2025

⁂ *Remerciements* ⁂

En premier lieu, nous tenons à remercier Dieu Le Tout Puissant de nous avoir dotées de force, de patience et de capacité afin de réaliser ce modeste travail.

Nous tenons également
à exprimer notre profonde gratitude envers :

Nos parents
qui n'ont jamais cessé de croire en nous et de nous encourager.

Notre encadrant ***Dr AKILAL Karim***
pour ses conseils et son aide précieuse.

L'ensemble des membres du jury
pour avoir accepté d'examiner notre travail.

Nos amies
pour leur présence constante et leur soutien moral tout au long de ce parcours.

Toute personne
ayant contribué de près ou de loin à l'élaboration de ce projet.

⁂ *Dédicaces* ⁂

Nous dédions ce modeste travail :

À nos très chers
parents.

À nos aimables frères et sœurs.

À nos précieuses
amies.

À nos enseignants bienveillants.

Imene et Melissa

RÉSUMÉ

Les maladies cardiovasculaires font partie des principales causes de décès dans le monde. Leur prédiction et leur prévention restent un défi pour les médecins. Cependant, les avancées récentes en intelligence artificielle et en apprentissage automatique offrent des perspectives nouvelles et prometteuses pour la prédiction de ces maladies en particulier, et pour le domaine médical de façon générale.

Avant de mettre en avant notre approche, nous avons analysé divers travaux de la littérature dans ce domaine.

Ce travail propose un modèle combinant deux (02) algorithmes d'apprentissage automatique, qui sont les suivants : *Naïve Bayes* et *KNN*, avec la méthode *"Stacking"* en utilisant une *régression logistique*. Afin d'améliorer la précision de la prédiction, nous avons appliqué *SMOTE-NC* pour la gestion du déséquilibre des classes, et *MRMR* pour la sélection des caractéristiques les plus pertinentes. Nous avons pu prouver l'efficacité de notre modèle après l'avoir testé et évalué selon plusieurs métriques d'évaluation, telles que : l'exactitude, la précision, le rappel, la F1-mesure, l'AUC et la courbe ROC.

Mots-clés : Maladies cardiovasculaires, Intelligence Artificielle, Apprentissage Automatique, Prédiction, Stacking, Naïve Bayes, KNN, Régression Logistique, SMOTE-NC, MRMR.

ABSTRACT

Cardiovascular diseases are among the main causes of death in the world. Their prediction and prevention remain a challenge for doctors. However, recent artificial intelligence and machine learning developments offer new perspectives for diseases prediction in particular, and the medical field in general.

Before highlighting our approach, we have analysed various works from literature in this field.

This work proposes a model combining two machine learning algorithms, namely : *Naive Bayes* and *KNN*, alongside the *"Stacking"* method using *logistic regression (LR)*. In order to improve the prediction precision, we have applied *SMOTE-NC* for classes imbalance management and *MRMR* for pertinent feature selection. We were able to prove the efficiency of our model by testing it using various evaluation metrics such as : accuracy, precision, recall, F1-score, AUC and ROC curve.

Keywords : Cardiovascular Diseases, Artificial Intelligence, Machine Learning, Prediction, Stacking, Naive Bayes, KNN, Logistic Regression, SMOTE-NC, MRMR.

Table des matières

Table des figures

Liste des tableaux

Liste des algorithmes

Liste des abréviations

AB Adaptive Boosting

AHP Analytic Hierarchy Process

ANN Artificial Neural Network

ANOVA ANalysis Of VAriance

AUC Area Under the ROC Curve

AVC Accident Vasculaire Cérébral

BGGLM Boosted Generalized Linear Model

BGLM Bayesian Generalized Linear Model

BN Bayesian Network

BSD Berkeley Software Distribution

BT Bagged Tree

CatBoost Categorical Boosting

CDTL Cluster-based Decision Tree Learning

CIT Conditional Inference Tree

CNN Convolutional Neural Network

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DL Deep Learning

DNN Deep Neural Network

DQN Deep Q-Network

DT Decision Tree

ECG ElectroCardioGram

EPA Environmental Protection Agency

ET Extra Tree

FDA Flexible Discriminant Analysis

FFC Fédération Française de Cardiologie

FN Faux Négatif

FP Faux Positif

FPR False Positive Rate

GA Genetic Algorithm

GBDT Gradient Boosting Decision Tree

GBM Gradient Boosting Machine

GPU Graphics Processing Unit

HRFLM Hybrid Random Forest with Linear Model

HDL High Density Lipoprotein

IA Intelligence Artificielle

IM Infarctus du Myocarde

IMC Indice de Masse Corporelle

INSP Institut National de la Santé Publique

IoT Internet of Things

KNN K-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

LDL Low Density Lipoprotein

LGBM Light Gradient Boosting Machine

LM Linear Model

LR Logistic Regression

LSTM Long Short-Term Memory

MANN Model Averaged Neural Network

MARS Multivariate Adaptive Regression Splines

MCC Matthews Correlation Coefficient

MCV Maladies Cardiovasculaires

ML Machine Learning

MLP Multi-Layer Perceptron

MPC Multi-Layer Perceptron

MRMR Minimum Redundancy Maximum Relevance

NB Naive Bayes

NN Neural Network

Noyau RBF Radial Basis Function Kernel

OMS Organisation Mondiale de la Santé

PCA Principal Component Analysis

PPO Proximal Policy Optimization

ReLU Rectified Linear Unit

RF Random Forest

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SAE Stacked AutoEncoder

SBS Sequential Backward Selection

SFS Sequential Forward Selection

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Oversampling TEchnique

SMOTE-NC Synthetic Minority Oversampling TEchnique Nominal Continuous

SVE Soft Voting Ensemble

SVM Support Vector Machine

SVR Support Vector Regression

TPR True Positive Rate

TPU Tensor Processing Unit

t-SNE t-Distributed Stochastic Neighbor Embedding

UCI University of California, Irvine

VN Vrai Négatif

VP Vrai Positif

XGBoost eXtreme Gradient Boosting

Introduction générale

De nos jours, les maladies cardiovasculaires (MCV) constituent un enjeu de santé publique majeur à l'échelle mondiale. Selon l'Organisation Mondiale de Santé (OMS) [129], elles ont causé la mort de **17,9** millions de personnes en 2019, soit **32%** de tous les décès dans le monde. Parmi ces décès, **85%** étaient dus aux infarctus du myocarde (IM) ou aux accidents vasculaires cérébraux (AVC). Un autre type de ces pathologies, connu sous le nom de maladies coronariennes ont causé **371 506** décès aux États Unis en 2022 [125]. En Algérie, elles représentent la première cause de mortalité avec un taux de **34%** par an, selon les chiffres de l'Institut National de la Santé Publique (INSP) [110]. Ces chiffres alarmants ont poussé les chercheurs à trouver des solutions pour prévenir et détecter ces maladies. Parmi ces travaux de recherche, certains se sont orientés vers l'intelligence artificielle et ses branches.

L'intelligence artificielle (IA) désigne la capacité des algorithmes intégrés aux systèmes et outils à apprendre à partir de données afin d'exécuter des tâches habituellement associées à l'intelligence humaine [130].

Parmi les techniques les plus utilisées en IA, l'apprentissage automatique ou machine learning (ML). Il permet aux ordinateurs d'apprendre à partir de données sans programmation directe [88]. Cela peut servir à trouver des solutions aux problèmes complexes dans divers domaines, tels que la médecine pour la prédiction des maladies, ou l'environnement en vue de la détection de catastrophes naturelles.

L'IA et le ML offrent des perspectives nouvelles et prometteuses dans le domaine médical. En effet, il est désormais possible de prédire les maladies cardiovasculaires à partir des données cliniques, ce qui facilite la prise de décision médicale, notamment en matière de prévention, et afin de prendre les mesures nécessaires et de traiter le problème avant qu'il soit trop tard. Ceci aide à réduire les cas du dernier stade, et les victimes d'une détection tardive de ce genre de maladies à fort impact.

Cependant, malgré ces remarquables avancées, il reste encore un long chemin à parcourir et des lacunes à combler. Il est important de procurer des résultats fiables aux médecins afin de leur permettre de détecter toute anomalie, en remplaçant les techniques traditionnelles de diagnostic, qui entraînent parfois des erreurs et des complications, par d'autres plus intelligentes et modernes. C'est la raison pour laquelle, les chercheurs ne lésinent point sur les efforts dans le but de fournir des systèmes de plus

en plus robustes et stables et tenter de répondre à la question suivante : l'IA peut-elle révolutionner la médecine en assurant une prédiction efficace de la survenue d'un problème de santé chez un patient ?

Ce mémoire s'inscrit dans cette dynamique, en développant un modèle de prédiction de MCV basé sur des algorithmes de machine learning en plus d'autres techniques d'optimisation des performances. Il est entraîné et évalué sur le dataset *"Cleveland"* et est capable de classifier les patients en deux (02) catégories : malade (1) ou sain (0). L'objectif est de concevoir un modèle efficace, fiable et compréhensible, pouvant être utilisé dans le cadre médical.

Ce travail est structuré en trois (03) chapitres :

— Le premier chapitre *« Généralités sur les maladies cardiovasculaires et l'IA »*, sera consacré à l'introduction à notre thème de recherche. Il comprendra des concepts généraux et des détails à ne pas négliger sur les MCV, l'IA, le ML et le DL, ainsi que quelques problèmes et solutions liés à ces deux derniers.

— Le second chapitre, intitulé *« État de l'art »*, contiendra des résumés d'articles soigneusement choisis, suivis de tableaux synthétiques et de discussions de ces travaux.

— Le troisième et dernier chapitre, nommé *« Approche proposée : Implémentation, Évaluation, et Discussion »*, sera dédié à la présentation, l'explication, et l'évaluation de l'approche proposée. Il sera conclu par une discussion et comparaison des résultats expérimentaux avec ceux de quelques travaux, précédemment présentés dans le deuxième chapitre et ayant utilisé le même dataset *"Cleveland"* avec une classification binaire.

En dernier lieu, viendra la conclusion générale qui présentera l'intérêt de l'approche, ses principales limites, et les perspectives de travaux futurs.

Chapitre 1

Généralités sur les maladies cardiovasculaires (MCV) et l'intelligence artificielle (IA)

1.1 Introduction

Les maladies cardiovasculaires (MCV) représentent les principales causes de décès dans le monde. Elles regroupent différentes pathologies qui affectent le cœur et les vaisseaux sanguins, comme l'infarctus du myocarde (IM), l'accident vasculaire cérébrale (AVC) et l'hypertension artérielle. La prévention et le diagnostic précoce sont cruciaux pour les systèmes de santé, ce qui nécessite des approches plus performantes.

Dans ce contexte, l'intelligence artificielle (IA) apparaît comme un outil puissant afin d'améliorer la prédiction et la prise en charge des maladies cardiovasculaires (MCV).

Ce chapitre propose une vue d'ensemble des maladies cardiovasculaires (MCV), en détaillant leurs types, leurs symptômes, les facteurs de risque, et quelques conseils de prévention. Il introduit également les concepts fondamentaux de l'IA et de ses branches, leurs problèmes, et leurs solutions ainsi que des exemples d'algorithmes et quelques techniques d'évaluation pour le cas d'une classification.

1.2 Généralités sur les maladies cardiovasculaires (MCV)

Cette section exposera les différents aspects généraux sur les maladies cardiovasculaires (MCV) pour mieux éclaircir notre thème.

1.2.1 Définition du système cardiovasculaire

Selon la Fédération Française de Cardiologie (FFC) : « Constitué du cœur et des vaisseaux (les artères et les veines), le système cardiovasculaire a pour fonction de distribuer aux organes, par le sang, l'oxygène et les nutriments indispensables à leur vie, tout en éliminant leurs déchets » [82].

En d'autres termes, il s'agit d'un appareil circulatoire assurant la circulation du sang dans le corps. Il permet le transport des gaz respiratoires, des nutriments, des hormones ainsi que l'élimination des déchets métaboliques afin de maintenir l'équilibre corporel [82].

1.2.2 Anatomie

L'anatomie du système cardiovasculaire comprend principalement les composants suivants :

a) **Le cœur** : C'est un organe musculaire central agissant comme une pompe permettant de faire circuler le sang dans tout le corps. Il est constitué de quatre (04) cavités : deux (02) oreillettes (droite et gauche) qui reçoivent le sang provenant des veines, et deux (02) ventricules (droit et gauche) qui propulsent le sang vers les artères [81]. Il pèse environ 350 g et est à peu près de la taille d'un poing fermé d'adulte [95]. La figure 1.1 présente le schéma anatomique du cœur.

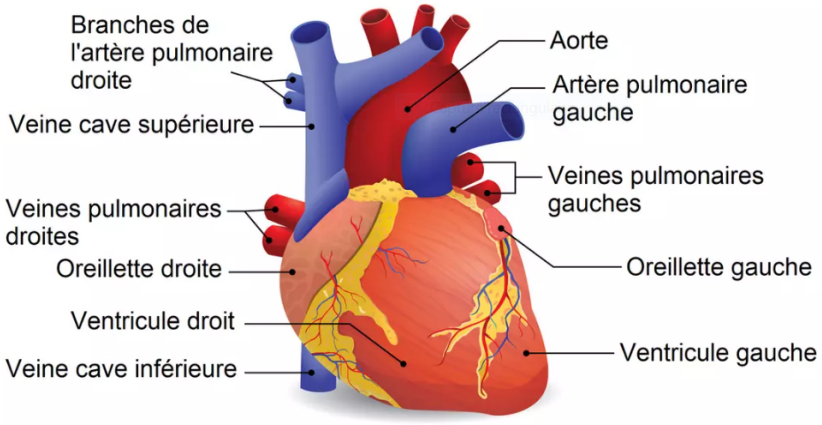


FIGURE 1.1 – Schéma anatomique du cœur [157].

b) **Les vaisseaux sanguins** : Un vaisseau sanguin est un petit conduit qui transporte le sang dans tout l'organisme [6]. Nous distinguons :

— *Les artères :* Elles transportent le sang du cœur vers les organes et les tissus. Ce sang est généralement riche en oxygène sauf pour l'artère pulmonaire qui transporte du sang désoxygéné [27].

— *Les veines :* Elles transportent le sang des organes vers le cœur. Le sang qui y circule est généralement pauvre en oxygène, à l'exception de la veine pulmonaire qui transporte le sang oxygéné des poumons vers le cœur [6].

— *Les capillaires :* Ils sont les vaisseaux sanguins les plus fins. Ils relient les artères aux veines. Ils permettent les échanges de gaz, de nutriments, et de déchets entre le sang et les cellules des tissus [86].

c) **La circulation sanguine** : C'est le mouvement continu du sang à travers les vaisseaux sanguins, propulsé par les contractions du cœur. Elle assure l'apport d'oxygène et de nutriments aux cellules de l'organisme et l'élimination des déchets métaboliques. Elle se divise en deux (02) circuits : la circulation pulmonaire et la circulation systémique [3]. La figure 1.2 met en lumière le schéma de la circulation sanguine.

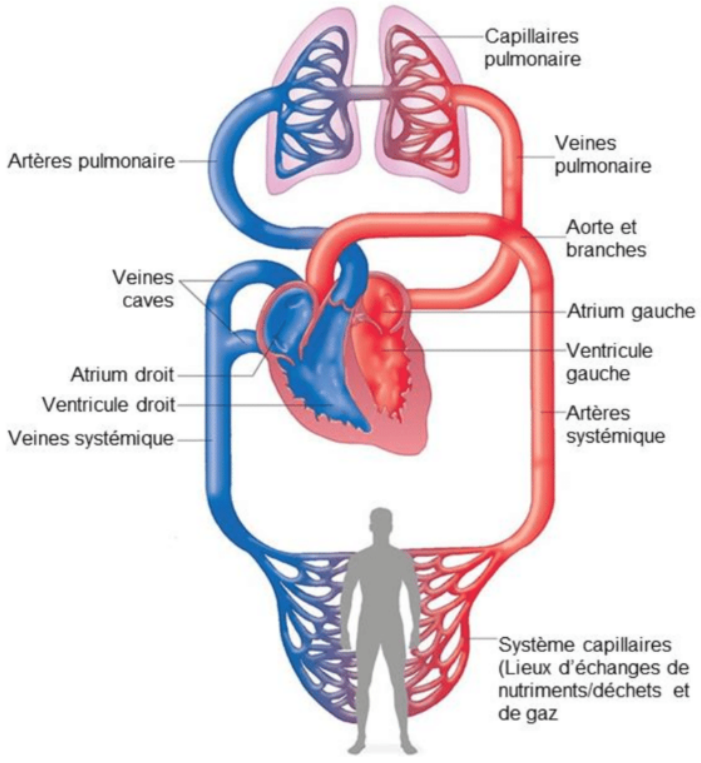


FIGURE 1.2 – Schéma de la circulation sanguine [108].

1.2.3 Fonctionnement du système cardiovasculaire

Après avoir présenté les composants du système cardiovasculaire, il est nécessaire de comprendre comment ils interagissent entre eux.

En effet, le cœur agit comme une pompe et assure la circulation du sang dans deux circuits distincts :

— *La circulation pulmonaire :* Après avoir reçu le sang désoxygéné de l'*oreillette droite*, le *ventricule droit* l'envoie aux poumons via l'artère pulmonaire afin de l'oxygéner.

— *La circulation systémique :* Le ventricule gauche reçoit le sang oxygéné de l'*oreillette gauche* et le propulse dans l'*aorte* pour le distribuer à l'ensemble du corps.

Donc, les artères transportent le sang oxygéné vers les tissus, où les capillaires assurent les échanges de gaz, de nutriments, et de déchets métaboliques au niveau cellulaire, tandis que les veines récupèrent le sang appauvri en oxygène et le ramènent au cœur [3].

1.2.4 Maladies cardiovasculaires (MCV)

Pour une meilleure compréhension de notre thème, il est nécessaire d'aborder les aspects généraux essentiels sur les maladies cardiovasculaires (MCV) qui seront présentés dans cette sous-section.

1.2.4.1 Définition

Les maladies cardiovasculaires (MCV) constituent un ensemble de troubles qui affectent le système cardiovasculaire causant ainsi, des dommages dans différentes parties du corps humain [129].

1.2.4.2 Types de maladies cardiovasculaires (MCV)

Les maladies cardiovasculaires (MCV) peuvent se catégoriser en plusieurs types dont nous citons :

a) **Les maladies coronariennes** : Ce sont des pathologies affectant les artères du cœur, causées par un rétrécissement ou une obstruction de ces dernières avec une insuffisance d'irrigation du muscle cardiaque en sang oxygéné. Cela est souvent dû à l'accumulation de plaques d'athérome (amas graisseux), ce qui peut entraîner notamment l'angine de poitrine (une douleur thoracique violente causée par un manque d'oxygène au niveau du cœur), ainsi que L'Infarctus du Myocarde (IM) ou la crise cardiaque (l'occlusion totale d'une artère coronaire qui provoque la mort cellulaire d'une zone du muscle cardiaque) [8]. La figure 1.3 illustre une artère coronaire partiellement obstruée à cause du dépôt de plaques, ce qui désigne un signe d'une maladie coronarienne.

FIGURE 1.3 – Maladies coronariennes [8].

b) **Les maladies cérébrovasculaires** : Ce sont des problèmes affectant la circulation du sang dans le cerveau. Cela est souvent dû quand un vaisseau sanguin est bouché ou rompu, empêchant l'oxygène d'arriver au cerveau, ce qui peut endommager les cellules nerveuses. Ces pathologies incluent notamment les Accidents Vasculaires Cérébraux (AVC) ischémiques (vaisseau cérébrale bouché) ou hémorragiques (vaisseau rompu) et d'autres affections comme les anévrismes ou les anomalies vasculaires [42]. La figure 1.4 met en lumière une occlusion artérielle cérébrale.

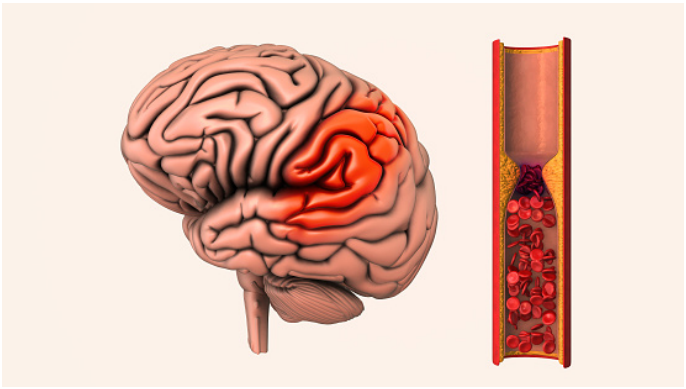


FIGURE 1.4 – Maladies cérébrovasculaires [2].

c) **L'hypertension artérielle** : C'est une condition où la pression du sang dans les artères est trop élevée. Cela force le cœur à fournir un effort supplémentaire pour assurer la circulation sanguine, ce qui peut endommager les artères et affecter des organes vitaux (cœur, cerveaux, etc). Elle est caractérisée par une pression sanguine atteignant ou dépassant 140 mmHg pour la systolique (pression lors de la contraction du cœur) et/ou 90 mmHg pour la diastolique (pression lors de sa relaxation) [131]. La figure 1.5 montre que le fonctionnement du cœur est étroitement lié aux fluctuations de la pression artérielle.

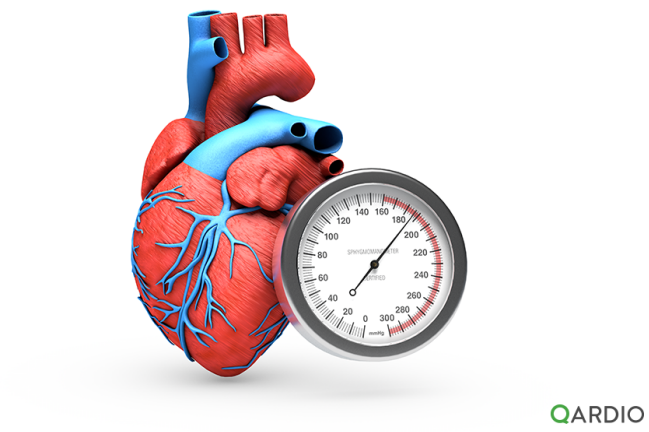


FIGURE 1.5 – Hypertension artérielle [4].

d) **L'insuffisance cardiaque** : C'est un syndrome qui représente l'incapacité du cœur à assurer un débit sanguin suffisant pour répondre aux besoins des organes. Cela peut être dû à une faiblesse ou une rigidification du muscle cardiaque [12, 96]. La figure 1.6 montre la différence entre un cœur sain et une insuffisance cardiaque.

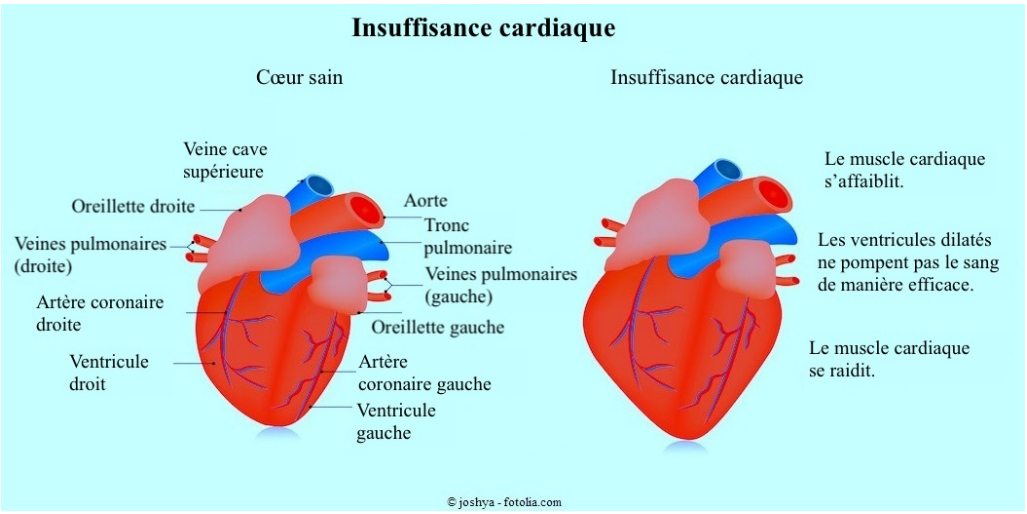


FIGURE 1.6 – Insuffisance cardiaque [12].

e) **Les arythmies cardiaques** : Ce sont des troubles du rythme cardiaque, il s'agit de perturbations dans la fréquence ou la régularité des battements du cœur, qui peuvent être trop rapides (tachycardie), trop lents (bradycardie) ou irréguliers (fibrillation auriculaire) entraînant des difficultés dans la circulation sanguine [119]. La figure 1.7 présente un cœur normal ainsi qu'une fibrillation auriculaire.

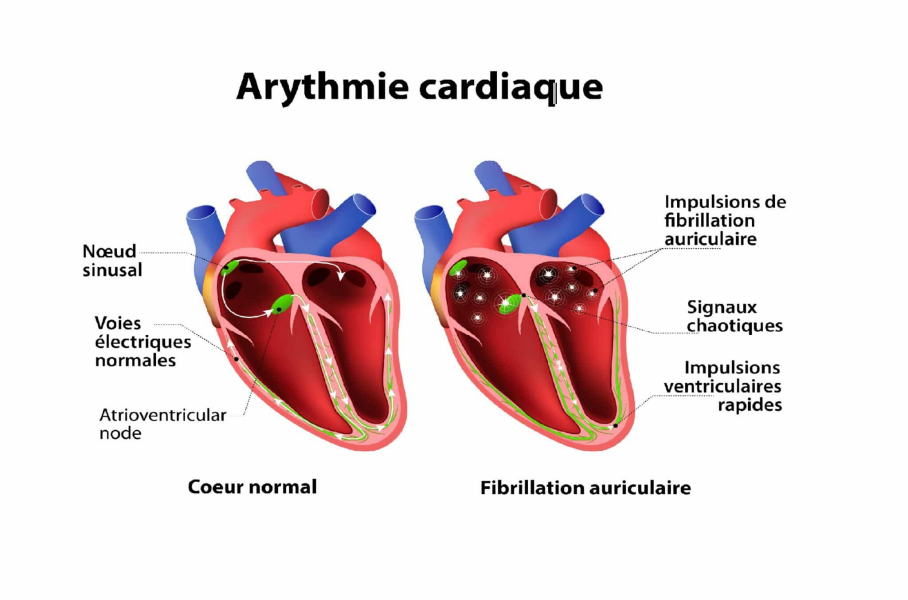


FIGURE 1.7 – Arythmie cardiaque [140].

f) **Les maladies des artères périphériques** : Ce sont des complications affectant les artères situées en dehors du cœur et du cerveau, surtout celles des jambes mais elles peuvent aussi concerner les bras, l'abdomen ou la tête. Elles sont dues à un rétrécissement ou l'obstruction des artères souvent en raison de l'athérosclérose (accumulation de dépôts graisseux sur les parois artérielles). Cela provoque des douleurs lors de l'effort [83, 126]. La figure 1.8 illustre le développement de cette maladie du normal au thrombus.

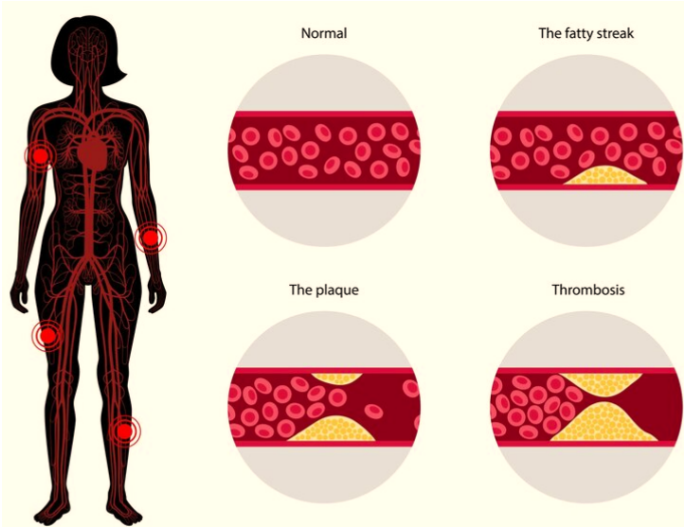


FIGURE 1.8 – Maladies des artères périphériques [84].

g) **Les malformations cardiaques congénitales** : Ce sont des défauts dans la structure, voire le fonctionnement du cœur, qui sont présents depuis la naissance. Ces anomalies surviennent à la suite d'un développement cardiaque anormal pendant la grossesse, qui peuvent être cyanogènes (mélange de sang oxygéné et non oxygéné ce qui provoque une cyanose (peau bleutés)) ou non cyanogènes (troubles circulatoires sans cyanose) [35]. La figure 1.9 représente une anomalie cardiaque congénitale.

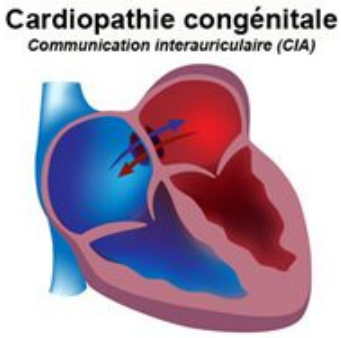


FIGURE 1.9 – Malformations cardiaques congénitales [1].

h) **Les cardiomyopathies** : Ce sont des maladies qui touchent le muscle cardiaque et qui sont différentes des autres troubles. Elles se catégorisent en trois (03) types principaux en fonction des anomalies anatomopathologiques observées, dont nous avons la cardiomyopathie dilatée (cœur élargi et faible), hypertrophique (épaississement anormal du muscle) et restrictives (cœur rigide et peu souple) [153], comme le montre la figure 1.10.

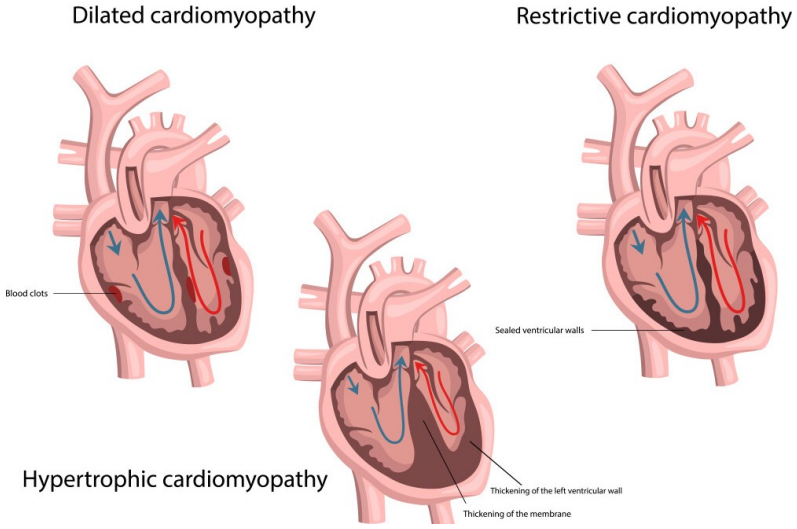


FIGURE 1.10 – Cardiomyopathies [5].

i) **Les maladies valvulaires** : Ce sont des problèmes affectant le dysfonctionnement des valves cardiaques (mitrale, aortique, tricuspide, pulmonaire) qui peuvent être trop étroites ou mal fermées, provoquant des fuites, ce qui perturbe la circulation sanguine normale [21, 24].

La figure 1.11 révèle le rétrécissement aortique calcifié.

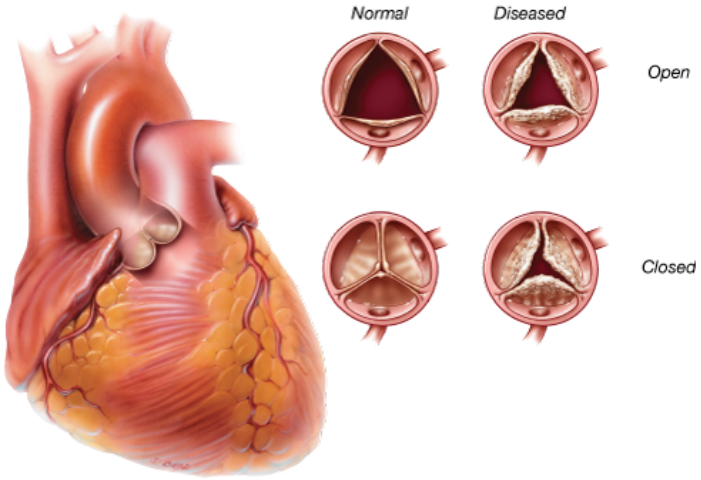


FIGURE 1.11 – Maladies valvulaires [24].

j) **Les maladies veineuses et thrombo-emboliques** : Ce sont des complications qui touchent les veines et la circulation sanguine, liées à la formation de caillots (thrombus) ou à un mauvais retour veineux. Cela provoque des problèmes graves comme une thrombose veineuse profonde (caillot sanguin qui se forme dans une veine profonde, généralement dans les jambes) ou une embolie pulmonaire (caillot sanguin qui bloque une artère des poumons), ce qui est très dangereux [117, 138]. La figure 1.12 montre l'évolution d'un caillot sanguin dans les veines des jambes.

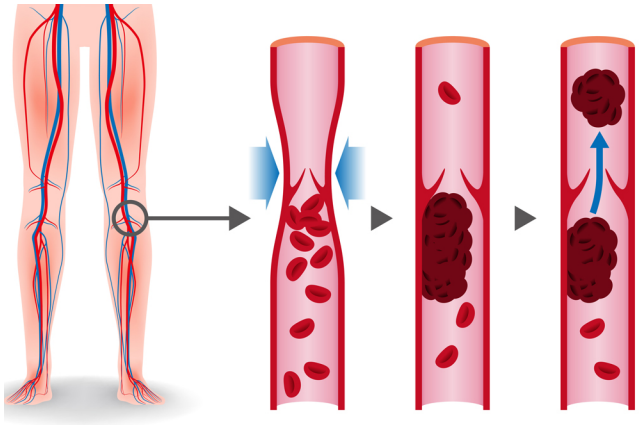


FIGURE 1.12 – Maladies veineuses et thrombo-emboliques [138].

1.2.4.3 Facteurs de risque

Les maladies cardiovasculaires (MCV) résultent fréquemment de la combinaison de plusieurs facteurs de risque, permettant de développer des complications cardiaques et vasculaires. Certains de ces facteurs échappent à notre contrôle, tandis que d'autre peuvent être maîtrisés ou limités par l'adoption d'un mode de vie plus sein et plus équilibrés. Parmi ces facteurs [52, 121], nous citons :

a) **Les facteurs de risque non modifiables** : Ce sont des paramètres inchangeables et indépendants du mode de vie et qui jouent un rôle dans la prédisposition aux maladie cardiovasculaires (MCV), tels que :

- *L'âge :* La fréquence des maladies cardiovasculaires (MCV) s'élève de manière très rapide avec l'âge. La majorité des décès dus aux MCV concernent essentiellement les personnes âgées, y compris les centenaires. Cela en raison du vieillissement naturel des vaisseaux sanguins et du cœur.
- *Le sexe :* Certaines affections touchent différemment les hommes et les femmes, sachant que les hommes avant 55 ans sont plus vulnérables aux maladies cardiovasculaires contrairement aux femmes chez lesquelles le risque s'accroît après la ménopause en raison de diminution des hormones protectrices.
- *Les facteurs génétiques :* Avoir des membres de la famille ayant souffert des MCV, influence la susceptibilité d'un individu à développer ces maladies.

b) **Les facteurs de risque modifiables** : Ce sont des paramètres changeables et dépendants du mode de vie qui peuvent être gérés pour réduire le risque de ce genre de pathologies, comme :

— **Comportements et habitudes**

Ce genre de maladies est lié à des routines et des choix de vie qui peuvent accroître le risque. Parmi ces facteurs [122], nous citons :

- *Le tabagisme :* La fumée d'une cigarette présente un risque très important, à cause des substances chimiques et toxiques contenues dans ses composants qui endommagent les parois artérielles, ce qui facilite l'apparition des problèmes cardiaques comme l'athérosclérose et les caillots sanguins.
- *La malnutrition :* La surconsommation des trois (3) blancs (graisse, sel et sucre) favorise l'augmentation du cholestérol, de la pression artérielle, et aussi du diabète.
- *La sédentarité :* Passer la plupart du temps avec un manque d'activité physique fragilise le cœur et provoque des complications comme l'obésité.

- *L'obésité et le surpoids :* L'accumulation incontrôlée de graisse corporelle, souvent identifiée par un indice de masse corporelle (IMC), entraîne d'autres risques comme l'hypertension, l'inflammation des vaisseaux sanguins, le diabète, etc. Cela favorise les maladies cardiovasculaires.
- *La consommation excessive d'alcool :* En plus d'autres problèmes, la surconsommation d'alcool a un impact négatif sur la santé en favorisant le risque de plusieurs maladies cardiovasculaires comme les arythmies cardiaques.
- *Le stress chronique :* La pression excessive sur le cœur entraîne une augmentation des taux de cortisol et d'adrénaline ainsi qu'une aggravation des troubles cardiovasculaires due par exemple à l'hypertension.

— **Maladies et troubles médicaux associés**

Les maladies cardiovasculaires (MCV) sont influencées aussi par des pathologies et des complications médicales qui augmentent ce risque [122], telles que :

- *L'hypertension artérielle :* La pression artérielle élevée force le cœur à pomper le sang d'une façon anormale, ce qui le fatigue. Elle abîme également les artères et représente donc, un risque de développer des maladies cardiovasculaires.
- *Le diabète :* Dans le cas d'une hyperglycémie, cette maladie chronique affecte les vaisseaux sanguins et les nerfs qui assurent la fonction cardiaque, ce qui complique la situation et provoque par exemple des AVC voire des crises cardiaques.
- *L'hypercholestérolémie :* Elle représente un taux élevé de cholestérol dans le sang. Une accumulation importante de « mauvais cholestérol » LDL (Low Density Lipoprotein) déclenche le risque d'une athérosclérose, et réduit la circulation sanguine.
- *Le syndrome métabolique :* C'est un ensemble de complications physiologiques et métaboliques dont trois (03) facteurs parmi cinq (05) chez un même individu, notamment l'obésité, l'hypertension, l'hyperglycémie, hypertriglycéridémie et la baisse de HDLc (High Density Lipoprotein cholesterol) entraînent des problèmes cardiovasculaires [23].
- *L'apnée du sommeil :* Cette condition représente un trouble respiratoire chronique dans lequel une personne cesse de respirer périodiquement pendant son sommeil, ce qui provoque une baisse de l'oxygénation du sang ainsi que d'autres problèmes cardiovasculaires [101].
- *La fibrillation atriale ou la fibrillation auriculaire :* Elle représente le premier facteur de risque d'origine cardiaque, avec un risque multiplié par quatre (04). Elle se caractérise par des battements du cœur irréguliers et souvent très rapides [115].

c) **Les facteurs de risque environnementaux et sociaux**

Les maladies cardiovasculaires sont impactées par divers facteurs environnementaux et sociaux qui peuvent accroître leurs développement [45], tels que :

- *La pollution atmosphérique :* Les contaminants de l'air (dioxyde d'azote, monoxyde de carbone, etc) pénètrent profondément dans les poumons et passent dans la circulation sanguine, cela provoque l'inflammation des vaisseaux sanguins ainsi que divers types de maladies cardiovasculaires (MCV).
- *L'eau potable :* Il est prouvé que plusieurs métaux trouvés dans l'eau potable comme le plomb et l'arsenic, peuvent causer des problèmes cardiaques ou aggraver les symptômes cardiovasculaires.
- *Les accidents de chaleur excessive :* Dans le cas de coup de chaleur, il est fortement possible d'avoir des dommages sévères et permanents aux organes essentiels (crise cardiaque, choc hypovolémique, etc), ce qui influence la santé cardiovasculaire.
- *Les facteurs socio-économiques :* L'incidence des MCV est plus élevée chez la population défavorisée aux plan matériel et social, ou chez les populations à faible revenu, et ce, dans les pays développés comme dans les pays en développement.
- *L'urbanisation et la mondialisation :* Ces événements influencent les comportements quotidiens en encourageant la sédentarité, l'alimentation industrialisée, la pollution, etc, et favorisent ainsi le développement de maladies cardiovasculaires (MCV).

1.2.4.4 Symptômes

Les symptômes des maladies cardiovasculaires (MCV) varient selon le type. Dans le tableau 1.1, nous citerons les signes les plus courants qui aident à identifier la maladie afin de prendre les mesures nécessaires [90, 102].

Catégorie	**Symptômes**
Symptômes généraux communs	- Fatigue inhabituelle et faiblesse inattendue. - Essoufflement ou respiration difficile. - Vertiges ou étourdissements. - Certains types de douleurs.
Symptômes de maladies coronariennes	- Angine de poitrine. - Palpitations. - Nausées et vomissements. - Sensation d'indigestion.

Symptômes de maladies cérébrovasculaires	- Difficulté cognitive (prise de décision) et motrice (problèmes de mouvement). - Perte de sensation au niveau du visage, d'un bras ou d'une jambe, souvent d'un seul côté du corps. - Perte de vision par hasard. - Maux de tête excessifs.
Symptômes de l'hypertension artérielle	- Souvent asymptomatique, mais certains signes comme les bourdonnements d'oreilles, les saignements du nez, etc, peuvent alerter.
Symptômes d'une insuffisance cardiaque	- Œdème (gonflement des pieds, des chevilles voire même des jambes). - Toux persistante sèche ou avec des glaires. - Prise de poids rapide liée à la rétention de liquide. - Perte d'appétit.
Symptômes d'arythmie	- Palpitations. - Évanouissement et syncope (perte de conscience). - Transpiration excessive. - Gêne ou blocage dans la gorge.
Symptômes des maladies des artères périphériques	- Claudication. - Ulcères ou lésions sur la peau qui tardent à guérir. - Pieds froids et bleutés. - Pouls faibles ou absents dans les jambes ou les pieds.
Symptômes des malformations cardiaques	- Souvent asymptomatiques, mais certains signes comme les infections pulmonaires à répétition et le retard de croissance chez les enfants peuvent en signaler l'éventualité.
Symptômes des maladies veineuses et thrombo-emboliques	- Apparition des varices. - Crampes dans les jambes. - Peau sèche, rouge ou irritée. - Sensation de chaleur locale.

TABLE 1.1 – Symptômes des maladies cardiovasculaires (MCV) [90, 102].

L'apparition de ces symptômes nécessite un examen médical pour confirmer si le patient est atteint d'une maladie cardiovasculaire ou non, car cela ne signifie pas toujours l'existence de la pathologie.

Il est recommandé aussi de consulter immédiatement un professionnel de santé pour prévenir les complications graves.

1.2.4.5 Quelques conseils de prévention des maladies cardiovasculaires (MCV)

Dans le but de prévenir les maladies cardiovasculaires, il est important d'adopter un mode de vie sain et de prendre soin de sa santé au quotidien, en visant à éliminer ou à réduire autant que possible les facteurs de risques modifiables, notamment :

a) **Comportements, habitudes, maladies et troubles médicaux associés**

L'intégration de bonnes habitudes a un impact positif pour la prévention des maladies cardiovasculaires (MCV) [40], telles que :

- L'abandon du tabac et de consommation d'alcool.
- Le suivi d'un plan de régime alimentaire équilibré, riche en fruits, légumes et fibres.
- Programmation du temps pour les activités physiques régulières (30 minutes par jours au moins cinq (5) jours par semaine), comme la marche.
- La gestion du stress, à travers des techniques de relaxation comme le yoga.
- Un suivi médical régulier pour dépister ou traiter l'hypertension, l'hyperglycémie, hypercholestérolémie, etc.

b) **Préventions environnementales et sociales**

Afin de réduire les facteurs de risque environnementaux et sociaux, il est essentiel de [45] :

- Réduire l'exposition à la circulation des véhicules et à la pollution de l'air ambiant.
- Boire de l'eau pure.
- Promouvoir des systèmes de soins actifs d'avertissement et de réponse à la chaleur qui aident à identifier ce risque à l'avance pour faire le nécessaire.
- Réguler la quantité de sel et de sucre dans les produits de l'industrie alimentaire.
- Éviter l'empoisonnement au monoxyde de carbone en aérant bien votre maison et en vérifiant vos appareils de chauffage.
- Etc.

En adoptant ces mesures préventives, chacun peut réduire, voire éliminer significativement le risque d'avoir une maladie cardiovasculaire (MCV) et améliorer sa qualité de vie.

Étant donné les risques majeurs de ce genre de pathologies sur la santé publique, l'IA permet d'intégrer diverses techniques pour un diagnostic plus précis et une prévention préliminaire.

1.3 Généralités sur l'IA

Cette section sera dédiée à l'exploration des concepts de base de l'intelligence artificielle et de l'apprentissage automatique, y compris l'apprentissage profond.

1.3.1 Intelligence artificielle

L'intelligence artificielle (IA) puise ses origines dans les travaux d'Alan Turing durant les années 1940, puis a été introduite en 1956, lors de la conférence de Dartmouth par John McCarthy [156].

Il s'agit d'un ensemble de théories et de techniques qui vise à créer des machines capables de réaliser des tâches habituellement associées à l'intelligence humaine [39].

La figure 1.13 révèle la relation entre l'IA, l'apprentissage automatique (ML), et l'apprentissage profond (DL).

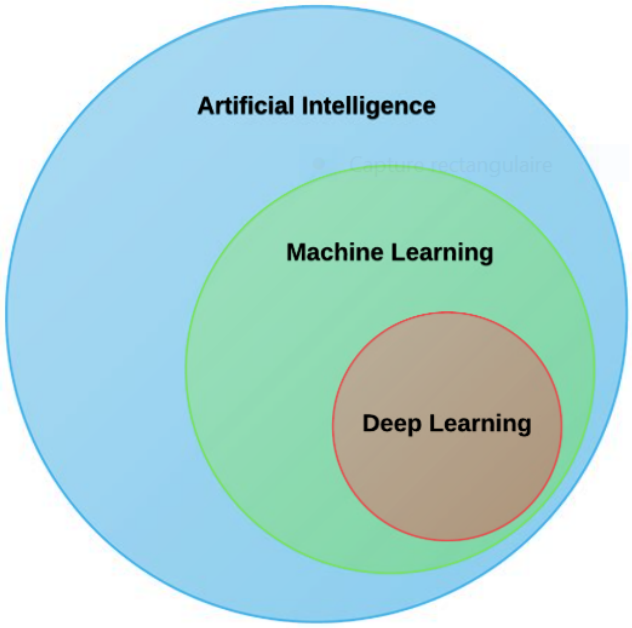


FIGURE 1.13 – Relation entre IA, apprentissage automatique (ML), et apprentissage profond (DL) [128].

1.3.2 Apprentissage automatique

L'apprentissage automatique ou Machine Learning (ML) est une branche de l'IA qui permet aux ordinateurs d'imiter la manière dont les humains apprennent grâce à des algorithmes afin de faire des tâches auxquelles ils n'ont pas été explicitement programmés [91].

1.3.3 Types de ML

Le ML est classé en plusieurs catégories [31] :

1.3.3.1 Apprentissage supervisé

L'algorithme est programmé sur un ensemble de données étiquetées. C'est-à-dire, chaque entrée est associée à une sortie connue.

Il existe des algorithmes de :

— **Classification :** Ces algorithmes prédisent des catégories.

Exemples : Régression logistique (LR) [60], machine à vecteurs de support (SVM) [77], forêt aléatoire (RF) [62], K plus proches voisins (KNN) [57], Naïve Bayes (NB) [77], eXtreme Gradient Boosting (XGBoost) [150], etc.

— **Régression :** Ces algorithmes prédisent une valeur continue.

Exemples : Régression linéaire [143], régression ridge [26], régression LASSO [76], gradient boosting [54], etc.

1.3.3.2 Apprentissage non-supervisé

L'algorithme est entraîné sur des données non-étiquetées pour extraire des structures ou des tendances cachées sans l'intervention humaine.

Il existe des algorithmes de :

— **Clustering :** Ces algorithmes regroupent les données similaires.

Exemples : K-Means [55], DBSCAN [66], Hierarchical Clustering [111], etc.

— **Réduction de dimensionnalité :** Ces algorithmes visent à transformer les données en un espace de plus faible dimension.

Exemples : PCA [50], t-SNE [106], etc.

1.3.3.3 Apprentissage semi-supervisé

C'est une technique qui combine l'apprentissage supervisé et non-supervisé en exploitant un petit ensemble de données étiquetées et un grand ensemble de données non-étiquetées.

Exemples : Self-Training [59], Graph-Based Algorithms [25], Semi-Supervised SVM [104], etc.

1.3.3.4 Apprentissage par renforcement

L'algorithme apprend grâce aux interactions avec un environnement en recevant des retours comme récompenses ou punitions.

Exemples : Q-Learning [97], Deep Q-Network (DQN) [56], Proximal Policy Optimization (PPO) [72], Actor-Critic Methods [63], etc.

1.3.4 Quelques algorithmes de ML

Il existe plusieurs algorithmes d'apprentissage automatique (ML). Dans cette partie, nous présentons les plus utilisés.

1.3.4.1 Arbre de décision (DT)

L'arbre de décision (DT) est une méthode d'apprentissage supervisé qui divise les données en branches en fonction des valeurs des caractéristiques, créant ainsi une structure arborescente qui aide à classifier les instances. Bien que les modèles d'arbres de décision soient interprétables, ils sont sujets au sur-apprentissage, en particulier avec des ensembles de données de haute dimension [134].

La figure 1.14 présente un exemple d'un arbre de décision.

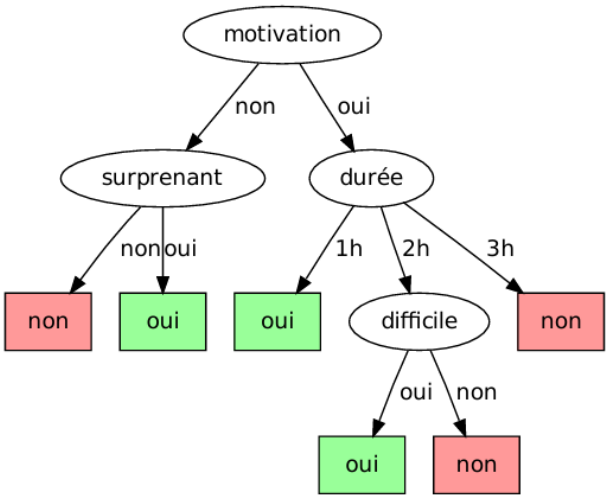


FIGURE 1.14 – Exemple d'un arbre de décision (DT) pour la question " Cette présentation est-elle intéressante ? " [33].

1.3.4.2 Forêt aléatoire (RF)

C'est une méthode d'apprentissage ensembliste supervisé qui construit plusieurs arbres de décision, et les fusionne pour produire une prédiction plus précise et stable. Elle fonctionne bien avec des données de haute dimension et peut gérer efficacement les classes déséquilibrées. Les RFs ont été utilisées dans la prédiction des maladies cardiaques par plusieurs chercheurs [134], notamment J.Nageswara Rao et al (2021) [62], qui ont souligné sa robustesse et sa fiabilité.

La figure 1.15 suivante montre une structure typique d'une forêt aléatoire.

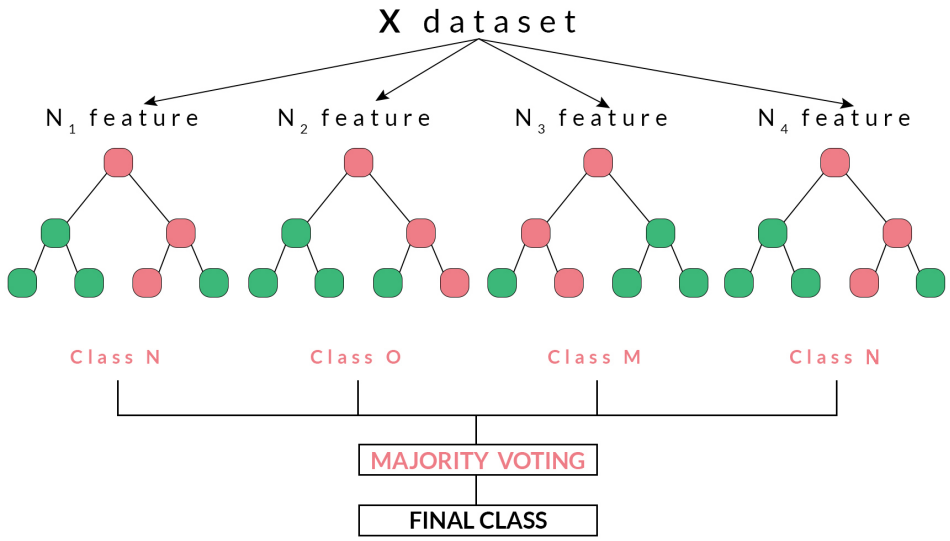


FIGURE 1.15 – Forêt aléatoire [11].

1.3.4.3 Machine à vecteurs de support (SVM)

L'algorithme SVM (Support Vector Machine) est un classificateur puissant qui fonctionne en trouvant un hyperplan optimal pour séparer les données en différentes classes. Il est capable à gérer des ensembles de données de haute dimension [134]. La figure 1.16 permet de visualiser les concepts clés d'une SVM (hyperplan, marge, et vecteurs de support).

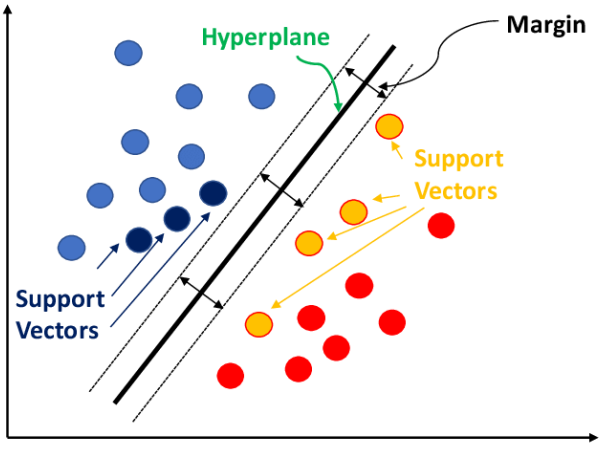


FIGURE 1.16 – Machine à vecteurs de support (SVM) [123].

1.3.4.4 K plus proches voisins (KNN)

L'un des algorithmes d'apprentissage automatique supervisé les plus largement utilisés pour les problèmes de classification, la reconnaissance de motifs, et la régression. Cet algorithme permet d'identifier les voisins des données en utilisant la distance entre les points de données [17]. La figure 1.17 illustre le fonctionnement du KNN pour une classification ternaire. Chaque point sera classé comme rouge (1), bleu (2) ou vert (3) en fonction de sa distance par rapport à chaque groupe [145].

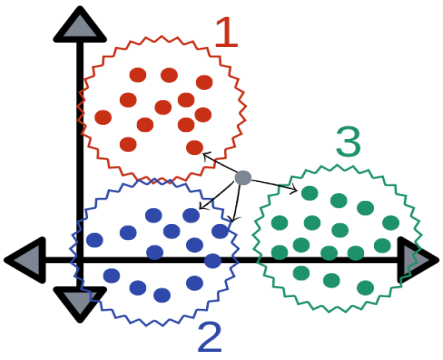


FIGURE 1.17 – K plus proches voisins (KNN) [145].

1.3.4.5 Régression logistique (LR)

L'un des algorithmes les plus simples en apprentissage automatique supervisé, utilisé largement pour les problèmes de classification binaire, y compris la prédiction des maladies cardiaques. La régression logistique fonctionne en estimant la probabilité d'un résultat binaire (par exemple, la présence ou l'absence d'une maladie cardiaque) en se basant sur des caractéristiques d'entrée. Malgré sa simplicité, elle s'est avérée efficace dans certains cas, notamment lorsque les données sont linéairement séparables [134]. La figure 1.18 met en lumière une courbe sigmoïde d'un modèle de LR.

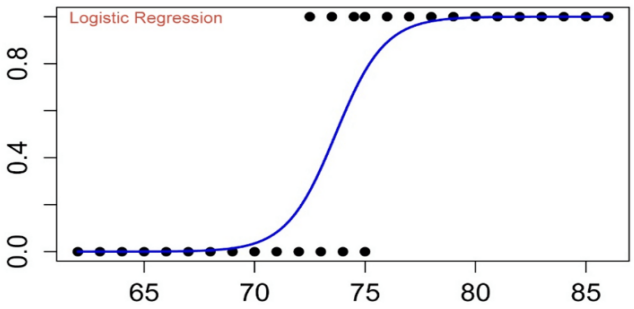


FIGURE 1.18 – Régression logistique (LR) [14].

1.3.4.6 Naïve Bayes (NB)

NB est un algorithme d'apprentissage automatique supervisé probabiliste. Il permet de résoudre des problèmes de classification et est basé sur le théorème de Bayes, en considérant que les variables sont conditionnellement indépendantes, étant donné l'étiquette de la classe. La formule de cette supposition peut être simplifiée comme suit [133, 147] :

$$P(Y \mid \mathbf{X}) = \frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(\mathbf{X})} \tag{1.1}$$

Où : X représente les "d" attributs, Y désigne la classe, et P(X) c'est la probabilité marginale de X.

La figure 1.19 illustre l'utilisation du théorème de Bayes (probabilité conditionnelle) pour une classification ternaire des formes géométriques et de leurs couleurs (carré en vert, cercle en rouge, et triangle en bleu).

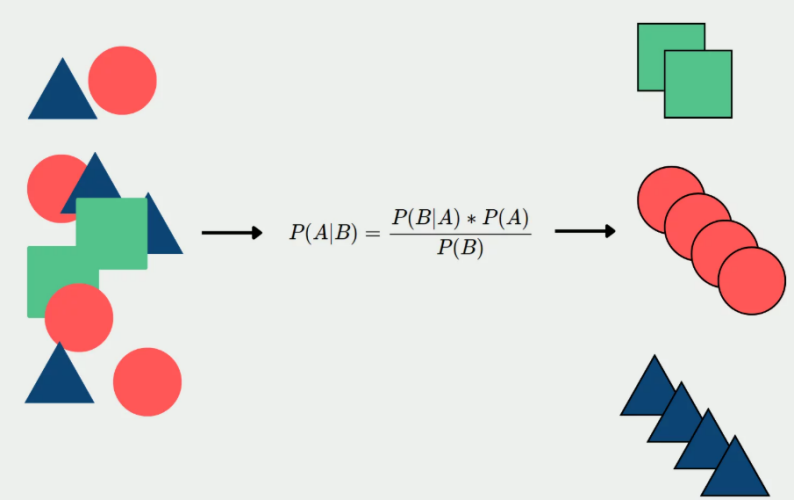


FIGURE 1.19 – Naïve Bayes (NB) [7].

1.3.5 Apprentissage profond : une branche avancée du ML

L'apprentissage profond ou le deep learning (DL) est un sous-ensemble du ML qui utilise des réseaux de neurones multicouches, appelés réseaux de neurones profonds, pour simuler la capacité de prise de décision complexe du cerveau humain [92].

1.3.6 Architectures de l'apprentissage profond

L'apprentissage profond (DL) peut adopter de nombreuses architectures différentes, variant selon le domaine d'application.

1.3.6.1 Perceptrons multicouches (MLP)

Le perceptron multicouche est un type de réseau neuronal artificiel organisé en plusieurs couches. Il doit avoir au moins trois (03) couches : une couche d'entrée, au moins une couche cachée, et une couche de sortie [112], comme le montre la figure 1.20.

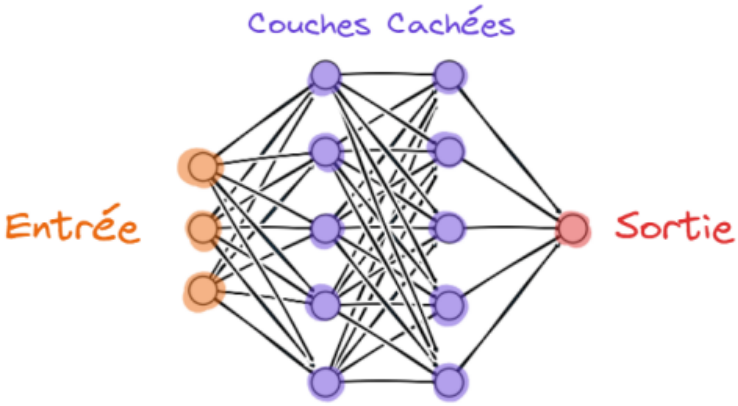


FIGURE 1.20 – Perceptrons multicouches [144].

1.3.6.2 Réseaux de neurones convolutifs (CNN)

Un réseau de neurones convolutifs est une architecture de réseau neuronal profond. Il est particulièrement utilisé afin de classifier des images [36].

La figure 1.21 illustre un schéma simplifié de l'architecture d'un CNN.

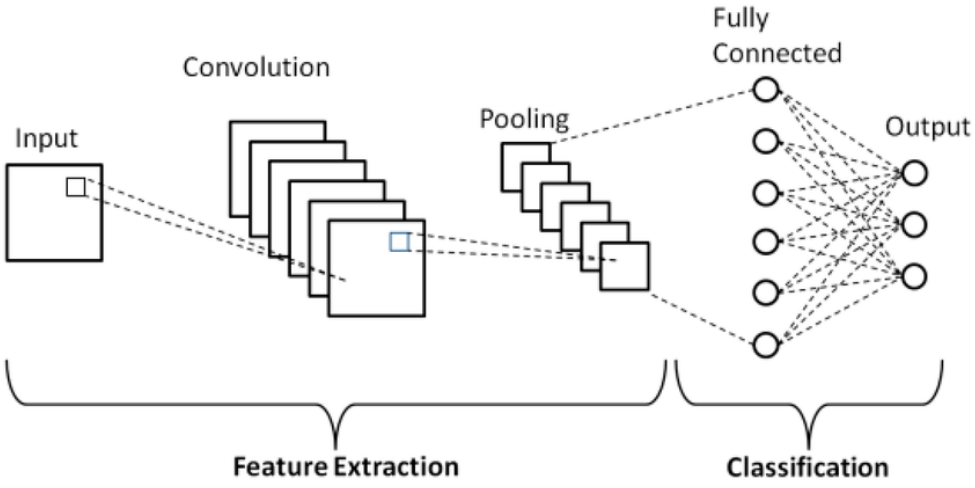


FIGURE 1.21 – Schéma simplifié de l'architecture d'un CNN [136].

1.3.6.3 Réseaux de neurones récurrents (RNN) et LSTM

Les réseaux de neurones récurrents sont une classe de modèles d'apprentissage automatique conçus pour traiter des données séquentielles, où les sorties dépendent des entrées précédentes [151], comme le montre la figure 1.22.

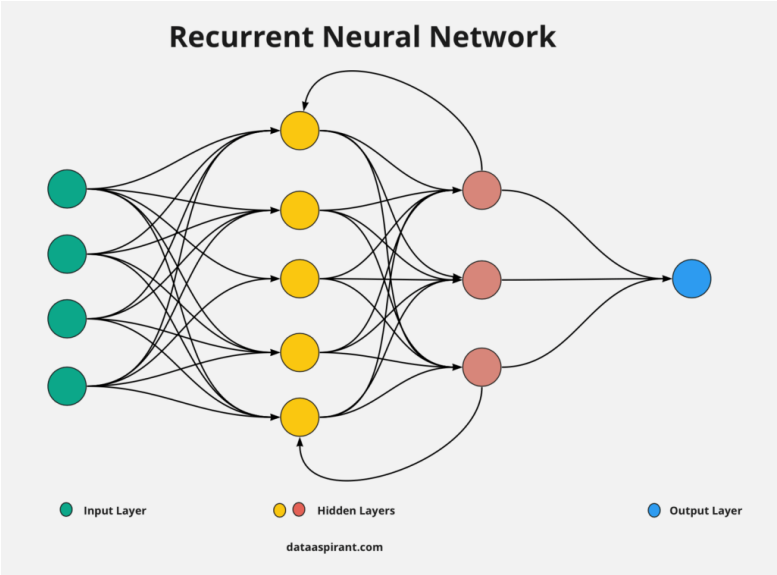


FIGURE 1.22 – Réseau de neurones récurrent [37].

Les réseaux LSTM (Long Short-Term Memory) sont une variante de RNN qui introduisent des mécanismes de portes pour gérer efficacement les informations sur de longues séquences [15].

La figure 1.23 comporte un schéma représentatif de LSTM.

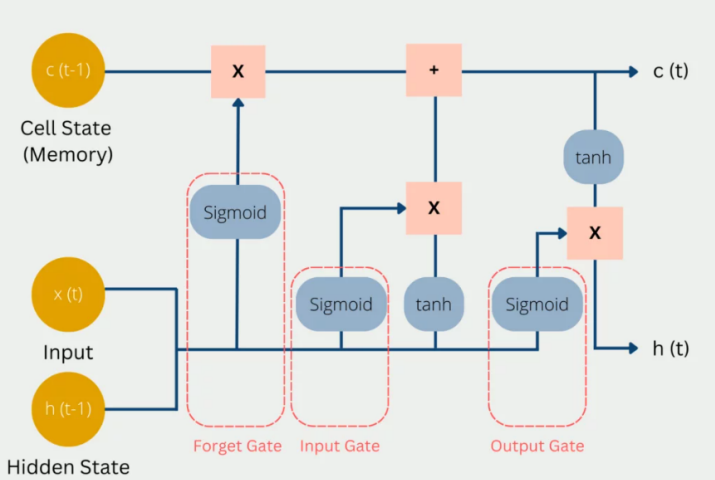


FIGURE 1.23 – Schéma représentatif de LSTM [99].

1.3.6.4 Transformeurs

Les transformeurs sont des architectures d'apprentissage profond introduites en 2017 [46]. Ils sont principalement utilisés dans le traitement automatique des langues et servent de base aux grands modèles de langage. Ils peuvent aussi traiter les images, les vidéos, ou le son, parfois simultanément [116]. La figure 1.24 présente l'architecture d'un transformeur.

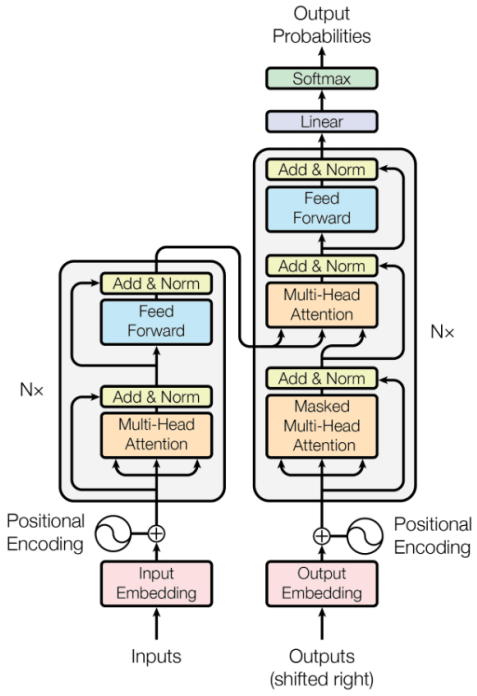


FIGURE 1.24 – Architecture d'un transformeur [46].

1.3.7 Quelques problèmes du DL et leurs solutions

Malgré ses performances remarquables, il est important de préciser que l'apprentissage profond présente quelques problèmes, comme montré dans le tableau 1.2, ainsi que leurs solutions [48].

Problèmes et défis du DL	**Solutions et avancées technologiques**
Besoin de grandes quantités de données.	- Utilisation des techniques de régularisation (dropout, batch normalization, etc).
Coût computationnel élevé.	- Optimisation et accélération des calculs (GPU, TPU, quantization).
Sur-apprentissage (overfitting)	- Apprentissage par transfert (Transfer Learning).
Interprétabilité des modèles.	- AutoML et optimisation des hyperparamètres.

TABLE 1.2 – Quelques problèmes du DL et leurs solutions [48].

1.3.8 Évaluation des performances des modèles de l'IA (classification)

Les modèles de l'IA (ML ou DL) passent par une étape très importante qui est l'évaluation des performances. Cette étape diffèrent selon la catégorie de l'apprentissage (classification, régression, clustering, etc).

Dans notre recherche, nous nous concentrons sur l'évaluation des modèles de classification en utilisant la matrice de confusion ainsi que d'autres métriques.

1.3.9 Matrice de confusion

La matrice de confusion est un tableau qui permet de citer pour chaque classe le nombre de bonnes et de mauvaises prédictions par rapport aux vraies étiquettes afin d'évaluer les performances des modèles de classification utilisés. Elle s'appuie sur quatre (04) types de résultats, notamment [133] :

— **Vrai Positif (VP) :** lorsque le modèle prédit "positif" et qu'il l'est réellement.

— **Faux Négatif (FN) :** lorsque le modèle prédit "négatif", alors qu' il est en réalité "positif".

— **Faux Positif (FP) :** lorsque le modèle prédit "positif", alors qu'il est en réalité "négatif".

— **Vrai Négatif (VN) :** lorsque le modèle prédit "négatif" et qu'il l'est réellement.

Le tableau 1.3 ci-après illustre une matrice de confusion pour un problème de classification binaire.

		Classe prédite	
		+	−
Classe réelle	+	f_{++} (VP)	f_{+-} (FN)
	−	f_{-+} (FP)	f_{--} (VN)

Table 1.3 – Matrice de confusion [133].

Remarque : À partir de cette matrice, nous pouvons calculer d'autres métriques d'évaluation qui sont présentées ci-après dans le tableau 1.4.

1.3.10 Métriques d'évaluation

Les métriques utiles pour mesurer objectivement la performance des modèles de classification sont résumées dans le tableau 1.4 [30, 80, 133, 155].

Métrique	**Description**	**Formule**
Exactitude (Accuracy)	Elle désigne la proportion des prédictions correctement classées par rapport au nombre total des cas évalués.	$$\text{Exactitude} = \frac{VP + VN}{VP + VN + FP + FN} \quad (1.2)$$
Précision (Precision)	C'est le rapport entre les prédictions positives correctement classées et le nombre total des cas assignés comme positifs.	$$\text{Précision} = \frac{VP}{VP + FP} \quad (1.3)$$
Rappel (Recall)	C'est le rapport entre les prédictions positives correctement classées et le nombre total des cas réellement positifs.	$$\text{Rappel} = \frac{VP}{VP + FN} \quad (1.4)$$
F1-mesure (F1-score)	Elle représente la moyenne harmonique entre la précision et le rappel.	$$F_1 = \frac{2pr}{p + r} = \frac{2 \times VP}{2 \times VP + FP + FN} \quad (1.5)$$
Courbe ROC (Receiver Operating Characteristic)	Il s'agit d'un graphique qui permet de représenter le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour donner une vue d'ensemble des performances du modèle de classification.	$$\text{TPR} = \frac{VP}{VP + FN} \quad \text{et} \quad \text{FPR} = \frac{FP}{FP + VN} \quad (1.6)$$
AUC (Area Under the ROC Curve)	Elle désigne l'aire sous la courbe ROC qui permet de mesurer la capacité du modèle à discriminer entre classes.	$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \quad (1.7)$$

TABLE 1.4 – Métriques d'évaluation [30, 80, 133, 155].

Remarque : Nous disons qu'un modèle de classification est bon lorsqu'il possède une exactitude, une précision, une F1-mesure, et un rappel élevés, ainsi qu' une AUC proche de 1. De façon similaire, des valeurs faibles de ces métriques indiquent une performance insuffisante du modèle. Chacune de ces mesures apporte une interprétation spécifique de la capacité du modèle à classifier correctement.

1.3.11 Sur-apprentissage, sous-apprentissage, et solutions

Dans le contexte du ML classique ou profond, il existe généralement deux (02) problèmes courants, le sur-apprentissage (overfitting) et le sous-apprentissage (underfitting). Le compromis entre ces problèmes est un défi majeur lors de la création des modèles prédictifs que nous devons réduire, et résoudre en utilisant des techniques adéquates comme le tableau 1.5 l'illustre [13, 137].

	Sur-apprentissage	**Sous-apprentissage**
Problème	Le modèle montre une excellente performance sur les données d'apprentissage, mais une faible capacité à généraliser sur de nouvelles données.	Le modèle est très simple avec une mauvaise performance sur les données d'apprentissage, ainsi qu'une faible capacité à généraliser sur de nouvelles données.
Causes principales	• Phase d'entraînement trop longue. • Trop de paramètres.	• Phase d'apprentissage très courte. • Peu de paramètres.
Solutions	• Arrêt d'entraînement avant que la performance ne cesse de s'améliorer (arrêt précoce). • Réduction du réseau afin d'exclure le bruit présent dans l'ensemble d'entraînement. • Augmentation de la taille du jeu de données. • Régularisation en introduisant une contrainte sur les paramètres du modèle.	• Augmentation de la complexité du modèle. • Augmentation du temps d'entraînement. • Optimisation des hyperparamètres.

TABLE 1.5 – Sur-apprentissage, sous-apprentissage, causes principales, et quelques solutions [13, 137]

Remarque : Il est fréquent de rencontrer des problèmes de sous-apprentissage et de sur-apprentissage soit dans la régression, dans la classification, ou dans l'apprentissage profond (DL), comme le montre la figure 1.25.

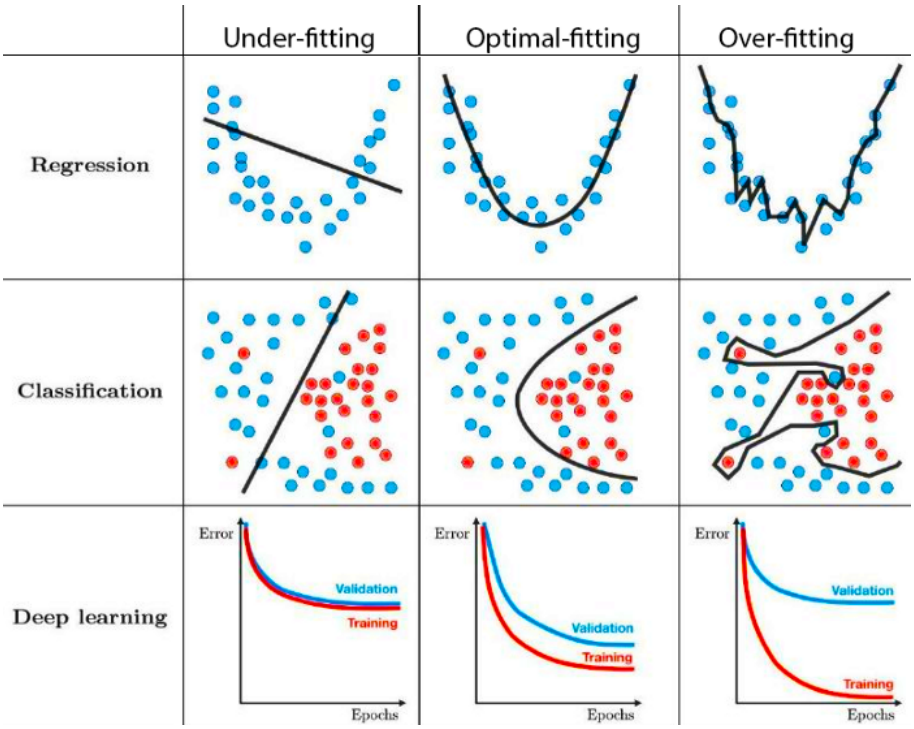


FIGURE 1.25 – Sur-apprentissage et sous-apprentissage [16].

1.3.12 ML dans la prédiction des maladies cardiovasculaires (MCV)

Le ML se concentre sur le développement d'algorithmes capables d'apprendre en analysant de vastes ensembles de données comme les données médicales.

Dans le cadre de la prédiction des maladies cardiovasculaires (MCV), il est utilisé pour détecter à l'avance le risque d'avoir cette maladie à partir de données comme l'âge, le sexe, la pression artérielle, le cholestérol, l'historique médical, et le mode de vie.

Les algorithmes de ML, tels que les forêts aléatoires (RF) et les machines à vecteurs de support (SVM), offrent des prédictions plus ou moins satisfaisantes. Toutefois, leur mise en application nécessite une validation rigoureuse pour garantir leur fiabilité et leur efficacité en milieu clinique, ce qui peut réduire les taux de mortalités en prenant les mesures nécessaires [134].

1.3.13 Avantages et inconvénients du ML en santé

Le ML est utilisé dans divers domaines tels que la médecine en présentant des avantages mais aussi des inconvénients comme le montre le tableau 1.6 suivant :

Avantages	**Inconvénients**
• Détection précoce des maladies grâce à l'analyse de grands volumes de données. • Prédiction des épidémies. • Personnalisation et recommandation des traitements en fonction du profil des patients. • Autonomisation des patients et gain de temps pour les professionnels de santé. • Optimisation des ressources et réduction des erreurs médicales, etc.	• Besoin de données de haute qualité pour éviter les erreurs. • Manque d'interprétabilité des modèles complexes comme les réseaux de neurones. • Risque de biais si les données ne sont pas bien équilibrées. • Les erreurs de prédiction entraînent des traitements inappropriés ou un retard de prise en charge, ce qui affecte négativement la vie des patients.

TABLE 1.6 – Avantages et inconvénients du ML en santé [53].

1.4 Conclusion

Dans ce chapitre, nous avons abordé les concepts fondamentaux en relation avec notre thème, incluant des généralités sur les maladies cardiovasculaires (MCV). Nous avons mis l'accent sur ce système en identifiant l'anatomie, le fonctionnement, les types de ses pathologies, les divers facteurs de risques, les symptômes et quelques conseils de prévention contre ces maladies.

Nous avons également présenté un aperçu introductif des principes de base de l'IA et de ses sous-domaines (ML et DL), notamment les différents types de ML et quelques algorithmes courants tels que les forêts aléatoires (RF). Nous avons introduit aussi le concept d'apprentissage profond (DL), ses architectures en explorant quelques problèmes et solutions associés à ce type d'apprentissage. De plus, nous avons identifié les défis de généralisation en IA : le sur-apprentissage et le sous-apprentissage ainsi que quelques solutions adéquates après avoir mis en évidence comment évaluer les performances des modèles de classification.

Enfin, nous avons déterminé l'impact et l'utilité de l'apprentissage automatique (ML) dans la prédiction des maladies cardiovasculaires (MCV), les avantages et les inconvénients de ce dernier dans le domaine médical.

Dans le chapitre suivant, nous examinerons l'état de l'art sur la prédiction des maladies cardiovasculaires (MCV), afin d'arriver à proposer une nouvelle approche qui aide à améliorer la performance des modèles prédictifs en analysant les méthodes actuelles et les perspectives à venir.

Chapitre 2

État de l'art

2.1 Introduction

Ce chapitre sera consacré à l'état de l'art. D'abord, nous aborderons quelques travaux connexes ayant utilisé l'apprentissage automatique ou l'apprentissage profond dans le cadre de la prédiction des maladies cardiovasculaires. Puis, nous les synthétiserons dans des tableaux, pour enfin, terminer avec une discussion des travaux de chacun des deux (02) types d'apprentissage.

2.2 Travaux connexes

Ces dernières années, les travaux de recherche sur l'IA et les maladies cardiovasculaires sont en développement continu et rapide, comme le montre la figure 2.1 suivante [61] :

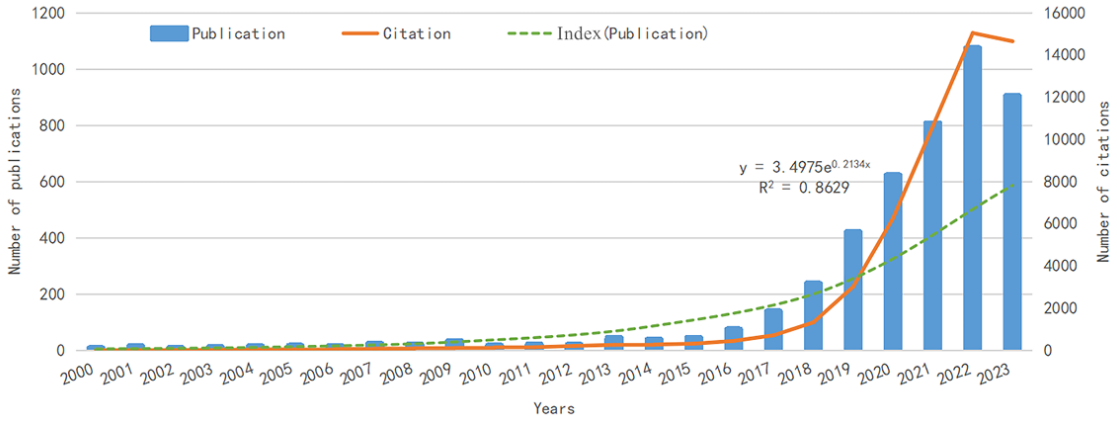


FIGURE 2.1 – Tendance mondiale des publications en recherche pathologique sur les maladies cardiovasculaires par l'IA au cours des 23 dernières années [61].

2.2.1 Classification des travaux connexes analysés

Dans notre travail, nous explorerons les articles les plus significatifs sur la prédiction des maladies cardiovasculaires, vu qu'il existe de nombreux travaux sur ce sujet. Ces derniers, nous pouvons les classifier en deux (02) grandes familles selon les approches proposées *"apprentissage automatique (ML)"* et *"apprentissage profond (DL)"* comme le montre la figure 2.2

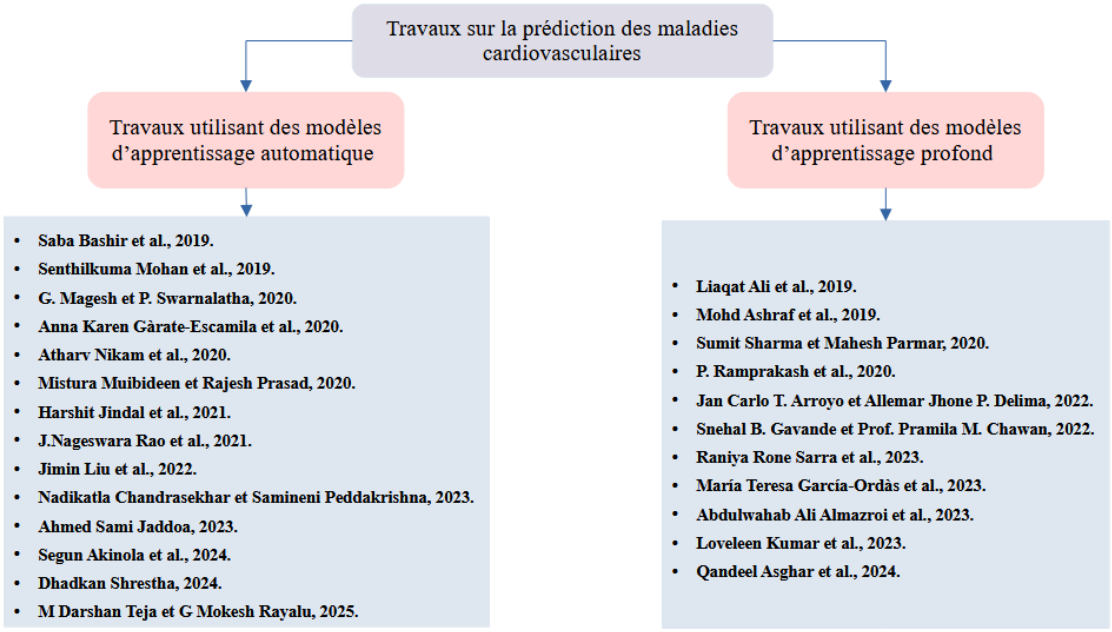


FIGURE 2.2 – Classification des travaux connexes analysés.

2.2.2 Présentation des travaux connexes utilisant des modèles d'apprentissage automatique (ML)

Parmi les travaux analysés proposant des approches de l'apprentissage automatique, nous avons :

« Improving Heart Disease Prediction Using Feature Selection Approaches, 2019 »

Saba Bashir et al. [77], ont suggéré d'utiliser des techniques de sélection de variables combinées avec des algorithmes d'apprentissage automatique.

Ils ont appliqué la méthode *MRMR* (Minimum Redundancy Maximum Relevance) sur un dataset médical d'UCI Repository à l'aide de l'outil *"RapidMiner"* afin de pondérer les variables les plus pertinentes, avant de tester les cinq (05) algorithmes suivants : Decision Tree *(DT)*, Logistic Regression *(LR)* avec *SVM*, Naive Bayes *(NB)* et Random Forest *(RF)*.

Il a été constaté que les meilleurs résultats ont été obtenus avec *Naive Bayes* avec une exactitude de **84,24%** et *Logistic Regression-SVM* avec une exactitude de **84,85%**.

La figure 2.3 montre l'architecture de la méthodologie proposée.

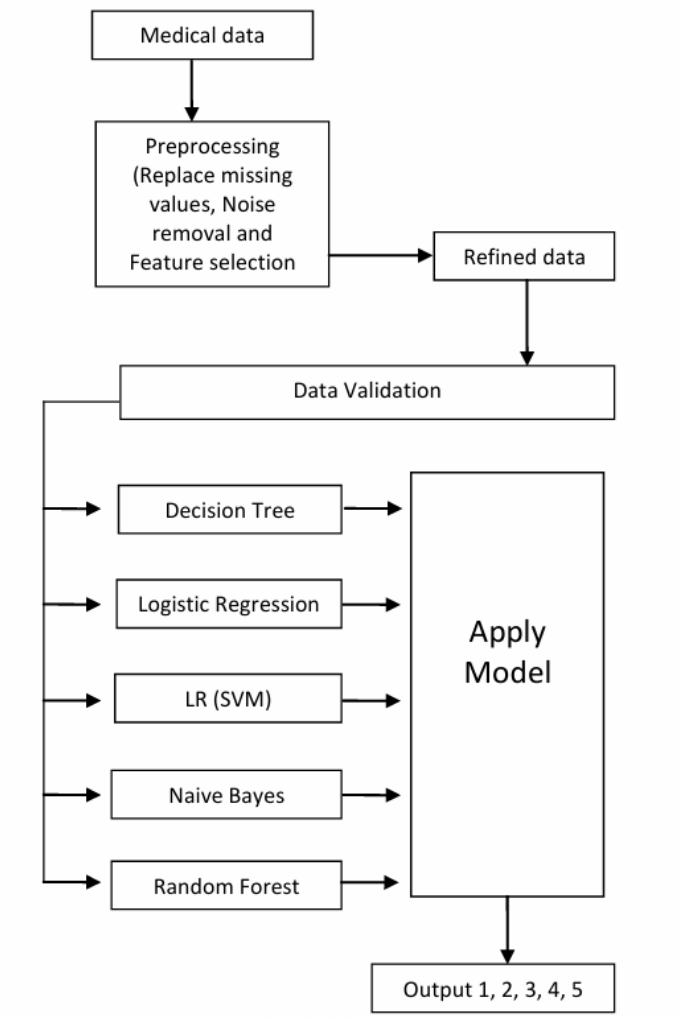


FIGURE 2.3 – Architecture de la méthodologie proposée par Saba Bashir et al [77].

« Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, 2019»

Senthilkumar Mohan et al. [79] ont présenté une méthode hybride, nommée *HRFLM* (Hybrid Random Forest with Linear Model), combinant les forêts aléatoires *(RF)* et un modèle linéaire *(LM)*, afin d'améliorer la précision de la prédiction des maladies cardiaques.

Les auteurs ont précisé qu'ils ont utilisé le dataset *"Cleveland"*, après avoir effectué un prétraitement, et divisé en huit (08) sous-ensembles avec DT. Selon les résultats, l'approche proposée est plus performante que les méthodes déjà existantes, telles que : les arbres de décision (DT), SVM, les réseaux de neurones, etc. En effet, son exactitude aurait atteint **88,47%**.

Ci-dessous 2.4, un schéma explicatif de l'approche *HRFLM* :

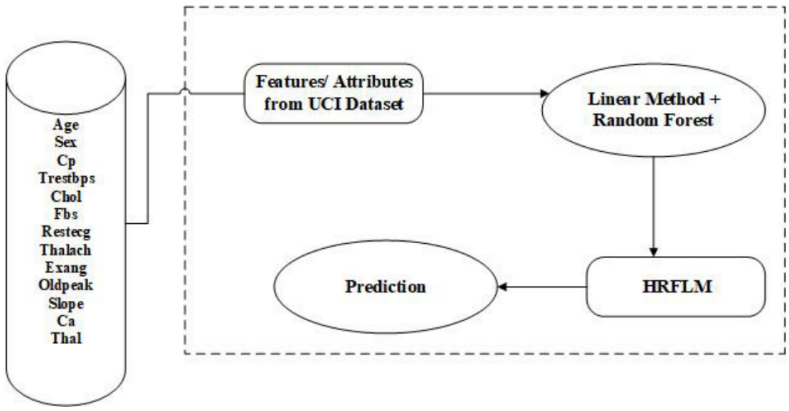


FIGURE 2.4 – Schéma explicatif de l'approche HRFLM de Senthilkumar Mohan et al [79].

« Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction, 2020 »

G. Magesh et P. Swarnalatha [109] ont proposé une approche pour la prédiction des maladies cardiovasculaires basée sur la sélection optimale des caractéristiques. Il s'agit d'une technique appelée *CDTL* (Cluster-based Decision tree Learning) qui vise à améliorer les performances des classificateurs d'apprentissage automatiques, sachant que la plupart des modèles rencontrent des problèmes de sélection des caractéristiques, de fractionnement des attributs, et des ensembles de données déséquilibrés. Les auteurs l'ont expérimenté sur le jeu de données *"Cleveland"* contenant 303 enregistrements et 14 attributs. Cette nouvelle méthodologie repose sur cinq (05) étapes clés :

- Répartition du jeu de données selon la distribution de la variable cible.
- Création de nouvelles combinaisons entre classes à l'aide des enregistrements les plus représentés.
- Sélection des caractéristiques pertinentes à l'aide d'entropie [1].
- Répartition des données en clusters entropiques.
- Évaluation des performances du meilleur classificateur *(RF)* en utilisant les caractéristiques informatives que pour les classes **0** et **1**.

Le résultat obtenu aurait atteint une exactitude de **89,30%** contre **76,70%** *sans CDTL*. De ce fait, les auteurs ont constaté que leur méthodologie surpasse les autres modèles (*SVM*, *LM*, *DT*) en termes d'exactitude en minimisant le taux d'erreur.

1. Entropie : C'est une mesure utilisée en théorie de l'information qui caractérise le niveau de désordre ou d'incertitude dans un ensemble de données [120].

«Classification models for heart disease prediction using feature selection and PCA, 2020»

Anna Karen Garàte-Escamila et al. [50] ont proposé d'utiliser une méthode combinant le test du chi-carré (Chi^2) avec l'analyse en composantes principales (PCA).

Cette étude a exploité trois (03) datasets : *"Cleveland"*, *"Hungarian"* et *"Cleveland-Hungarian"* qui est une combinaison des deux (02) précédents. Étant donné le nombre important de caractéristiques, il a été jugé nécessaire de, d'abord, sélectionner les variables les plus pertinentes à l'aide du test du *chi-carré*, puis de réduire la dimension du dataset.

Six (06) classifieurs ont été testés : *DT*, *GBT*, *LR*, *MPC*, *NB* et *RF*. Parmi tous ces algorithmes, *RF* a donné les meilleurs résultats avec une exactitude de **98,7%** pour *"Cleveland"*, **99,0%** pour *"Hungarian"*, et **99,4%** pour *"Cleveland-Hungarian"*.

La figure 2.5 montre le schéma de l'approche proposée.

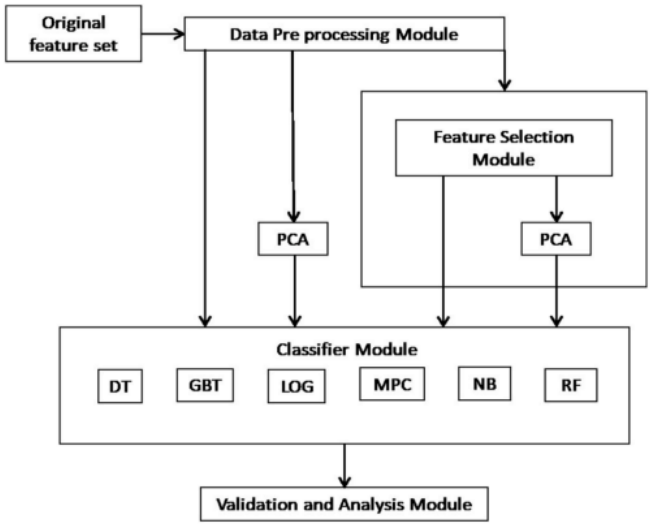


FIGURE 2.5 – Schéma de l'approche proposée par Anna Karen Garàte-Escamila et al [50].

« Cardiovascular Disease Prediction Using Machine Learning Models, 2020 »

Atharv Nikam et al. [51] ont proposé l'ajout d'une caractéristique à partir des données déjà existantes dans le dataset, et qui est l'indice de masse corporelle *(IMC)*, calculé à partir du poids et de la taille du patient.

Sept (07) classificateurs ont été testés : *LR*, *KNN*, *NB*, *DT*, *les réseaux de neurones*, *XGBoost* et *LGBM*.

Parmi tous ces classificateurs, *l'arbre de décision (DT)* semble être le plus performant avec une précision de **73,13%**. De plus, l'*IMC* est identifié comme une caractéristique clé pour améliorer les performances de prédiction.

« A Fast Algorithm For Heart Disease Prediction Using Bayesian Network Model, 2020»

Mistura Muibideen et Rajesh Prasad [124], ont proposé l'application des réseaux bayésiens *(BN)*.

Les auteurs ont opté pour le dataset *"Cleveland"*, qu'ils ont d'abord nettoyé et prétraité. Ensuite, ils ont construit le modèle à l'aide de la bibliothèque `bnlearn` du langage R, et ont procédé à l'apprentissage des tables de probabilités conditionnelles pour chaque variable, avant de terminer avec l'implémentation du modèle avec Python.

Selon les résultats, *BN* a atteint une exactitude de **85%**, dépassant ainsi celle de *Naive Bayes* qui est de **80%**.

Ci-dessous 2.6, le diagramme de flux de la conception du réseau :

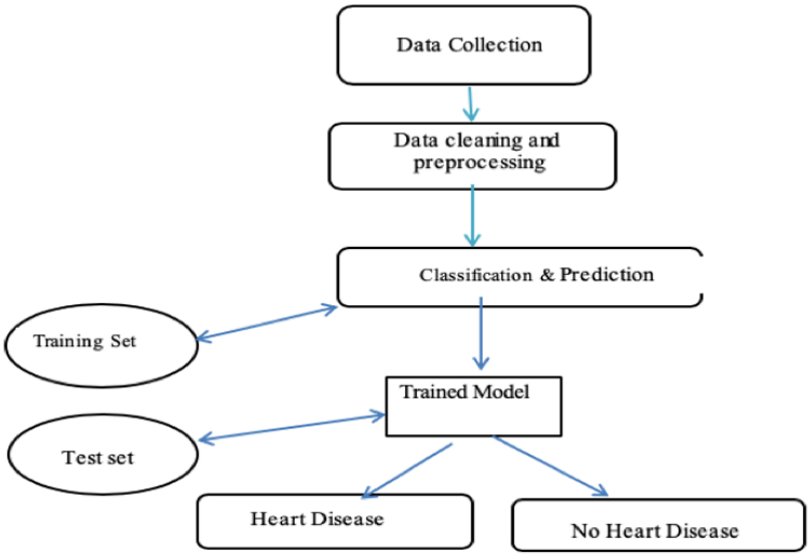


FIGURE 2.6 – Diagramme de flux de la conception du réseau [124].

« Heart disease prediction using machine learning algorithms, 2021 »

Harshit Jindal et al. [57] ont proposé de tester trois (03) méthodes de machine learning, qui sont les suivantes : *la régression logistique (LR)*, *KNN* et *Random Forest Classifier*, sur le dataset *"Heart Disease"*, provenant d'UCI Repository.

Après avoir nettoyé et normalisé les données du dataset, les trois (03) algorithmes ont été testés séparément. Les résultats montrent que *KNN* est le plus efficace, avec une exactitude de **88,52%**.

Les auteurs ont précisé que les classes du dataset sont légèrement déséquilibrées.

« Cardiovascular Disease Prediction Using Machine Learning Techniques, 2021 »

J.Nageswara Rao et al. [62] ont proposé un système de prédiction des maladies cardiaques utilisant différentes techniques d'apprentissage automatique.

Le dataset utilisé dans cette étude est *"Cleveland"* après un prétraitement de ses données, sur lequel ont été testés les algorithmes suivants : *KNN*, *Naive Bayes*, *DT*, *RF*, *SVM* et *LR*.

Selon les résultats, *RF* est le plus précis, avec une exactitude de **90,16%**, tandis que les autres algorithmes montrent des performances inférieures, tels que *KNN*, avec **67,21%** d'exactitude.

« Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion, 2022»

Jimin Liu et al. [60] ont développé une nouvelle approche à base des méthodes d'apprentissage ensembliste qui est un peu avancée par rapport aux modèles traditionnels. Cette dernière est connue sous le nom du *"stacking"*. En exploitant un *jeu de données fusionné* (Cleveland, Hongrie, Suisse, Long Beach VA et Stalog) comprenant onze (11) caractéristiques et 918 échantillons sans doublons, l'idée met en évidence tout un processus. Après la répartition des données en données d'entraînement et de test, vient la normalisation.

Par la suite, une technique appelée *SHAP* (SHapley Additive exPlanations) a été utilisée pour sélectionner et expliquer le choix des caractéristiques importantes en tenant compte de l'intégration du modèle *GBDT*. De ce fait, dix (10) caractéristiques ont été choisies pour les étapes restantes.

L'évaluation de dix (10) classificateurs (*SVM*, *KNN*, *LR*, *RF*, *ET* (Extra Tree), *GBDT*, *XGBoost*, *LightGBM*, *CatBoost* et *MLP*) avait pour but de choisir les modèles les plus pertinents en respectant la condition *« bon mais différent »* afin de les considérer comme apprenants de base du modèle *"stacking"*. Les cinq (05) meilleurs modèles incluent *LR*, *RF*, *Extra Tree*, *MLP* et *CatBoost*. Ils ont également utilisé la *LR* comme méta-classificateur pour éviter le sur-apprentissage et en sortir avec une prédiction finale. Les hyperparamètres du modèle ont été optimisés avec *"optuna"*.

Les résultats obtenus montrent que le modèle proposé est meilleur en termes d'exactitude **(89,86%)**, de F1-mesure **(91,36%)** et d'AUC **(95%)**. En outre, les auteurs ont appliqué les mêmes étapes sur un autre jeu de données *"Heart Attack"* afin de vérifier sa capacité de généralisation. Finalement, ceci a souligné également une amélioration significative atteignant **84,62%** d'exactitude, **92%** d'AUC ainsi que **86%** de rappel et de F1-mesure.

La figure 2.7 illustre le schéma du processus de prédiction proposé.

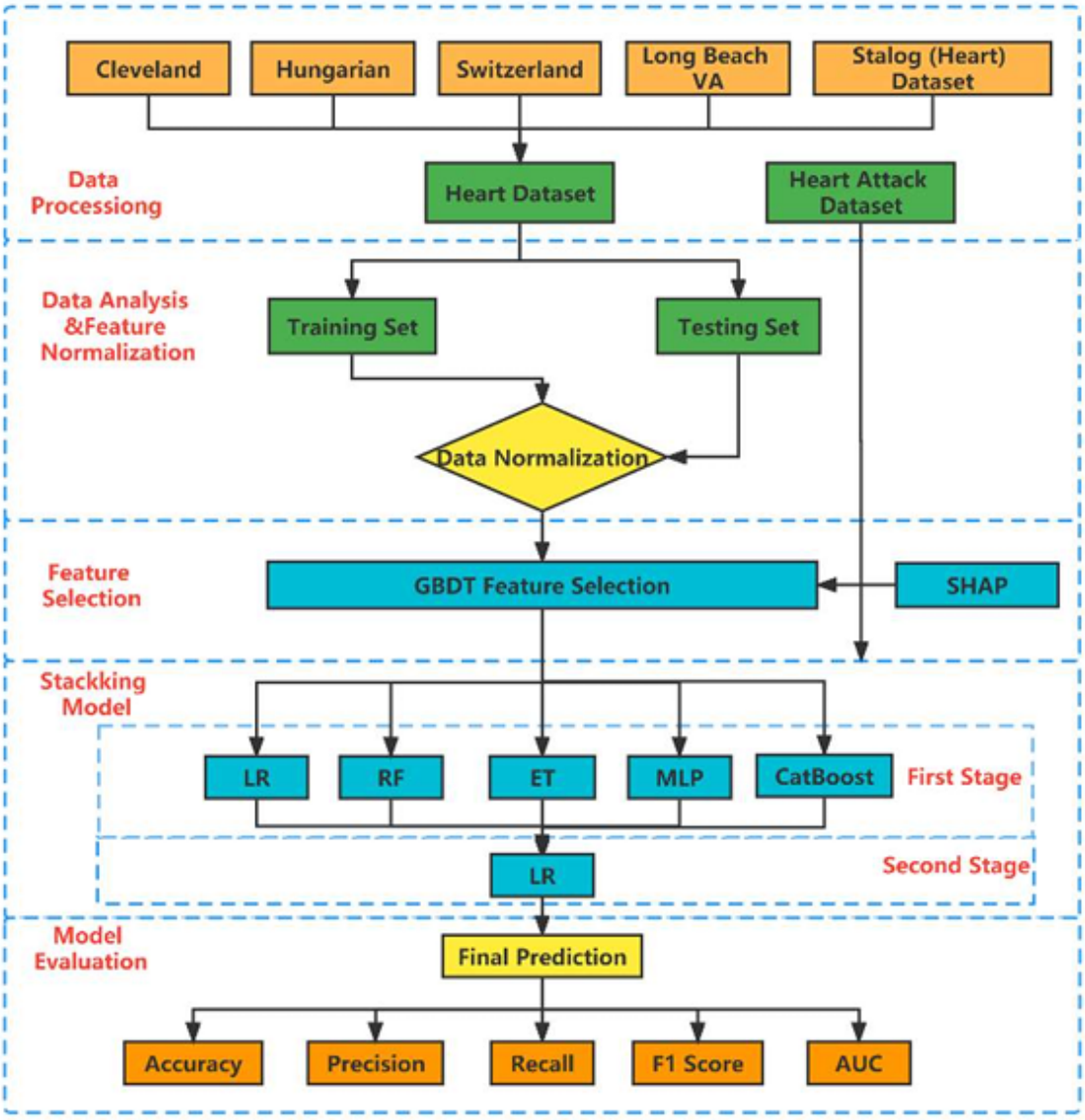


FIGURE 2.7 – Processus de prédiction proposé par Jimin Liu et al [60].

« Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization, 2023 »

Nadikatla Chandrasekhar et Samineni Peddakrishna [34] ont proposé une approche qui consiste à améliorer l'exactitude de prédiction des maladies cardiaques, basée sur six (06) algorithmes de l'apprentissage automatique (*RF*, *KNN*, *LR*, *NB*, *GB* et *AB*) combinés à l'aide du classificateur *SVE* (Soft Voting Ensemble).

Dans leur étude, ils ont détaillé les différentes phases qui mènent au but, de la collection des données jusqu'à la prédiction finale. Leurs expérimentations ont été implémentées sur deux (02) ensembles de données (*"Cleveland"* et *"IEEE Dataport"*) et évaluées par plusieurs métriques (matrice de confusion, exactitude, précision, rappel, F1-mesure, AUC-ROC, etc). Ces chercheurs ont utilisé *GridSearchCV* pour optimiser les hyperparamètres et une *validation croisée* (5-folds).

En outre, ils ont évalué en premier lieu les algorithmes précédents individuellement. Le modèle le plus performant ainsi que son exactitude dans chaque ensemble de données sont représentés dans le tableau 2.1.

Jeu de données utilisé	**Meilleur modèle testé**	**Exactitude**
Cleveland	RF	90,16%
IEEE Dataport	AB	90%

TABLE 2.1 – Meilleur modèle testé et son exactitude dans chaque ensemble de données [34].

En second lieu, ils ont également testé leur proposition et ont obtenu les meilleurs résultats dans les deux (02) jeux de données, atteignant **93,44%** (Cleveland) et **95%** (IEEE Dataport) d'exactitude. Ils sont arrivés à une bonne idée en comparant que ce soit par rapport aux algorithmes testés individuellement ou par rapport aux approches existantes dans la littérature tel qu'ils ont précisé dans leur article. La figure 2.8 illustre le *SVE* proposé :

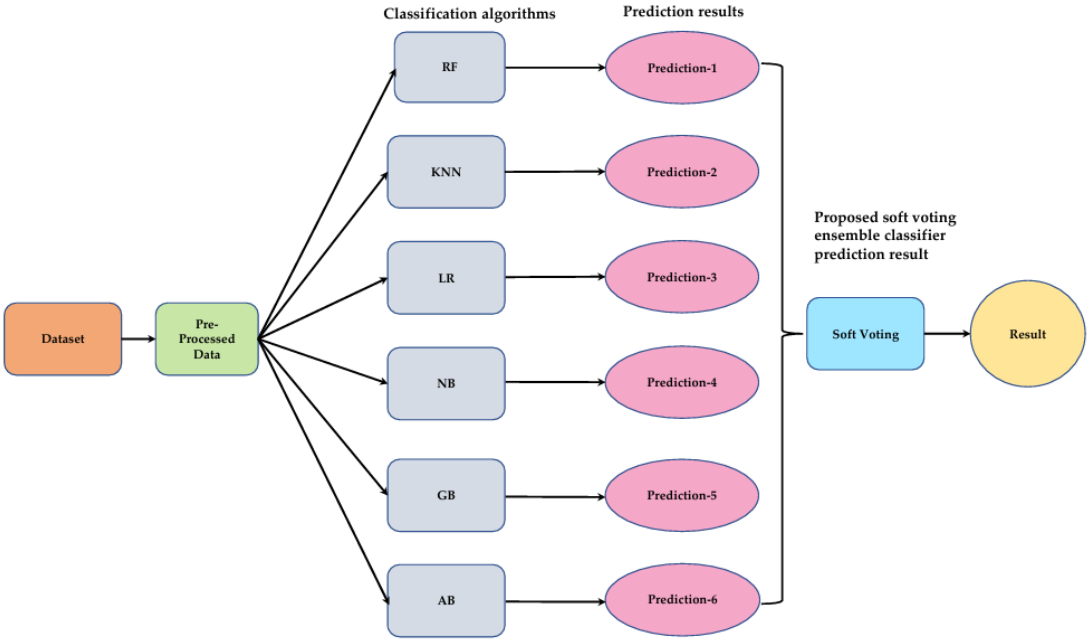


FIGURE 2.8 – SVE proposé par Nadikatla Chandrasekhar et Samineni Peddakrishna [34].

« Heart Disease Prediction System Using (SMOTE Technique) Balanced Dataset and Decision Tree Classifier, 2023 »

Ahmed Sami Jaddoa [94] a proposé d'utiliser l'algorithme *DT* avec la technique de suréchantillonnage *SMOTE* afin de prédire la présence ou l'absence d'une maladie cardiaque à l'aide des données bien équilibrées. Cette étude a été faite sur l'ensemble de données *"Cleveland Heart Disease Dataset"*. Après le prétraitement des données, deux (02) approches ont été testées : *DT avec SMOTE* (méthode proposée) et *DT sans SMOTE*. Comme résultat, la méthode proposée s'est avérée la plus performante en termes des cinq (05) métriques utilisées pour l'évaluation (exactitude, précision, sensibilité, spécificité et F-mesure), atteignant **91,4%** d'exactitude contre **73,3%** sans gestion du déséquilibre.

Ci-dessous 2.9, le système de prédiction proposé :

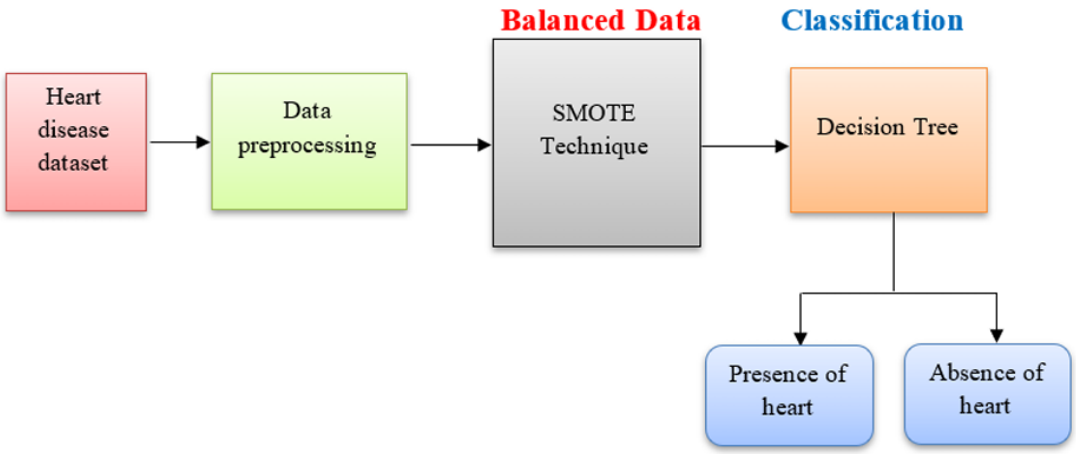


FIGURE 2.9 – Système de prédiction proposé par Ahmed Sami Jaddoa [94].

« Enhancing cardiovascular disease prediction : A hybrid machine learning approach integrating oversampling and adaptive boosting techniques, 2024 »

Dans cet article, **Segun Akinola et al.** [78] ont combiné l'algorithme *AdaBoost* avec *SMOTE*, puis associé aux classificateurs *RF*, *XGBoost* et *ET*, afin d'améliorer la précision des prédictions des maladies cardiaques.

Le dataset utilisé dans cette étude est *"Heart Failure Prediction Dataset"* comprenant des attributs mesurés à partir d'électrocardiogrammes (ECG) enregistrés pour différentes personnes ayant des rythmes cardiaques variés.

Les résultats des algorithmes de ML combinés avec *AdaBoost-SMOTE* donnent des précisions plus élevées. Par exemple, l'algorithme *RF sans AdaBoost-SMOTE* a une précision de **96,87%** tandis qu'*avec AdaBoost-SMOTE* a une précision de **97.35%**.

« Advanced Machine Learning Techniques for Predicting Heart Disease : A Comparative Analysis Using the Cleveland Heart Disease Dataset, 2024 »

Dhadkan Shrestha [150] présente une étude comparative intéressante des modèles de l'apprentissage automatique traditionnels *(LR)* et avancés (*RF*, *Gradient Boosting*, *XGBoost* et les réseaux de type *LSTM*). L'objectif principal de cette étude est de trouver le modèle ayant la meilleure performance dans le cadre de la prédiction des maladies cardiovasculaires. Le jeu de données utilisé est celui de *"Cleveland"*.

Le prétraitement de données effectué inclut la gestion des valeurs manquantes en les remplaçant par la médiane, la transformation des caractéristiques catégorielles en entiers ainsi que la binarisation de la variable cible en *"présence (1)"* et *"absence (0)"* de maladie. Par la suite, l'auteur a divisé l'ensemble de données en données d'entraînement **(80%)** et de test **(20%)**. En outre, les modèles illustrés auparavant ont été entraînés en sachant que la technique *SHAP* a été utilisée pour *XGBoost* et *LSTM* afin de comprendre l'impact des caractéristiques, puis évaluer ces modèles à l'aide de diverses métriques notamment l'exactitude, la précision, le rappel, la F1-mesure et l' AUC-ROC. Les résultats obtenus révèlent que le *XGBoost* est le plus efficient par rapport aux autres avec une exactitude de **90%** et une AUC-ROC de **94%**, tandis que *LSTM* s'avère être le moins performant et le moins adapté aux données tabulaires statiques.

« Optimizing heart disease diagnosis with advanced machine learning models : a comparison of predictive performance, 2025 »

Dans cet article, **M Darshan Teja et G Mokesh Rayalu** [155] ont testé quinze (15) algorithmes d'apprentissage automatique différents, notamment *LR*, *RF*, *SVR*, *KNN*, *GBM*, *NN*, *XGBoost*, *MANN*, *FDA*, *CIT*, *BT*, *NB*, *MARS*, *BGGLM* et *BGLM* afin d'arriver à une analyse détaillée des performances de chacun de ces modèles. Dans cette étude, ils ont utilisé un *ensemble de données fusionné* (Cleveland, Suisse, Hongrie, Long Beach et Stalog) contenant 1190 enregistrements et douze (12) caractéristiques. Sept (07) parmi ces douze (12) ont été sélectionnées à l'aide de la matrice de corrélation et des recherches précédentes qui prouvent leur importance pour diagnostiquer les maladies cardiaques et améliorer les prédictions. Ce jeu de données a été divisé en **80%** pour l'entraînement et **20%** pour le test, puis ils ont appliqué les quinze (15) algorithmes précédents.

L'évaluation de ces modèles a été faite avec diverses métriques incluant la matrice de confusion, l'exactitude, la précision, le rappel, la F1-mesure et l'AUC-ROC. Comme résultat, *XGBoost* et *BT* affichent les meilleures exactitudes de **93%** après *RF* et *KNN* en deuxième position atteignant **91%**. Une validation croisée a été utilisée pour vérifier leur capacité de généralisation. De ce fait, ils ont remarqué que *RF* a gardé ses performances élevées, *XGBoost* a connu une légère réduction mais pour

KNN une diminution importante était observée, ceci implique un sur-apprentissage. En termes d'AUC, *RF*, *BT*, *XGBoost* et *GBM* atteignant respectivement **95%**, **95%**, **94%** et **92%**. En comparant avec les autres algorithmes testés, ces modèles ensemblistes étaient les plus performants.

La figure 2.10 présente la méthodologie adoptée pour l'analyse extraite de l'article [155].

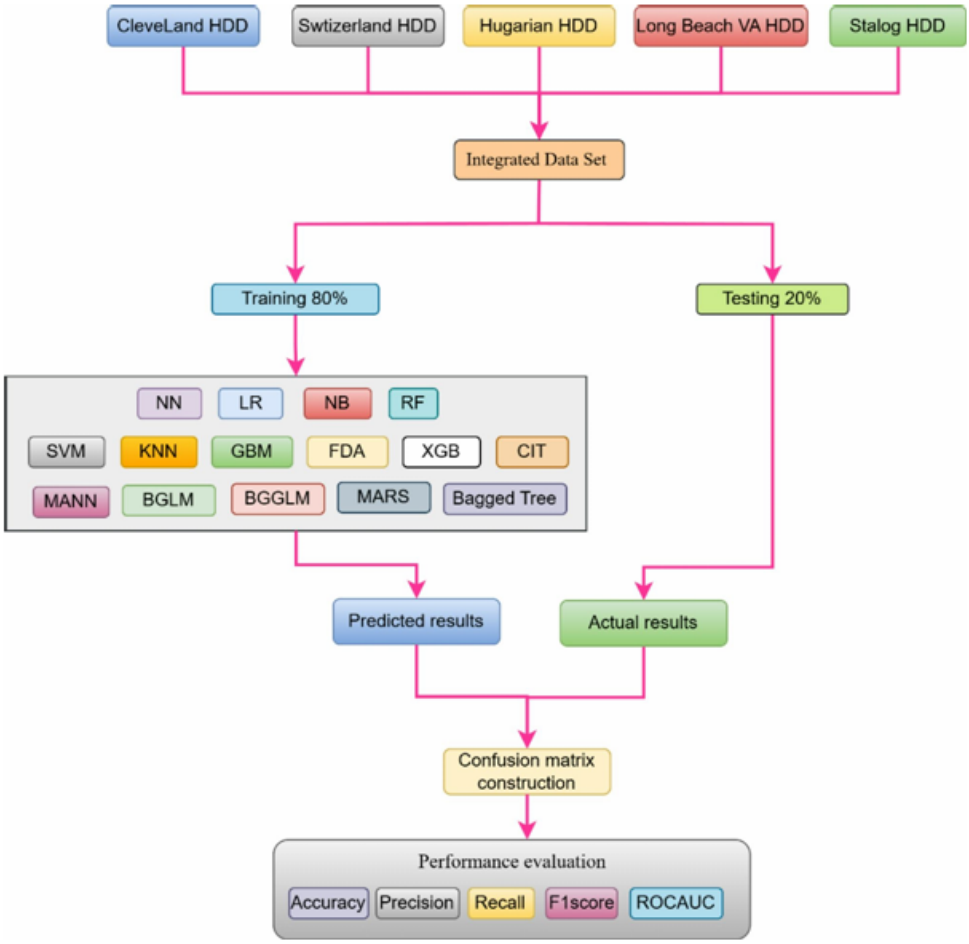


FIGURE 2.10 – Méthodologie adoptée pour l'analyse proposée par M Darshan Teja et G Mokesh Rayalu [155].

2.2.3 Synthèse des travaux connexes utilisant des modèles d'apprentissage automatique

Dans ce qui suit, un tableau récapitulatif 2.2 des articles présentés auparavant en matière de prédiction des maladies cardiovasculaires, basés sur les techniques de l'apprentissage automatique. Ceci afin d'avoir des idées générales sur l'approche proposée, le ou les ensemble(s) de données mis en œuvre, les résultats d'évaluation, et enfin quelques observations essentielles.

Papier	**Méthode proposée**	**Dataset utilisé**	**Évaluation**	**Observations importantes**
Saba Bashir et al., 2019 [77].	MRMR + méthodes de ML.	Dataset médical de la base UCI Repository.	Exactitude : **84,24%** avec Naive Bayes. **84,85%** avec Logistic Regression-SVM.	- Ces techniques n'ont pas été appliquées à des datasets médicaux en temps réel ni exploitées sous forme d'ensembles.
Senthilkuma Mohan et al., 2019 [79].	HRFLM (Hybrid Random Forest with Linear Model).	Cleveland.	Exactitude : **88,47%**.	- Risque de sur-apprentissage dû à la division du dataset en sous-ensembles. - Le déséquilibre des sous-ensembles pourrait biaiser les résultats du modèle.
G. Magesh et P. Swarnalatha, 2020 [109].	Cluster-based Decision Tree Learning (CDTL).	Cleveland.	Exactitude : **89,39%**. Taux d'erreur : **9,70%**.	- RF avec CDTL atteint de meilleures performances par rapport aux autres algorithmes testés avec une réduction de **13,6%** du taux d'erreur. - CDTL est basé sur l'utilisation des enregistrements à forte distribution ainsi qu'une partition entropique. - Réduction de la dimensionnalité sans perte de performances prédictives. - CDTL gère le déséquilibre des classes par une sélection intelligente des attributs selon la distribution des classes. - Malgré ce résultat, le CDTL semble compliqué en implémentation en raison de ses multiples étapes qui nécessitent une expertise en clustering, théorie de l'information ainsi qu'en optimisation des modèles de ML.
Anna Karen Gàrate-Escamila et al., 2020 [50].	Test du chi-carré + ACP + RF.	Cleveland, Hungarian, Cleveland-Hungarian.	Exactitude : **98,7%** pour Cleveland, **99,0%** pour Hungarian, **99,4%** pour Cleveland-Hungarian.	- Déséquilibre remarquable des classes des datasets Hungarian et Cleveland-Hungarian.

Atharv Nikam et al., 2020 [51].	Ajout d'une caractéristique + Méthode de machine learning.	Non précisé.	Meilleure exactitude : **73,13%** avec DT.	- Le dataset n'a pas été clairement identifié. - Modeste exactitude, ce qui limite l'usage de ce modèle en pratique.
Mistura Muibideen et Rajesh Prasad, 2020 [124].	Réseaux Bayésiens.	Cleveland.	Exactitude : **85%**.	- Le manque de flexibilité de `bnlearn` a poussé les auteurs à implémenter leur modèle avec python.
Harshit Jindal et al., 2021 [57].	Méthodes de machine learning.	Heart Disease Dataset.	Meilleure exactitude : **88,52%** avec KNN.	- Léger déséquilibre des classes du dataset.
J.Nageswara Rao et al., 2021 [62].	Random Forest (RF).	Cleveland.	Exactitude : **90,16%**.	- Temps de prédiction assez long.
Jimin Liu et al., 2022 [60].	SHAP + Stacking (LR, RF, Extra Tree, MLP et CatBoost).	- Jeu de données fusionné. - Heart Attack.	Exactitude : **89,86%**, précision : **92,5%**, rappel : **90,24%**, F1-mesure : **91,36%** et AUC : **95%** pour le jeu de données fusionné. Exactitude : **84,62%**, précision : **86%**, rappel : **86%**, F1-mesure : **86%** et AUC : **92%** pour Heart Attack.	- Le modèle proposé surpasse tous les modèles testés individuellement en termes d'exactitude et de F1-mesure car il combine plusieurs algorithmes performants et distincts. - SHAP est une technique de sélection et d'interprétation de l'impact des caractéristiques. - L'utilisation de LR comme un méta-classificateur aide à éviter le sur-apprentissage. - Quelques métriques d'évaluation diminuent légèrement, qui peuvent être négligées en cas du stacking telles que la précision, mais une amélioration significative dans la plupart. - Malgré que cette approche atteint les meilleurs résultats, les problèmes de complexité et de lenteur peuvent être présents en raison de l'utilisation combinée de plusieurs techniques (SHAP + stacking avec cinq (05) modèles de base).

Nadikatla Chandrasekhar et Samineni Peddakrishna, 2023 [34].	SVE avec (RF, KNN, LR, NB, GB et AB).	- Cleveland. - IEEE Dataport.	Exactitude : **93,44%** pour Cleveland. Exactitude : **95%** pour IEEE Dataport.	- SVR surpasse RF, KNN, LR, NB, GB et AB ainsi que des techniques existantes dans la littérature, appliquées sur les deux (02) ensembles de données. - L'utilisation de "GridSearchCV" et de la validation croisée améliore les performances. - Malgré ce résultat significatif, les jeux de données utilisés présentent peu de données. - Il est également conseillé d'utiliser, par exemple : des ensembles de données plus complets, l'apprentissage profond, des dispositifs IoT ou autres modèles plus efficaces.
Ahmed Sami Jaddoa, 2023 [94].	DT avec SMOTE.	Cleveland.	Exactitude : **91.4%**. Précision : **94.1%**. Sensibilité : **91.4%**. Spécificité : **91.3%**. F-mesure : **92.8%**.	- Une amélioration significative de **18,1%** d'exactitude a été observée après l'utilisation de SMOTE.
Segun Akinola et al., 2024 [78].	SMOTE + AdaBoost + Random Forest.	Heart Failure Prediction Dataset.	Précision : **97.35%**.	- Gestion du déséquilibre avec SMOTE. - Amélioration des résultats grâce à l'hybridation. - Le modèle est moins opaque que le deep learning, ce qui représente un point faible.
Dhadkan Shrestha, 2024 [150].	XGBoost.	Cleveland.	Exactitude : **90%**. AUC : **94%**.	- Le XGBoost désigne le modèle le plus performant, contrairement à LSTM qui est le moins efficace par rapport aux autres. - Les valeurs "SHAP" aident à comprendre l'influence des variables. - Le choix des modèles à tester n'était pas fait aléatoirement, mais selon leur importance et leur distinction afin d'arriver à une évaluation complète des techniques traditionnelles et avancées. - Malgré que XGBoost atteint une performance significative, l'ancienneté et le manque de données essentielles dans le jeu de données utilisé présentent des limites.

M Darshan Teja et G Mokesh Rayalu, 2025 [155].	Modèles d'ensemble tels que XGboost et BT.	Jeu de données fusionné.	Pour XGBoost (**93%** d'exactitude, **94%** d'AUC, **92%** de précision, de rappel et de F1-mesure). Pour BT (**93%** d'exactitude, **95%** d'AUC, **92%** de rappel, **93%** de précision et de F1-mesure).	- Les méthodes d'ensembles (XGBoost, BT, etc) surpassent les autres algorithmes testés. - La sélection des caractéristiques importantes a été effectuée. - Utilisation de la validation croisée pour vérifier la capacité de généralisation des modèles. - La méthodologie utilisée répond à quelques limites existantes de la littérature.

TABLE 2.2 – Synthèse des travaux connexes utilisant des modèles d'apprentissage automatique.

2.2.4 Discussion

Les travaux précédents montrent la diversité d'approches utilisant des méthodes de machine learning (ML), parfois seules ou combinées avec d'autres algorithmes, dont nous comptons six (06) articles pour la première catégorie et huit (08) articles pour la seconde.

Certains de ces travaux soulignent l'importance du prétraitement, notamment de la sélection des variables les plus pertinentes [50, 109, 155], tandis qu'un autre montre que l'ajout de nouvelles variables cliniques, pourrait améliorer la prédiction, à condition qu'elles soient validées par la littérature médicale ou approuvées par des experts du domaine [51].

Nous avons aussi constaté que certaines approches ont obtenu de très bons résultats, avec des exactitudes dépassant **90%**, rivalisant ainsi avec les approches basées sur l'apprentissage profond.

2.2.5 Présentation des travaux connexes utilisant des modèles d'apprentissage profond (DL)

Parmi les articles analysés proposant des approches de l'apprentissage profond, nous avons :

« An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network, 2019 »

La plupart des travaux précédents sur la prédiction des maladies cardiaques, se focalisent sur le prétraitement de caractéristiques seulement et que le modèle le plus utilisé était *ANN* grâce à sa capacité de traitement des problèmes linéaires et non linéaires.

Afin de répondre à ce genre de remarques, **Liaqat Ali et al.** [64] ont manipulé et pré-traité des données provenant du *"Cleveland heart disease dataset"*. Ceci, pour but de travailler sur un nouveau système de diagnostic automatisé, qui s'articule autours d'une double démarche, à savoir l'affinement des caractéristiques et la résolution des problèmes de sur-apprentissage et de sous-apprentissage. L'idée est d'appliquer un test statistique χ^2 de classement et de sélection des variables importantes. Ensuite, faire apprendre un modèle d'apprentissage profond *2-DNN* avec un paramétrage optimal à l'aide d'une recherche exhaustive *(grid search)*. Ce processus peut améliorer les performances de classification des patients atteints, ou pas, de ce genre de maladies graves.

L'étape de l'évaluation s'est faite à l'aide de six (06) métriques : l'exactitude, le rappel, la spécificité, le coefficient de corrélation de Matthews (MCC), l'AUC, et la courbe ROC. Ces chercheurs sont arrivés à un résultat satisfaisant, après quelques expérimentations. Ils ont essayé de comparer leur approche proposée avec :

— DNN conventionnel.

— D'autres modèles de l'apprentissage automatique notamment, *AB*, *RF*, *ET*, *SVM linéaire* et *SVM avec noyau RBF*.

— Des méthodes de la littérature, telles que *Neural network ensembles* (Resul et al., 2009), *NB* (Newton Cheung, 2001) et *ANN-Fuzzy-AHP* (Samuel et al., 2017).

Ils ont également validé sa capacité de généralisation en utilisant une *validation croisée* à vingt (20) plis.

En outre, ils ont observé que leur modèle affiche les meilleurs résultats de prévision avec **93,33%** d'exactitude, **85,36%** de rappel, **100%** de spécificité, **0,872** de MCC et **0,94** d'AUC-ROC.

Dans cette étude, ils ont rapporté qu'ils n'ont pas analysé le temps de traitement, ce qui souligne un défi.

La figure 2.11 tirée de l'article [64], représente le diagramme de système de diagnostic proposé.

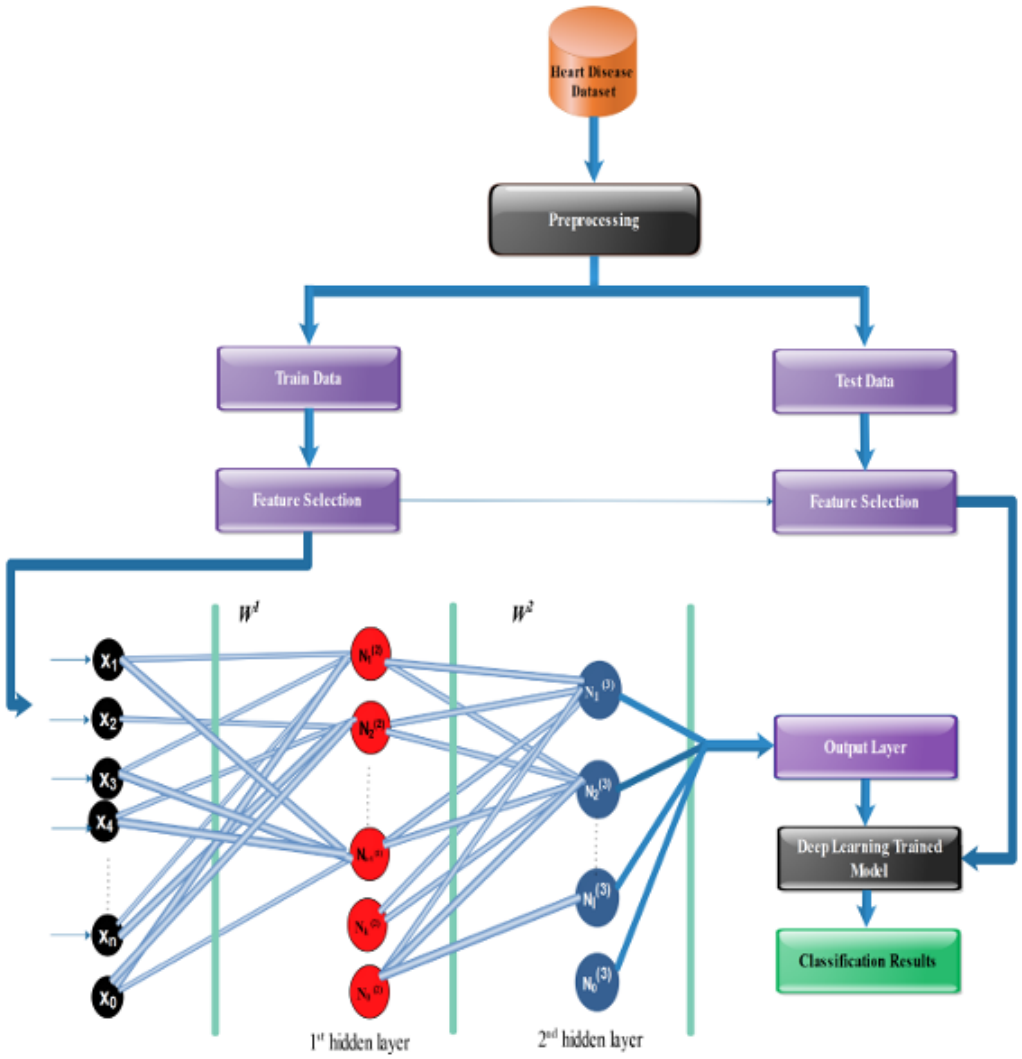


FIGURE 2.11 – Diagramme de système de diagnostic proposé par Liaqat Ali et al [64].

« Improved Heart Disease Prediction Using Deep Neural Network, 2019 »

Mohd Ashraf et al. [68] ont proposé l'utilisation des réseaux de neurones profonds à deux (02) couches cachées pour la prédiction des maladies cardiaques. Les auteurs ont utilisé le dataset *"Cleveland"*, après l'avoir prétraité.

Ce modèle a atteint une exactitude minimale de **87,64%**, surpassant ainsi les autres méthodes, telles que : *SVM*, *RF*, *ANN*, etc.

« Heart Diseases Prediction using Deep Learning Neural Network Model, 2020 »

Sumit Sharma et Mahesh Parmar [149] ont proposé un modèle fondé sur les réseaux de neurones profonds optimisés avec *"Talos"*, qui est une méthode récente d'optimisation hyperparamétrique, ayant pour but de permettre aux utilisateurs de continuer à travailler avec des modèles *"Keras"*. Ils ont utilisé le dataset *"Heart Disease"* provenant du référentiel UCI.

Cette approche montre de meilleurs résultats, avec une exactitude atteignant **90,78%**, dépassant ainsi les méthodes classiques, telles que : *KNN*, *SVM*, *Naive Bayes*, etc.

« Heart Disease Prediction Using Deep Neural Network, 2020 »

P. Ramprakash et al. [70] ont proposé un modèle construit à l'aide de réseaux de neurones profonds *(DNN)* à deux (02) couches et d'un modèle statistique *chi-carré*. Ils ont utilisé le dataset *"Cleveland"*, ensuite, divisé en **80%** pour l'entraînement et **20%** pour le test.

Les résultats montrent que le modèle proposé est meilleur, avec une exactitude atteignant **94%**, surpassant ainsi les méthodes traditionnelles, telles que les réseaux de neurones artificiels.

La figure 2.12 extraite de l'article [70] illustre le diagramme du modèle proposé.

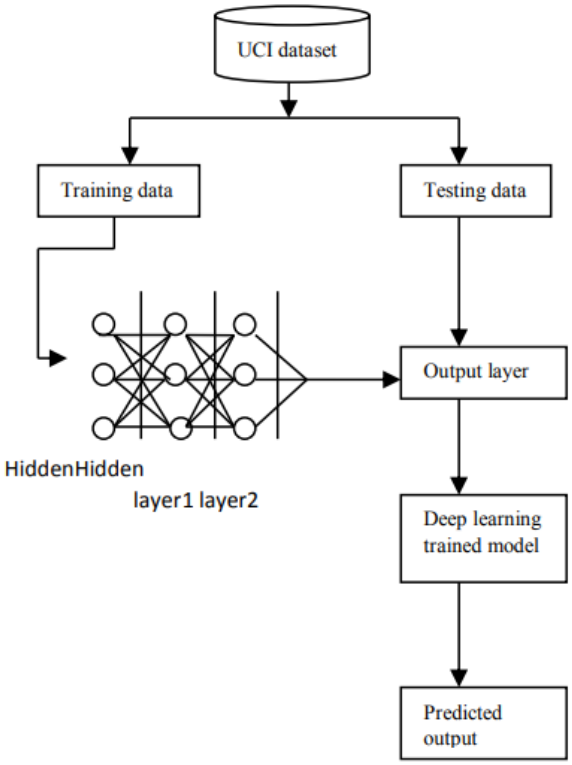


FIGURE 2.12 – Diagramme du modèle proposé par P. Ramprakash et al [70].

« An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction, 2022 »

Parmi les problèmes rencontrés dans le cadre de prédiction des maladies cardiovasculaires en utilisant les réseaux de neurones artificiels *(ANN)*, celui lié au choix optimal du nombre de couches et de neurones à mettre en œuvre. De ce fait, **Jan Carlo T. Arroyo et Allemar Jhone P. Delima** [22] ont proposé d'ajuster les paramètres de *ANN* à l'aide de l'algorithme génétique *(GA)* pour le rendre plus performant. Dans cette étude, ils ont opté pour l'utilisation d'un jeu de données comprenant 70000 enregistrements et douze (12) attributs. Le prétraitement de ces données est effectué en supprimant les duplications, les valeurs manquantes ainsi que les valeurs extrêmes, et en convertissant l'âge de jours en années ainsi que la taille de centimètres en pieds. Ceci afin d'être prêtes pour les diviser en données d'entraînement **(70%)** et de test **(30%)**. Par la suite, vient l'étape de conception du *GA-ANN* en suivant cinq (05) phases principales (initialisation, évaluation, sélection, croisement et mutation) comme décrites dans la figure 2.13.

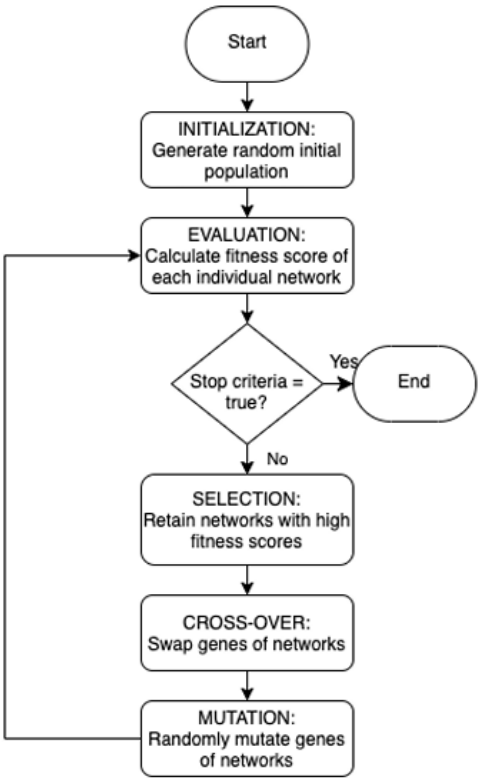


FIGURE 2.13 – Processus GA-ANN proposé par Jan Carlo T. Arroyo et Allemar Jhone P. Delima [22].

Après l'ajustement des paramètres (**3** couches, **64** neurones, *"softmax"* comme fonction d'activation et *"adagrad"* comme optimiseur), le modèle proposé atteint **73,3%** d'exactitude.

Afin d'évaluer l'efficacité de ce modèle, les auteurs ont comparé le résultat obtenu avec ceux de *ANN simple*, *LR*, *DT*, *RF*, *SVM* et *KNN*. En outre, ils ont observé une amélioration significative de **5,08%** par rapport à *ANN simple*, et que leur modèle surpassait tout les autres. Ceci montre que *GA* a optimisé *ANN*.

« Prediction of heart disease using neural network, 2022 »

Snehal B. Gavande et Prof. Pramila M. Chawan [85] ont développé un système (application web) basé sur les réseaux de neurones profonds *(DNN)* pour améliorer la précision de prédiction des maladies cardiaques. Cela pour but d'éviter les mauvais diagnostics des patients malades et de prendre les mesures nécessaires avant qu'il soit trop tard. Ils ont entraîné deux (02) modèles à l'aide de deux (02) techniques différentes : les *ANNs* (Artificial Neural Network) avec une (01) ou deux (02) couches et les *DNNs* (Deep Neural Network) ayant plusieurs couches. Le jeu de données choisi était celui de *"Cleveland"*. Avant l'étape de l'apprentissage, ils ont suivi un processus de prétraitement, d'une analyse exploratoire, et de fractionnement de données afin de rendre ces dernières prêtes à être utilisées par la suite.

Comme résultat final, l'exactitude de *DNN* ainsi que celle de *ANN* sont respectivement **95%** et **70%**. Ceci montre l'efficacité de *DNN* par rapport à *ANN*, de plus aux autres métriques calculées (précision, rappel et F1-mesure) qui confirment également sa bonne performance. Le modèle proposé a été testé sur des données en temps réel.

La figure 2.14 illustre l'architecture du système proposé.

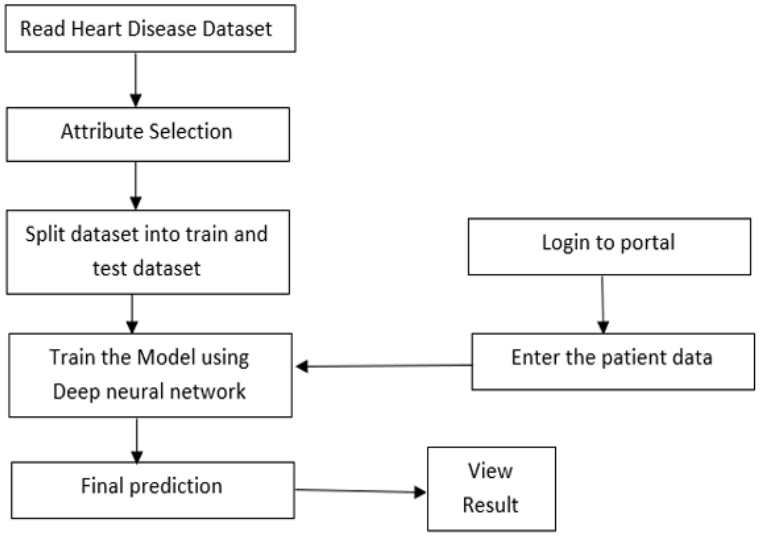


FIGURE 2.14 – Architecture du système proposé par Snehal B. Gavande et Prof. Pramila M. Chawan [85].

« Enhanced accuracy for heart disease prediction using artificial neural network, 2022 »

L'objectif principal de cette étude est d'avoir un modèle de prédiction des maladies cardiaques plus performant. Pour ce faire, **Raniya Rone Sarra et al.** [74] ont utilisé les réseaux de neurones artificiels *(ANN)*, avec une optimisation des hyperparamètres à l'aide de *random* et *grid search* ainsi que l'application d'une *validation croisée* à dix (10) plis pour vérifier et confirmer les performances du modèle. Ils ont manipulé les données issues du *"Cleveland"*.

Après l'évaluation, le modèle proposé a mené à de bons résultats, affichant **93,44%** d'exactitude, **93,30%** de rappel, **95%** d'AUC, ainsi que **93,35%** de précision et de F1-mesure. Ils ont observé une amélioration de **7,5%** par rapport à *SVM* et qu'il surpassait même les méthodes les plus performantes existantes dans la littérature pour la prédiction des maladies cardiovasculaires, telles que *DNN* (K. H. Miaoa and J. H. Miaoa, 2018) et *ANN* (Das et al., 2020). Ils ont également rapporté que la simple architecture de *ANN* contenant qu'une seule couche a permis de réduire le temps d'entraînement et de classification.

La figure 2.15 extraite de l'article [74], illustre le processus de validation croisée à dix (10) plis avec l'optimisation des hyperparamètres.

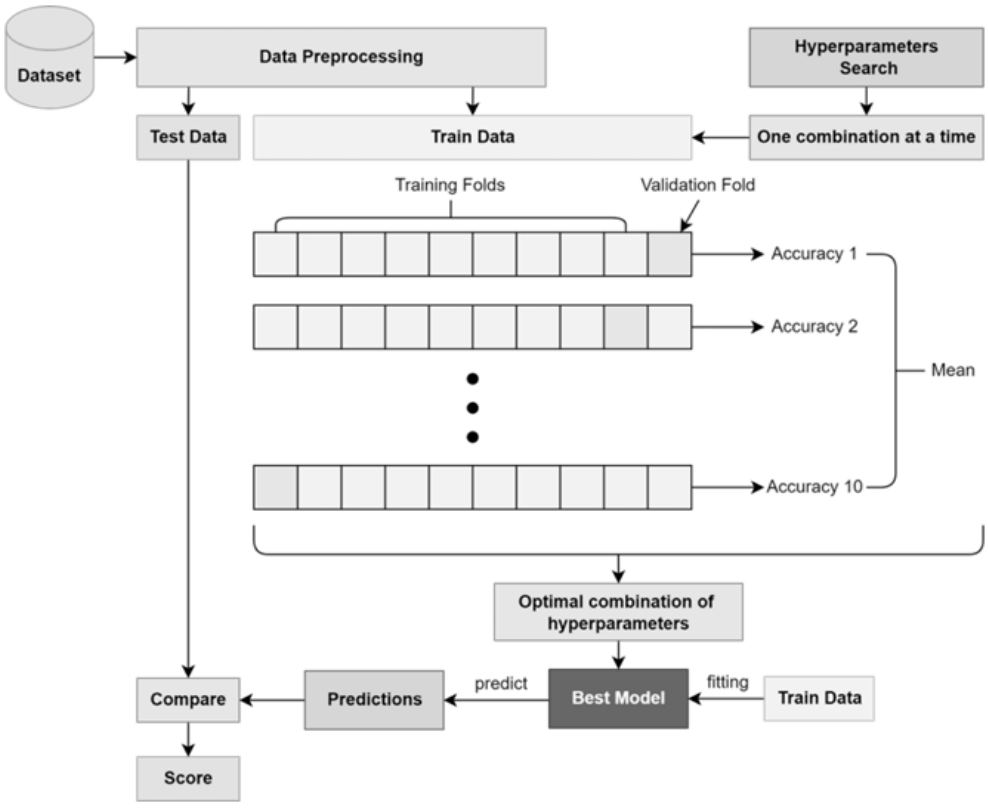


FIGURE 2.15 – Validation croisée à dix (10) plis avec l'optimisation des hyperparamètres [74].

« Heart disease risk prediction using deep learning techniques with feature augmentation, 2023 »

Dans cet article, **María Teresa García-Ordàs et al.** [67] ont proposé une nouvelle approche basée sur l'apprentissage profond et l'augmentation des caractéristiques. L'objectif clé de cette recherche était d'arriver à une méthode qui offre un pourcentage de succès élevé dans le cadre de la détection avancée du risque cardiovasculaire. Pour ce faire, ils ont utilisé la technique *SAE* (Sparse AutoEncoder) pour augmenter le nombre de variables du jeu de données mis en œuvre, contenant que onze (11) variables et 918 enregistrements. Puis, ils ont opté pour l'application de *MLP* (Multi Layer Perceptron) ou *CNN* (Convolutional Neural Network) pour la classification. Ils ont constaté que leur approche atteint une exactitude **89,543%** avec *MLP* et de **90,088%** avec *CNN*, ce qui représente un résultat significatif. Cette dernière surpasse les méthodes classiques de **4,4%** ainsi que les observations présentées dans l'état de l'art (*Adaptive Boosting*, *Bagging*, *Stacking*, *RF* et *CatBoost*). Lors de leur étude, ils ont comparé la performance de plusieurs méthodes classiques, notamment : *XGBoost*, *Gaussian Naive Bayes*, *AdaBoost*, *DT*, *KNN*, *MLP* et *RF*.

La figure 2.16 illustre l'approche proposée avec *CNN*.

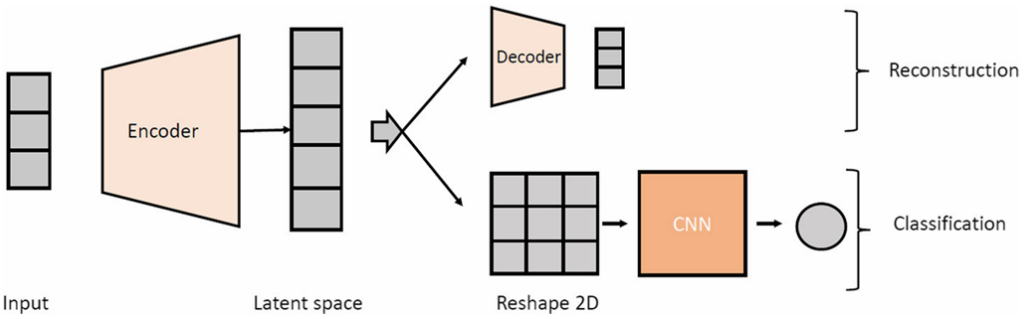


FIGURE 2.16 – Réseau neuronal multi-tâche composé d'un auto-encodeur parcimonieux et d'un classificateur CNN [67].

« A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning, 2023 »

Abdulwahab Ali Almazroi et al. [47] ont proposé une approche évaluée sur quatre (04) datasets différents : *"Cleveland"*, *"Hungarian"*, *"Switzerland"* et *"Long Beach"*, et qui repose sur un modèle basé sur les réseaux de neurones profonds. En effet, il s'agit plus précisément d'un modèle *Keras* avec des couches cachées variant de **3** à **9**, chacune comprenant **100** neurones et utilisant la fonction d'activation *ReLU*.

Les résultats montrent une exactitude atteignant **83%**. Cependant, les performances varient selon les ensembles de données en raison des différences dans les catégories d'attributs.

« Deep Learning Based Healthcare Method for Effective Heart Disease Prediction, 2023»

Loveleen Kumar et al. [65] ont proposé d'utiliser les réseaux de neurones convolutifs *(CNN)* à plusieurs couches : la couche de convolution pour extraire des caractéristiques, la couche de regroupement pour la réduction de la dimensionnalité, ainsi que des couches entièrement connectées pour effectuer la classification et estimer le risque attendu. Ils ont collecté des données en temps réel en utilisant des appareils *"Raspberry Pi"* et des *capteurs ECG*. Leur méthodologie proposée inclut le prétraitement de données recueillies, la conception de l'architecture *CNN*, l'entraînement du modèle, sa validation afin d'éviter le sur-apprentissage et de confirmer sa capacité de généralisation, l'évaluation des résultats, et enfin, l'interprétation des prédictions afin d'avoir une idée sur les caractéristiques les plus importantes, comme l'indique la figure 2.17.

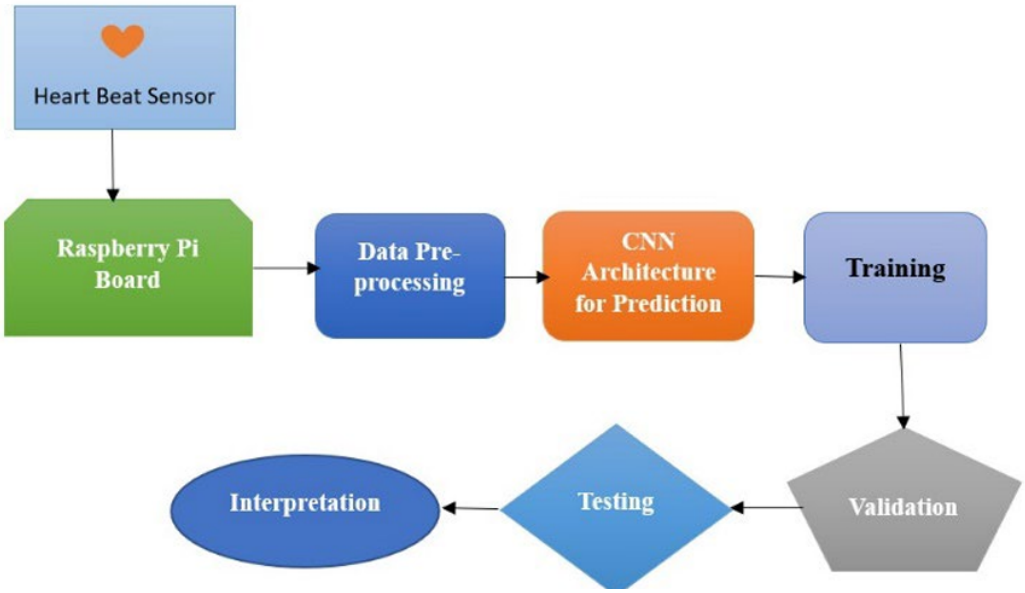


FIGURE 2.17 – Architecture du système proposé par Loveleen Kumar et al [65].

Après une étude comparative avec d'autres méthodes classiques (*KNN*, *RF*, *SVM* et *LR*), l'approche proposée atteint les meilleures performances avec une exactitude, une précision et un rappel de **86%**. Ils ont également énoncé que l'utilisation des données en temps réel et même des capteurs joue un rôle important pour un bon suivi mais nécessite une attention spécifique à la qualité, la confidentialité, et la sécurité des données collectées. Ceci représente un défi en plus du nombre limité des capteurs utilisés.

« Optimized Deep Learning Framework for Early Detection of Heart Disease, 2024 »

Afin de détecter à l'avance les maladies cardiaques et d'aider les professionnels de santé à prendre des décisions appropriées au bon moment, **Qandeel Asghar et al.** [73] ont présenté un framework optimal de l'apprentissage profond, basé sur les réseaux de neurones convolutifs *(CNN)*, intégrant un prétraitement des données (nettoyage, transformation, etc), une optimisation des hyperparamètres, une sélection des variables utiles pour la classification finale, ainsi que des auto-encodeurs empilés (Stacked AutoEncoders (*SAE*)) pour une bonne analyse de l'espace caché. Dans leur étude, ils ont utilisé un ensemble de données volumineux comprenant divers attributs, tels que les signaux électrocardiogrammes (ECG), la pression artérielle, le taux de cholestérol, etc.

Ils ont également comparé selon l'exactitude quelques algorithmes de l'apprentissage automatique (*XGBoost*, *NB*, *AD*, *KNN*, *MLP* et *RF*), puis *(MLP+SAE)* et *(CNN+SAE)* afin d'arriver à une évaluation complète, tout comme une idée générale sur les performances de ces différents modèles. Ils ont constaté que les méthodes d'ensembles présentent des performances importantes mais *CNN SAE200* atteint **90,088%** d'exactitude surpassant tous les autres algorithmes testés. Ils ont remarqué que l'utilisation de *SAE* optimise les résultats, ce qui indique leur importance pour avoir des prédictions plus précises.

Malgré les bons résultats de l'approche proposée, nous avons remarqué que le manque de transparence des modèles de l'apprentissage profond présente un défi qui peut entraver leur usage par les cliniciens. De ce fait, il est conseillé d'employer des techniques d'IA explicables comme *SHAP*, de l'apprentissage fédéré pour assurer la confidentialité des données et d'autres, afin d'avoir le modèle le plus performant et de faciliter leur emploi dans le milieu clinique.

2.2.6 Synthèse des travaux connexes utilisant des modèles d'apprentissage profond

Afin de présenter clairement les travaux de recherche analysés sur la prédiction des maladies cardiovasculaires, fondés sur l'apprentissage profond, nous avons synthétisé sous forme de tableau 3.3 les informations marquantes.

Papier	**Méthode proposée**	**Dataset utilisé**	**Évaluation**	**Observations importantes**
Liaqat Ali et al., 2019 [64].	χ^2+2-DNN.	Cleveland.	Exactitude : **93,33%**. Rappel : **85,36%**. Spécificité : **100%**. MCC : **0,872**. AUC : **94%**.	- Le modèle proposé atteint les meilleures performances, ce qui améliore la qualité de prise de décision. - Il peut résoudre les problèmes de sous-apprentissage et de sur-apprentissage. - Le test χ^2 élimine les caractéristiques les plus indépendantes (deux (02) variables ont été éliminées). - Ils ont utilisé "grid search" pour l'optimisation des hyperparamètres, ainsi qu'une validation croisée à vingt (20) plis afin de confirmer la généralisation du modèle proposé. - Le 2-DNN proposé surpasse le DNN traditionnel de 3,33%. - Le temps d'exécution n'a pas été analysé malgré son importance majeure en pratique clinique, un défi à résoudre dans les recherches futures. - La recherche exhaustive peut prendre du temps, alors, il est recommandé d'utiliser prochainement des techniques plus avancées et plus rapides comme les algorithmes génétiques.
Mohd Ashraf et al., 2019 [68].	2-DNN.	Cleveland.	Exactitude minimale : **87,64%**.	- L'approche proposée surpasse les méthodes traditionnelles, telles que : SVM, Naive Bayes et ANN.
Sumit Sharma et Mahesh Parmar, 2020 [149].	DNN optimisé avec Talos.	Heart Disease Dataset.	Exactitude : **90,78%**.	- Le modèle proposé est plus performant que les méthodes classiques, telles que : KNN, SVM, Naive Bayes, etc.
P. Ramprakash et al., 2020 [70].	2-DNN + chi-carré.	Cleveland.	Exactitude : **94%**.	- Meilleurs résultats que les méthodes traditionnelles, comme : les réseaux de neurones artificiels.

Jan Carlo T. Arroyo et Allemar Jhone P. Delima, 2022 [22].	GA-ANN.	Cardio-vascular Disease dataset.	Exactitude : **73,30%**.	- GA-ANN est plus performant que les autres modèles testés. - Amélioration de **5,08%** par rapport à ANN simple a été observée. - GA permet d'ajuster les paramètres afin d'arriver à un modèle plus efficace. - Malgré cette optimisation, le modèle atteint que **73,30%** d'exactitude, ce qui reste insuffisant pour la pratique clinique ainsi qu'il a été évalué à l'aide d'exactitude uniquement.
Snehal B. Gavande et Prof. Pramila M. Chawan, 2022 [85].	DNN.	Cleveland.	Exactitude : **95%**. Précision : • Classe 0 : **94%**. • Classe 1 : **96%**. Rappel : • Classe 0 : **97%**. • Classe 1 : **92%**. F1-mesure : • Classe 0 : **96%**. • Classe 1 : **94%**.	- DNN surpasse ANN de **25%**, ce qui représente une amélioration remarquable surtout en médecine. - Les DNNs ont plusieurs couches cachées, tandis que les ANNs ont qu'une (01) ou deux (02). - L'intégration du modèle développé dans une application web pratique aide à l'utiliser en temps réel. - Malgré le résultat obtenu, les DNNs désignent des "boîtes noires", c'est à dire leur fonctionnement interne est caché, ce qui reste incompréhensible par les médecins. Ils sont également compliqués en termes de calculs.
Raniya Rone Sarra et al., 2023 [74].	ANN amélioré.	Cleveland.	Exactitude : **93,44%**. Précision : **93,35%**. Rappel : **93,30%**. F1-mesure : **93,35%**. AUC : **95%**.	- ANN réduit le temps d'entraînement et de classification à moins d'une minute grâce à sa simplicité (une seule couche cachée). - L'ajustement des hyperparamètres et l'utilisation efficace de la validation croisée améliorent les performances du modèle. - Le modèle proposé surpasse les autres méthodes existantes en matière de prédiction des maladies cardiovasculaires. - Il est plus performant qu'un DNN sur un petit jeu de données.

María Teresa García-Ordàs et al., 2023 [67].	Approche multi-tâche avec SAE et CNN.	Jeu de données combiné (Cleveland, Hongrois, Suisse, Long Beach et Stalog).	Exactitude : **90,088%**.	- L'approche proposée surpasse les techniques classiques de **4,4%** ainsi que celles représentées dans l'état de l'art, ceci montre une amélioration importante et plus précise. - Gestion des données limitées en extrayant 200 nouvelles caractéristiques à l'aide de SAE. - Utilisation de la validation croisée à dix (10) plis ainsi que deux (02) tests statistiques (Kolmogorov Smirnov et t-test) pour confirmer la supériorité statistiquement significative par rapport aux méthodes observées. - Malgré que l'approche proposée utilise un modèle innovant, les nouvelles variables générées par SAE ne représentent pas une vérité médicale, elles sont créées mathématiquement pour la machine, ce qui influence la confiance clinique.
Abdulwahab Ali Almazroi et al., 2023 [47].	DNN.	Cleveland, Hungarian, Switzerland, et Long Beach.	Les meilleures exactitudes sont : **83,03%** pour Hungarian. **82,5%** pour Cleveland.	- L'exactitude varie selon le dataset. - Possibilité d'intégrer des images médicales, d'utiliser CNN et d'explorer des modèles hybrides.
Loveleen Kumar et al., 2023 [65].	CNN.	Données collectées en temps réel en utilisant des appareils "Raspberry Pi" et des capteurs ECG.	Exactitude : **86%**. Précision : **86%**. Rappel : **86%**.	- CNN atteint de bons résultats surpassant KNN, RF, SVM et LR. - Il est capable de classer les signaux ECG et de prédire efficacement les maladies cardiaques. - Il peut apprendre des schémas complexes. - La collecte des données en temps réel à l'aide des capteurs et des "Raspberry Pi" représente une bonne idée pour un suivi rapide et confortable pour les patients par rapport aux interventions invasives. - Malgré ces avantages, ceci exige une bonne qualité, une confidentialité, ainsi qu'une sécurité des données recueillies. - Le nombre limité de capteurs employés influence sur la quantité de données collectées.

Qandeel Asghar et al., 2024 [73].	CNN SAE200.	Un ensemble de données volumineux (son nom n'est pas mentionné).	Exactitude : **90,088%**.	- CNN SAE200 atteint l'exactitude la plus haute par rapport aux autres algorithmes testés. - Les méthodes ensemblistes sont plus efficaces par rapport à celles utilisées individuellement. - La préparation des données, l'optimisation des hyperparamètres, la sélection des caractéristiques, et le SAE sont utiles pour améliorer les performances de plusieurs algorithmes. - Les techniques de l'apprentissage profond sont pratiques en cas de données complexes et non linéaires. Cependant, elles sont considérées comme des boîtes noires, ce qui est un défi. - La confidentialité et la compréhension des données sont très importantes pour une intégration pratique en clinique.

TABLE 2.3 – Synthèse des travaux connexes utilisant des modèles d'apprentissage profond.

2.2.7 Discussion

D'après ces travaux réalisés à l'aide de divers algorithmes de l'apprentissage profond, nous observons que les approches proposées étaient à base de *DNN*s, de *CNN*s et de *ANN*s, qui ont été présents respectivement dans six (06), trois (03) et deux (02) articles différents.

Cependant, la plupart de ces modèles ont été implémentés avec des améliorations grâce à des méthodes appropriées, notamment celles qui éliminent les caractéristiques les moins pertinentes (χ^2 [64, 70]), celles qui extraient de nouvelles variables pour augmenter la taille du jeu de données (SAE [67]), celles qui optimisent les hyperparamètres (GA [22], GridSearch [64, 74], Talos [149], etc), ainsi que l'intégration de la validation croisée [64, 67, 74]. Ceci afin d'optimiser les performances des modèles de prédiction.

En évaluant les performances des approches proposées, nous constatons que la majorité atteignent des résultats significatifs, tel est le cas du modèle proposé par **Snehal B. Gavande et Prof. Pramila M. Chawan** [85] ayant **95%** d'exactitude. Cela revient à la capacité des algorithmes de l'apprentissage profond à modéliser les relations complexes, mais les considérer comme des boîtes noires est un défi, surtout dans ce domaine sensible. De ce fait, il est conseillé d'intégrer des techniques pour l'interprétation.

Nous remarquons également que la plupart des chercheurs ont utilisé l'ensemble de données *"Cleveland"* dans leurs études. Par ailleurs, les résultats diffèrent selon le jeu de données utilisé.

2.3 Idées-clés

Afin d'optimiser les performances des modèles de prédiction, il est important de :

— Utiliser un ensemble de données approprié contenant des attributs significatifs par rapport à la problématique.

— Effectuer un bon prétraitement de données pour éviter les résultats biaisés.

— Normaliser ou standardiser si nécessaire.

— Opter pour la validation croisée et des techniques de sélection ou d'augmentation des caractéristiques, d'optimisation des hyperparamètres, d'interprétabilité des données, etc.

— Choisir les modèles d'apprentissage automatique, d'apprentissage profond ou hybrides selon la simplicité ou la complexité des données d'entraînement.

2.4 Conclusion

Dans ce chapitre, nous avons passé en revue les articles les plus pertinents de la littérature sur la prédiction des maladies cardiovasculaires, et qui sont en développement continu ces dernières années. Afin d'avoir une idée préalable et claire sur les techniques utilisées dans ces travaux, nous les avons d'abord classifiés en deux (02) grandes familles, notamment ceux qui ont développé des modèles à base de l'apprentissage automatique et ceux à base de l'apprentissage profond. Pour chaque famille, nous avons mis en avant des présentations concises de différents travaux analysés, un tableau synthétique, ainsi qu'une discussion à la fin. Chaque tableau comprend la méthode proposée, l'ensemble de données utilisé, les résultats numériques obtenus à l'aide de diverses métriques d'évaluation et des observations pour chacun des articles. Ceci aide à bien présenter les nombreuses informations de manière structurée pour une comparaison facile. Les lacunes et les atouts des méthodes déjà en place, nous inspirent pour développer une nouvelle approche que nous aborderons en détail dans le prochain chapitre.

Chapitre 3

Approche proposée : Implémentation, Évaluation, et Discussion

3.1 Introduction

Ce chapitre sera consacré à la présentation et à l'explication de l'approche proposée. Nous commencerons d'abord par présenter notre problématique. Ensuite, nous aborderons brièvement notre approche. Après avoir défini les outils, le langage de programmation, et les bibliothèques utilisées, nous passerons à l'analyse de l'ensemble de données. Puis, viendra l'étape du prétraitement qui sera directement suivie de l'implémentation, pour enfin terminer avec l'évaluation et la discussion des résultats obtenus.

3.2 Problématique

Les maladies cardiovasculaires (MCV) désignent souvent la principale cause de mortalité à l'échelle mondiale. Pour diagnostiquer ce genre de maladies dangereuses, les médecins s'appuient généralement sur les symptômes observés ainsi que les informations rapportées dans le dossier médical du patient. La gravité réside surtout dans l'absence des symptômes préliminaires jusqu'à ce qu'il soit trop tard, ainsi que sur le grand nombre de facteurs que le médecin doit prendre en considération. Ceci peut provoquer des erreurs et même des retards dans le diagnostic, ce qui rend la tâche difficile et plus compliquée.

De ce fait, les chercheurs ont pensé à utiliser les différentes techniques de l'apprentissage automatique et profond pour détecter à l'avance ce genre de maladies, et sauver des vies.

Jusqu'à aujourd'hui, des recherches sont encore en cours afin d'améliorer les performances des modèles de prédiction pour les intégrer dans les systèmes de santé.

3.3 Approche proposée

Dans le chapitre précédent, nous avons présenté quelques méthodes permettant d'améliorer la performance des modèles de ML en prédiction des maladies cardiovasculaires (MCV), plus précisément, les maladies coronariennes. De cela, nous avons choisi quelques techniques qui nous ont semblé particulièrement pertinentes et complémentaires afin de les appliquer ensemble. Ces méthodes ont été utilisées dans trois (03) articles différents. Notre idée a pour but d'employer chaque technique de manière à résoudre des petits problèmes au cours de tout le processus de construction d'un modèle plus efficace.

L'approche proposée est basée sur l'apprentissage ensembliste. En effet, après la gestion du déséquilibre des classes avec *SMOTE-NC* [69] et la sélection des caractéristiques les plus importantes pour la prédiction avec l'algorithme *MRMR* [44], nous avons combiné deux (02) algorithmes de ML : *NB* et *KNN*, avec la méthode *"Stacking"* en utilisant *LR*.

La figure 3.1 représente l'architecture de l'approche proposée.

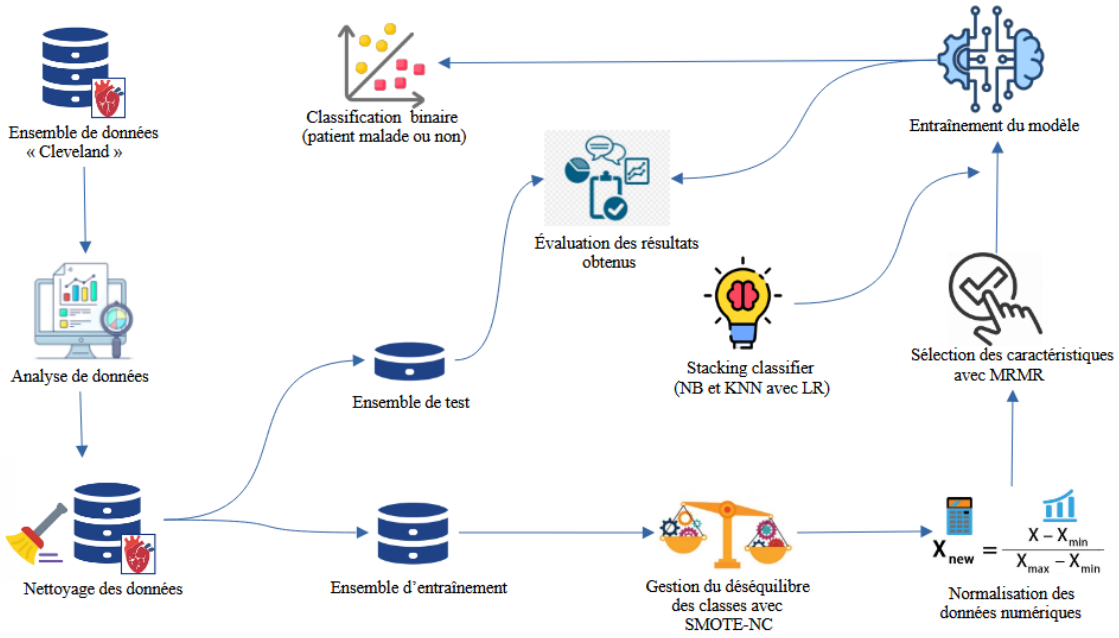


FIGURE 3.1 – Approche proposée.

3.4 Expérimentation

Cette section présente l'expérimentation réalisée afin d'évaluer l'efficacité de la méthode proposée qui débute par la description de l'environnement de développement, puis l'analyse des données, et enfin l'implémentation de l'approche.

3.4.1 Environnement de développement

Pour la réalisation de ce projet, nous avons utilisé divers outils et bibliothèques ainsi qu'un langage de programmation adapté à notre travail.

3.4.1.1 Outils

Les outils utilisés lors de la réalisation de cette approche sont les suivants :

a) **Anaconda :** C'est une distribution open source de Python et R conçue pour la science des données. Elle contient un gestionnaire de paquets et d'environnement pour l'interface de ligne de commande, appelé *"conda"*, une application de bureau construite sur ce dernier avec des options pour lancer d'autres applications de développement à partir des environnements gérés, appelé *"Anaconda Navigator"* et plus de 300 packages installés automatiquement [20].

FIGURE 3.2 – Anaconda.

b) **Spyder :** C'est un puissant environnement scientifique pour Python, qui est conçu par, et pour, des scientifiques, des ingénieurs et des analystes de données. Il combine les fonctionnalités avancées d'édition, d'analyse, de débogage, et de profilage d'un outil de développement complet avec les capacités d'exploration de données, d'exécution interactive, d'inspection approfondie, et de visualisation d'un logiciel scientifique [9].

FIGURE 3.3 – Spyder.

c) **Google Colab :** C'est un service hébergé de notebooks Jupyter qui ne nécessite aucune configuration et qui permet d'accéder sans frais à des ressources informatiques, y compris des GPU et des TPU. Colab est particulièrement adapté au machine learning (ML), à la data science, et à l'enseignement [87].

FIGURE 3.4 – Google Colab.

3.4.1.2 Langage de programmation

Dans notre travail, nous avons opté pour Python, qui est un puissant langage de programmation disposant de structures de données de haut niveau et permettant une approche simple mais efficace de la programmation orientée objet [139].

FIGURE 3.5 – Python.

3.4.1.3 Bibliothèques utilisées

Plusieurs bibliothèques ont été utilisées et qui sont les suivantes :

a) **Scikit-learn :** C'est une bibliothèque open source de ML qui prend en charge l'apprentissage supervisé et non supervisé. Elle fournit également divers outils pour l'ajustement des modèles, le prétraitement des données, la sélection des modèles, l'évaluation des modèles, ainsi que de nombreuses autres fonctionnalités [146].

b) **Imblearn :** Imbalanced-learn (importé en tant que imblearn) est une bibliothèque open source sous licence MIT, s'appuyant sur scikit-learn et fournit des outils pour traiter la classification avec des classes déséquilibrées [93].

c) **Pandas :** C'est une bibliothèque open source sous licence BSD qui offre des structures de données performantes et faciles à utiliser, ainsi que des outils d'analyse de données pour le langage de programmation Python [132].

d) **Matplotlib :** C'est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python. Elle rend les tâches simples faciles à réaliser, et les tâches complexes possibles [113].

   Parmi ses sous-modules, `matplotlib.pyplot`, qui est un ensemble de fonctions qui permet à Matplotlib de fonctionner de manière similaire à MATLAB [114].

e) **Numpy :** C'est une bibliothèque Python open source largement utilisée dans les domaines scientifiques et de l'ingénierie. Elle contient des structures de données, de tableaux multidimensionnels, ainsi qu'une vaste bibliothèque de fonctions qui opèrent efficacement sur ces structures de données [127].

f) **Seaborn :** C'est une bibliothèque pour créer des graphiques statistiques en Python. Elle s'appuie sur matplotlib et s'intègre étroitement avec les structures de données de pandas [148].

g) **MRMR :** C'est une bibliothèque python qui permet d'utiliser l'algorithme *MRMR* pour la sélection des caractéristiques pertinentes et complémentaires [103].

3.4.2 Analyse de données

L'objectif de cette partie est de mieux visualiser et comprendre l'ensemble de données.

3.4.2.1 Description de l'ensemble de données

L'ensemble de données que nous avons utilisé dans ce travail est le *"Cleveland Clinic Heart Disease Dataset"*, disponible sur Kaggle [142]. Il est composé de **303** instances et **13** variables explicatives qui correspondent à des informations médicales sur des patients, comme l'âge, le sexe, la pression artérielle, le taux de cholestérol, et une variable cible qui indique la présence ou non d'une maladie coronarienne. Dans ce jeu de données, la valeur **0** signifie l'absence de la maladie, tandis que les valeurs de **1** à **4** indiquent sa présence avec différents niveaux de gravité.

Le tableau 3.1 décrit les caractéristiques de cet ensemble de données.

Variable	**Description**	**Type**
Age	Âge du patient (Années).	Numérique
Sex	Sexe de la personne (1= homme, 0= femme).	Catégorielle
Cp	Type de douleur thoracique (1= angine typique, 2= angine atypique, 3= douleur non angineuse, 4= asymptomatique).	Catégorielle
Trestbps	Pression artérielle au repos (en mm Hg à l'admission à l'hôpital).	Numérique
Chol	Taux de cholestérol sérique en mg/dL (valeurs typiques : 126 à 564).	Numérique
Fbs	Glycémie à jeun > 120 mg/dL (1= vrai, 0= faux).	Catégorielle
Restecg	Résultat électrocardiographique au repos (0 = normal, 1 = anomalie de l'onde ST-T, 2 = hypertrophie ventriculaire gauche (critères d'Estes)).	Catégorielle
Thalach	Fréquence cardiaque maximale atteinte (valeurs typiques : 71 à 202).	Numérique
Exang	Angine induite par l'exercice (1 = oui, 0 = non).	Catégorielle
Oldpeak	Dépression ST induite par l'exercice par rapport au repos.	Numérique
Slope	Pente du segment ST à l'effort (1 = ascendante, 2 = plate, 3 = descendante).	Catégorielle
Ca	Nombre de vaisseaux majeurs (de 0 à 3) colorés par fluoroscopie.	Numérique
Thal	Résultat du test de stress nucléaire (3 = normal, 6 = défaut fixe, 7 = défaut réversible).	Catégorielle
Num	Variable cible (1, 2, 3 ou 4= présence de maladie cardiaque, 0= absence de la maladie).	Catégorielle

TABLE 3.1 – Description des variables de l'ensemble de données *"Cleveland"* utilisé pour la prédiction des maladies coronariennes [142].

3.4.2.2 Affichage de l'ensemble de données

Ci-dessous 3.6, un aperçu de l'ensemble de données contenant 303 enregistrements et 14 colonnes.

```
df
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2	7	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1	7	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1	3	1
302	38	1	3	138	175	0	0	173	0	0.0	1	?	3	0

303 rows × 14 columns

FIGURE 3.6 – Aperçu de l'ensemble de données.

3.4.2.3 Types de données

Il est nécessaire de connaître le type de chaque variable afin de choisir les traitements adaptés lors de la phase de prétraitement. Les types de données des colonnes du dataset sont présentés dans la figure 3.7.

```
#Affichage de type de données de chaque colonne
df.dtypes.to_frame().T
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	int64	int64	int64	int64	int64	int64	int64	int64	int64	float64	int64	object	object	int64

FIGURE 3.7 – Types de données.

3.4.2.4 Résumé statistique de l'ensemble de données

Pour mieux comprendre la distribution des données numériques, nous afficherons un résumé statistique des colonnes comme illustré dans la figure 3.8.

```
# Affichage d'un résumé statistique des colonnes
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.937294
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075	0.616226	1.228536
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	2.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	4.000000

FIGURE 3.8 – Résumé statistique de l'ensemble de données.

3.4.2.5 Matrice de corrélation

La matrice de corrélation permet de montrer le niveau de corrélation entre les variables d'un ensemble de données, comme montré dans la figure 3.9.

```
[11] # Matrice de corrélation
     df.corr()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1.000000	-0.097542	0.104139	0.284946	0.208950	0.118530	0.148868	-0.393806	0.091661	0.203805	0.161770	0.362605	0.127389	0.222853
sex	-0.097542	1.000000	0.010084	-0.064456	-0.199915	0.047862	0.021647	-0.048663	0.146201	0.102173	0.037533	0.093185	0.380936	0.224469
cp	0.104139	0.010084	1.000000	-0.036077	0.072319	-0.039975	0.067505	-0.334422	0.384060	0.202277	0.152050	0.233214	0.265246	0.407075
trestbps	0.284946	-0.064456	-0.036077	1.000000	0.130120	0.175340	0.146560	-0.045351	0.064762	0.189171	0.117382	0.098773	0.133554	0.157754
chol	0.208950	-0.199915	0.072319	0.130120	1.000000	0.009841	0.171043	-0.003432	0.061310	0.046564	-0.004062	0.119000	0.014214	0.070909
fbs	0.118530	0.047862	-0.039975	0.175340	0.009841	1.000000	0.069564	-0.007854	0.025665	0.005747	0.059894	0.145478	0.071358	0.059186
restecg	0.148868	0.021647	0.067505	0.146560	0.171043	0.069564	1.000000	-0.083389	0.084867	0.114133	0.133946	0.128343	0.024531	0.183696
thalach	-0.393806	-0.048663	-0.334422	-0.045351	-0.003432	-0.007854	-0.083389	1.000000	-0.378103	-0.343085	-0.385601	-0.264246	-0.279631	-0.415040
exang	0.091661	0.146201	0.384060	0.064762	0.061310	0.025665	0.084867	-0.378103	1.000000	0.288223	0.257748	0.145570	0.329680	0.397057
oldpeak	0.203805	0.102173	0.202277	0.189171	0.046564	0.005747	0.114133	-0.343085	0.288223	1.000000	0.577537	0.295832	0.341004	0.504092
slope	0.161770	0.037533	0.152050	0.117382	-0.004062	0.059894	0.133946	-0.385601	0.257748	0.577537	1.000000	0.110119	0.287232	0.377957
ca	0.362605	0.093185	0.233214	0.098773	0.119000	0.145478	0.128343	-0.264246	0.145570	0.295832	0.110119	1.000000	0.256382	0.518909
thal	0.127389	0.380936	0.265246	0.133554	0.014214	0.071358	0.024531	-0.279631	0.329680	0.341004	0.287232	0.256382	1.000000	0.509923
num	0.222853	0.224469	0.407075	0.157754	0.070909	0.059186	0.183696	-0.415040	0.397057	0.504092	0.377957	0.518909	0.509923	1.000000

FIGURE 3.9 – Matrice de corrélation.

3.4.2.6 Nombre d'occurrences

Cette étape consiste à connaître la répartition des classes cibles (présence ou absence de la maladie) pour vérifier l'équilibre des données. La figure 3.10 illustre le nombre d'occurrences de chaque classe de l'ensemble de données.

```
# Calcul de nombre d'occurrences
df['num'].value_counts()
```

num	count
0	164
1	55
2	36
3	35
4	13

dtype: int64

FIGURE 3.10 – Nombre d'occurrences.

3.4.3 Prétraitement

Avant d'utiliser les données, il est nécessaire de les préparer afin de les rendre exploitables par notre modèle.

Cette étape comprend la gestion des valeurs manquantes, dupliquées, et aberrantes, puis, la conversion des valeurs de la colonne cible en deux (02) classes.

3.4.3.1 Gestion des valeurs manquantes

Avant de passer à la répartition des données, il est essentiel de vérifier s'il existe des valeurs manquantes dans l'ensemble de données et de les gérer pour garantir la qualité des résultats du modèle.

Dans notre ensemble de données, nous avons trouvé six (06) valeurs manquantes (quatre (04) pour la variable "ca" et deux (02) pour "thal"). Ceci affecte grandement la qualité de la classification. Pour cela, nous avons décidé de les supprimer.

3.4.3.2 Gestion des duplications

Dans un ensemble de données, il est possible de trouver des lignes (enregistrements) redondantes ou presque redondantes dont la plupart des colonnes contiennent les mêmes valeurs.

De ce fait, *la déduplication* est couramment utilisée dans le but de la résolution de ce genre de problèmes qui peuvent fausser l'apprentissage et affecter la performance des modèles utilisés [133].

Après vérification, nous avons constaté qu'il n'existe pas de données dupliquées.

3.4.3.3 Gestion des valeurs aberrantes

Les valeurs aberrantes (outliers) sont des valeurs anormales, autrement dit, les données se distinguent nettement des autres par leurs caractéristiques globales ou par certaines de leurs valeurs spécifiques [133]. Cela a un impact négatif sur les algorithmes d'apprentissage automatique qui sont sensibles aux valeurs extrêmes.

Notre ensemble de données ne dispose pas de valeurs aberrantes.

3.4.3.4 Conversion des valeurs de la colonne cible en deux (02) classes

La variable cible *"num"* de notre ensemble de données représente l'absence de la maladie lorsque l'étiquette affiche **0** et sa présence en cas de **1**, **2**, **3** ou **4** [142].

Afin de réduire la complexité et de bien préparer les données à fonctionner correctement, nous avons pensé à binariser la colonne, de sorte que la valeur **0** indique l'absence de la maladie, contrairement à la valeur **1** qui signifie sa présence.

Après la conversion, la variable cible présente deux (02) classes (0 et 1), comme illustré dans la figure 3.11.

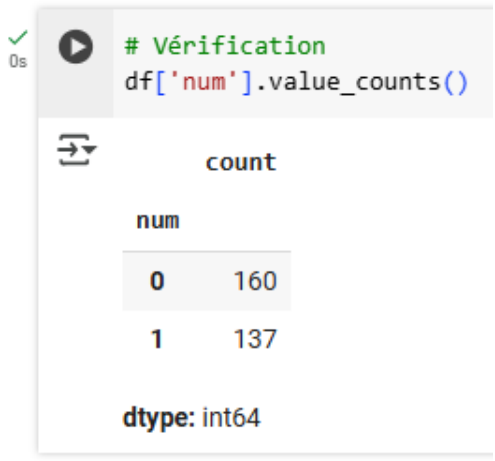


FIGURE 3.11 – Vérification de la conversion des valeurs de la colonne cible en deux (02) classes.

3.4.4 Répartition des données

Après l'étape de prétraitement, les données sont prêtes à être divisées en deux (02) sous-ensembles distincts, l'un permet d'entraîner le modèle (train set) tandis que l'autre sert à évaluer ses performances sur des nouvelles données (test set) en utilisant la fonction `train_test_split()` de `scikit-learn`.

Dans notre démarche, **80%** de nos données ont été sélectionnées comme ensemble d'apprentissage et les **20%** restantes conservées pour le test comme le montre la figure 3.12.

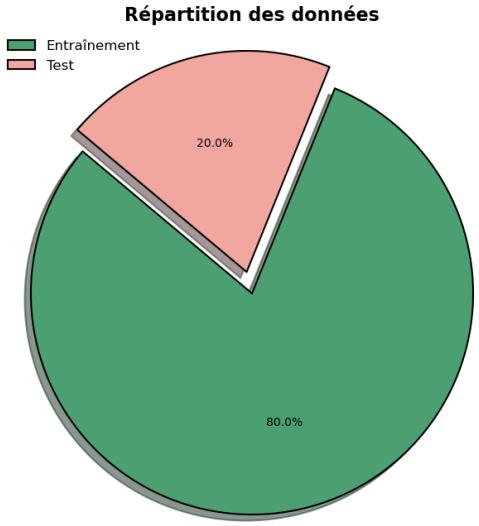


FIGURE 3.12 – Répartition des données.

3.4.5 Gestion du déséquilibre des classes avec SMOTE-NC

Selon l'analyse des données de l'ensemble utilisé pour l'entraînement, la distribution des classes est donnée dans la figure 3.13.

```
0s
# Pourcentage de présence de chaque classe dans l'ensemble d'entraînement
print('Le pourcentage de chaque classe est')
print(Y_train.value_counts(normalize=True) * 100)

Le pourcentage de chaque classe est
num
0    52.320675
1    47.679325
Name: proportion, dtype: float64
```

FIGURE 3.13 – Pourcentage de présence de chaque classe dans l'ensemble d'entraînement.

Cette répartition montre que la classe **1** est sous-représentée par rapport à la classe **0** de **4.64135%**, ce qui peut influencer les performances des algorithmes qui sont souvent biaisés vers la classe qui a plus d'exemples.

Le problème de données déséquilibrées en apprentissage automatique peut être résolu en utilisant l'une des approches de rééchantillonnage suivantes [75] :

— **Suréchantillonnage (Oversampling)**, qui permet d'augmenter le nombre d'exemples de la classe minoritaire, soit par des duplications simples, soit par génération de nouvelles instances synthétiques.
Exemples : SMOTE [69] et ADASYN [58].

— **Sous-échantillonnage (Undersampling)**, qui permet de réduire le nombre d'exemples de la classe majoritaire en suppriment certains de ses enregistrements.
Exemples : Random Undersampling [118] et Cluster Centroids [105].
La figure 3.14 illustre la difference entre le suréchantillonnage et le sous-échantillonnage.

FIGURE 3.14 – Suréchantillonage et sous-échantillonnage [75].

De ce fait, nous avons pensé à gérer ce petit déséquilibre à l'aide de la méthode *SMOTE* (Synthetic Minority Oversampling TEchnique) pour obtenir un modèle plus juste et plus efficace.

3.4.5.1 SMOTE (Synthetic Minority Oversampling TEchnique)

SMOTE est une technique de suréchantillonnage synthétique des classes minoritaires, ce qui revient à dire qu'elle génère de nouveaux exemples artificiels à partir des données déjà existantes de la classe peu fréquente en se basant sur les k plus proches voisins, et ce pour augmenter le pourcentage de cette classe afin d'arriver à un équilibre [69].

3.4.5.2 Fonctionnement de SMOTE

Contrairement aux méthodes de duplication simples, *SMOTE* fonctionne dans l'espace des caractéristiques. L'idée principale de cette technique consiste à :

1) Choisir aléatoirement un échantillon de la classe minoritaire et de trouver ses k plus proches voisins appartenant également à la même classe (classe minoritaire), sachant que le nombre de voisins à opter pour la suite des étapes dépend de pourcentage de suréchantillonnage ciblé.
2) Créer de nouveaux exemples synthétiques entre l'échantillon et ses voisins les plus proches après avoir tracé la ligne reliant ces deux (02) points et de calculer la différence entre eux.
3) Multiplier la différence par un nombre aléatoire entre 0 et 1 puis l'ajouter à la valeur initiale de l'échantillon choisi.
4) Répéter ce processus jusqu'à atteindre le nombre d'exemples synthétiques désirés afin d'arriver à un équilibre entre classes.

Le pseudo code suivant extrait de l'article [69] illustre le fonctionnement de cette technique :

Algorithm 1 SMOTE (T, N, k) [69]

Input : Number of minority class samples T ; Amount of SMOTE $N\%$; Number of nearest neighbors k
Output : $(N/100) * T$ synthetic minority class samples

(∗ If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. ∗)
if $N < 100$ **then**
    Randomize the T minority class samples
    $T = (N/100) * T$
    $N = 100$
end if
$N =$ (int)$(N/100)$ (∗ The amount of SMOTE is assumed to be in integral multiples of 100. ∗)
$k =$ Number of nearest neighbors
$numattrs =$ Number of attributes
$Sample[\,][\,]$: array for original minority class samples
$newindex$: keeps a count of number of synthetic samples generated, initialized to 0
$Synthetic[\,][\,]$: array for synthetic samples
    (∗ Compute k nearest neighbors for each minority class sample only. ∗)
for $i \leftarrow 1$ to T **do**
    Compute k nearest neighbors for i, and save the indices in the $nnarray$
    POPULATE(N, i, $nnarray$)
end for
    POPULATE(N, i, $nnarray$) (∗ Function to generate the synthetic samples. ∗)
while $N \neq 0$ **do**
    Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
    for $attr \leftarrow 1$ to $numattrs$ **do**
        Compute : $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        Compute : $gap =$ random number between 0 and 1
        $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
    end for
    $newindex++$
    $N = N - 1$
end while
Return (∗ End of Populate. ∗)
    (End of Pseudo-Code.)

3.4.5.3 Avantages et inconvénients de SMOTE

Comme tout algorithme de gestion de déséquilibre des classes, *SMOTE* présente des avantages mais aussi des inconvénients [19, 29, 38, 69], comme le montre le tableau 3.2.

Avantages	**Inconvénients**
• Résolution de problème de déséquilibre des classes d'un jeu de données. • Amélioration de la performance des algorithmes d'apprentissage automatique, particulièrement ceux sensibles au déséquilibre des classes en réduisant le biais envers la classe majoritaire. • Réduction du sur-apprentissage lié à la duplication simple grâce aux nouveaux exemples synthétiques créés. • Pas de perte d'informations importantes. • Simplicité d'implémentation.	• Risque de sur-apprentissage en cas d'un pourcentage très élevé de suréchantillonnage ou les exemples générés sont peu diversifiés. • Possibilité d'utiliser des exemples de la classe minoritaire prés de la classe majoritaire, ce qui augmente le risque de confusion entre classes. • Moins efficace pour les données de grandes dimensions. • Les exemples créés artificiellement peuvent présenter une distribution différente de la réalité, ce qui peut fausser l'apprentissage. • SMOTE classique est conçu pour les variables numériques, ce qui provoque des échantillons invalides en cas de présences des caractéristiques catégorielles dans l'ensemble de données.

Table 3.2 – Avantages et inconvénients de SMOTE [19, 29, 38, 69].

3.4.5.4 SMOTE-NC (Synthetic Minority Oversampling TEchnique Nominal Continuous)

La technique *SMOTE* a été généralisée pour traiter des jeux de données contenant des caractéristiques numériques et catégorielles. Cette approche nommée *SMOTE-NC* (Synthetic Minority Oversampling Technique Nominal Continuous) fonctionne comme suit [69] :

1) Calcul de la médiane des écarts types de toutes les variables numériques de la classe minoritaire.
2) Recherche des k plus proches voisins en se basant sur le calcul de la distance euclidienne entre le point sélectionné aléatoirement de la classe minoritaire et les autres points de cette même classe, en utilisant l'espace des caractéristiques continues, tout en incluant la médiane calculée précédemment en cas de différence entre les variables nominales du point et ses voisins les plus proches.
3) Génération des échantillons synthétiques de la même façon que dans l'approche *SMOTE* pour les variables numériques tandis que les autres (variables nominales) prennent la valeur la plus courante des k plus proches voisins.

Exemple de calcul des voisins les plus proches pour SMOTE-NC [69] :

F1 = 1 2 3 A B C [Let this be the sample for which we are computing nearest neighbors]

F2 = 4 6 5 A D E

F3 = 3 5 6 A B K

So, Euclidean Distance between F2 and F1 would be :

$$\text{Eucl} = \sqrt{(4-1)^2 + (6-2)^2 + (5-3)^2 + \text{Med}^2 + \text{Med}^2}$$

Med is the median of the standard deviations of continuous features of the minority class.

The median term is included twice for feature numbers 5 : B→D and 6 : C→E,

which differ for the two feature vectors : F1 and F2.

En s'appuyant sur ces informations extraites de l'article [69], nous avons choisi d'utiliser *SMOTE-NC* au lieu de *SMOTE* en raison de données mixtes (numériques et catégorielles) du jeu de données utilisé.

Après l'application de *SMOTE-NC*, la distribution des classes est comme montré dans la figure 3.15.

```
[47] # Pourcentage de présence de chaque classe dans l'ensemble d'entraînement après l'application de SMOTE-NC
     print("Le pourcentage de présence de chaque classe dans l'ensemble d'entraînement après l'application de SMOTE-NC")
     print(Y_train.value_counts(normalize=True) * 100)

Le pourcentage de présence de chaque classe dans l'ensemble d'entraînement après l'application de SMOTE-NC
num
0    50.0
1    50.0
Name: proportion, dtype: float64
```

FIGURE 3.15 – Résultat après l'application de SMOTE-NC.

Résultat : Classes équilibrées.

3.4.6 Mise à l'échelle des données (scaling)

La mise à l'échelle est un processus important en apprentissage automatique dans l'étape de prétraitement des données afin d'assurer le bon fonctionnement des différents algorithmes [152], notamment :

— **Normalisation Min-Max :** C'est une méthode qui permet de transformer les données et de les mettre à l'échelle dans un intervalle borné généralement entre **0** et **1** ou **-1** et **1** en suivant la formule ci-dessous :

$$x_{\text{normalisé}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

Où : x représente la valeur initiale, $x_{\min}$ et $x_{\max}$ correspondent respectivement aux valeurs minimales et maximales de la caractéristique.

— **Normalisation Z-score (standardisation) :** C'est une technique qui permet de centrer et réduire les données à l'aide de la moyenne et de l'écart type en utilisant la formule suivante :

$$x_{\text{standardisé}} = \frac{x - \mu}{\sigma} \tag{3.2}$$

Où : μ représente la moyenne de la caractéristique et σ correspond à l'écart type.

Remarque : L'importance de l'usage de ces techniques de la mise à l'échelle dépend principalement des algorithmes utilisés. C'est-à-dire qu'il existe des algorithmes sensibles à l'échelle des données qui nécessitent leur utilité dans le but d'améliorer les performances des modèles de classification tels que *KNN* et *SVM*.

Dans notre cas, nous avons choisi d'utiliser la normalisation *MinMax* afin de bien préparer les données numériques d'entraînement pour les algorithmes à employer dans les prochaines étapes.

```
# Normalisation des données numériques continues
from sklearn.preprocessing import StandardScaler, MinMaxScaler

# Colonnes à normaliser
cols_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']

# Initialiser le scaler
scaler = MinMaxScaler()
X_train= X_train.copy()
X_test = X_test.copy()
X_train[cols_to_scale] = scaler.fit_transform(X_train[cols_to_scale])
X_test[cols_to_scale] = scaler.transform(X_test[cols_to_scale])

X_train.head(5)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	0.208333	0	3	0.415094	0.214612	0	0	0.618321	0	0.000000	2	0	3
1	0.645833	0	1	0.528302	0.260274	0	0	0.763359	0	0.160714	1	0	3
2	0.833333	0	1	0.433962	0.257991	0	0	0.610687	0	0.321429	1	2	3
3	0.604167	1	2	0.245283	0.360731	0	2	0.679389	0	0.321429	2	0	3
4	0.375000	1	3	0.339623	0.289954	0	0	0.824427	0	0.000000	1	0	3

FIGURE 3.16 – Normalisation des données numériques.

Après avoir exécuté les instructions montrées dans la figure 3.16, nous avons obtenu des données numériques normalisées prêtes à être utilisées par la suite.

3.4.7 Sélection des caractéristiques importantes avec MRMR

La sélection des caractéristiques (feature selection) est une technique utilisée en apprentissage automatique afin de choisir les attributs les plus importants parmi les variables de l'ensemble de données initial, qui vont être utilisés pour la création des modèles simples, performants, plus efficaces, et plus rapides en entraînement [141]. Ce processus est divisé généralement en quatre (04) méthodes principales [107] :

— **Méthodes de filtrage (Filter methods)**, qui présélectionnent les attributs les plus pertinents pour la cible avant l'apprentissage en utilisant des critères statistiques simples tels que : la corrélation, la variance et l'information mutuelle [1], en gardant uniquement les meilleurs k attributs ayant les scores les plus élevés.
Exemples : Test ANOVA (Analysis Of Variance) [100] et Best First Search [98].

— **Méthodes d'enveloppes ou d'encapsulation (Wrapper methods)**, qui sélectionnent le sous-ensemble de données qui donne les meilleurs résultats après avoir utilisé un modèle de classification, et d'évaluer ses performances à l'aide de plusieurs métriques, telles que l'exactitude, la précision, le rappel et le F1 score. Elles capturent les relations complexes entre les variables.
Exemples : SFS (Sequential Forward Selection) [71], SBS (Sequential Backward Selection) [49] et les algorithmes génétiques [22].

— **Méthodes intégrées ou embarquées (Embedded methods)**, qui sélectionnent automatiquement les meilleures caractéristiques durant le processus d'entraînement du modèle en utilisant des mécanismes internes des algorithmes d'apprentissages choisis. C'est notamment le cas des arbres de décision (DT) (importance des caractéristiques) et les machines à vecteurs de support (SVM) avec la régularisation L1.

— **Méthodes hybrides**, qui combinent des méthodes de filtrage avec celles d'enveloppes ou intégrées pour éliminer d'abord les variables les moins importantes (filtrage), puis les utiliser dans des méthodes plus avancées (enveloppes ou intégrées) pour optimiser la sélection et équilibrer le compromis entre la vitesse, la précision, et la complexité.
Exemples : Chi2+Random Forest.

1. Information mutuelle : Une mesure qui s'appuie sur la théorie de l'information (Shannon, 1948) pour quantifier la dépendance entre deux (02) variables aléatoires

3.4.7.1 MRMR (Minimum Redundancy Maximum Relevance)

Il existe plusieurs méthodes de sélection des caractéristiques, notamment celle appelée *Minimum Redundancy Maximum Relevance* (MRMR) qui fait partir des méthodes de filtrage. Elle est basée sur l'idée d'avoir un sous-ensemble de données pertinentes, à fort pouvoir prédictif, peu redondantes, et à faible interdépendance. L'évaluation des caractéristiques sera conformément à leur type de données (discrètes ou continues) à l'aide de différentes formules mentionnées dans le tableau 3.3 [41].

Type	**Acronym**	**Full Name**	**Formula**
Discrete	MID	Mutual information difference	$\max_{i \in \Omega_S} \left[I(i,h) - \frac{1}{\lvert S \rvert} \sum_{j \in S} I(i,j) \right]$ (3.3)
	MIQ	Mutual information quotient	$\max_{i \in \Omega_S} \left\{ I(i,h) \Big/ \left[\frac{1}{\lvert S \rvert} \sum_{j \in S} I(i,j) \right] \right\}$ (3.4)
Continuous	FCD	F-test correlation difference	$\max_{i \in \Omega_S} \left[F(i,h) - \frac{1}{\lvert S \rvert} \sum_{j \in S} \lvert c(i,j) \rvert \right]$ (3.5)
	FCQ	F-test correlation quotient	$\max_{i \in \Omega_S} \left\{ F(i,h) \Big/ \left(\frac{1}{\lvert S \rvert} \sum_{j \in S} \lvert c(i,j) \rvert \right) \right\}$ (3.6)

TABLE 3.3 – Divers schémas de recherche de la caractéristique suivante selon les critères d'optimisation MRMR [41].

Où :

— S signifie le sous-ensemble de caractéristiques sélectionnées et |S| correspond à leur nombre.

— I(i,j) et I(i,h) représentent respectivement l'information mutuelle entre les deux (02) caractéristiques i et j, ainsi que celle entre i et la cible h.

— F(i,h) désigne le score de test F entre la caractéristique i et la cible h.

— c(i,j) est la corrélation entre les deux (02) caractéristiques i et j.

Remarque : I(i,h) et F(i,h) représentent le niveau de pertinence tandis que I(i,j) et c(i,j) permettent de mesurer la redondance.

Celle qui utilise l'information mutuelle pour l'évaluation est définit formellement [32] comme suit :

$$\text{MRMR} : \max \left[\frac{1}{|S|} \sum_{i \in S} I(i;h) - \frac{1}{|S|^2} \sum_{i \in S} \sum_{j \in S} I(i,j) \right] \tag{3.7}$$

3.4.7.2 Étapes de l'algorithme MRMR

Les étapes clés de l'algorithme itératif *MRMR* sont les suivantes [44] :

1) Évaluation de la pertinence de chaque caractéristique de l'ensemble de données initial.
2) Sélection de la caractéristique la plus pertinente.
3) Évaluation de la redondance entre la caractéristique choisie et celles qui sont restantes.
4) Calcul du score d'importance en utilisant la différence ou le quotient entre la pertinence et la redondance.
5) Sélection de la caractéristique qui dispose d'un score maximal et de l'ajouter à l'ensemble de caractéristiques sélectionnées.
6) Calcul de la redondance moyenne entre les caractéristiques restantes et celles qui ont été sélectionnées.
7) Calcul du score d'importance, la sélection de la caractéristique avec la valeur maximale, puis l'ajouter à l'ensemble de caractéristiques déjà sélectionnées, et ainsi de suite jusqu'à l'atteinte de nombre désiré de variables qui désigne le critère d'arrêt.

Le pseudo code de cet algorithme est comme suit :

Algorithm 2 MRMR (Minimum Redundancy Maximum Relevance) [44]

Entrée :
    X : Ensemble de variables
    h : Variable cible
    k : Critère d'arrêt (nombre désiré de variables à sélectionner)
Sortie :
    S : Sous-ensemble de variables sélectionnées
$S \leftarrow \emptyset$
Évaluer la pertinence de chaque variable $x_i \in X$ par rapport à la cible h
Sélectionner la variable la plus pertinente x_{p}
$S \leftarrow S \cup \{x_{\text{p}}\}$
$X \leftarrow X \setminus \{x_{\text{p}}\}$
while $|S| < k$ **do**
    for chaque variable $x_i \in X$ **do**
        Calculer la pertinence de x_i par rapport à la cible h
        Calculer la redondance moyenne entre x_i et les variables de S
        Calculer le score d'importance de x_i en utilisant la différence ou le quotient entre la pertinence et la redondance
    end for
    Sélectionner la variable ayant le score maximal $x_{\max}$
    $S \leftarrow S \cup \{x_{\max}\}$
    $X \leftarrow X \setminus \{x_{\max}\}$
end while
return S

3.4.7.3 Avantages et inconvénients de l'algorithme MRMR

Comme tout algorithme de sélection des caractéristiques, *MRMR* présente des avantages, mais aussi des inconvénients [41, 43] comme l'illustre le tableau 3.4.

Avantages	**Inconvénients**
• Maximisation de la pertinence avec la classe cible. • Minimisation de la redondance des caractéristiques. • Amélioration des performances des modèles de l'apprentissage automatique en sélectionnant que les caractéristiques les plus importantes, ce qui permet de réduire le sur-apprentissage. • Diversité des mesures d'évaluation comme l'information mutuelle et la corrélation conformément au type de données. • Simplicité de fonctionnement.	• Détermination de nombre de caractéristiques à sélectionner à l'avance ce qui représente un défi sans expertise du domaine. • Complexité élevée en termes de calcul pour les ensembles de données volumineux. • Nécessité de jeux de données supervisés. • Incapacité à détecter les relations non linéaires car il analyse les caractéristiques deux à deux. • L'évaluation dépend de la méthode de calcul.

TABLE 3.4 – Avantages et inconvénients de MRMR [41, 43].

Compte tenu de ces informations, nous avons pensé à utiliser cette technique pour sélectionner les dix (10) caractéristiques essentielles parmi treize (13) de l'ensemble de données *"Cleveland"* en utilisant la fonction `mrmr_classif` de la bibliothèque `mrmr` comme illustré dans la figure 3.17.

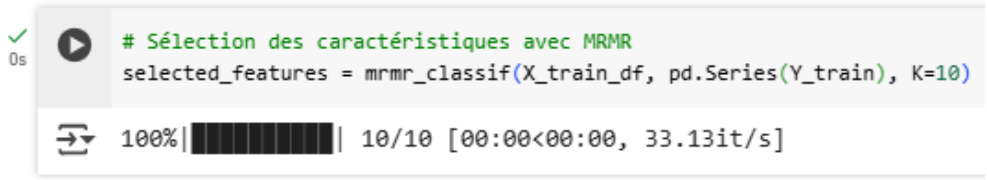


FIGURE 3.17 – Sélection des caractéristiques avec MRMR.

Le choix du nombre de caractéristiques (k=10) a été choisi après avoir effectué plusieurs essais afin d'arriver au meilleur paramètre.

Les dix (10) caractéristiques sélectionnées par cet algorithme sont : ['thal', 'ca', 'exang', 'oldpeak', 'thalach', 'sex', 'cp', 'slope', 'trestbps' et 'restecg'].

3.4.8 Construction et entraînement du modèle

Étant donné que nous avons utilisé le *stacking* pour la construction de notre modèle, il est avant tout, nécessaire de définir cette notion afin de mieux comprendre le fonctionnement de notre approche.

3.4.8.1 Définition du stacking

Le *stacking* est une technique d'apprentissage ensembliste qui combine plusieurs modèles de base, puis utilise leurs prédictions comme entrées pour un modèle de second niveau afin d'effectuer la prédiction finale [135].

3.4.8.2 Fonctionnement du stacking

Contrairement au *bagging* ou au *boosting*, qui agrègent les prédictions de manière simple (moyenne ou vote), le *stacking* utilise un méta-modèle pour une meilleure fusion des résultats. Voici son fonctionnement détaillé en **cinq (05) étapes** [28] :

1) **Division des données :** Les données sont divisées en ensemble de test et ensemble d'entraînement. La validation croisée (k-fold CV) sera appliquée sur ce dernier pour éviter les problèmes de sur-apprentissage.
2) **Entraînement des modèles de base (Niveau 0) :** Entraînement de plusieurs algorithmes différents indépendamment sur les données, appelés *"base learners"* et qui permettent d'apporter des perspectives complémentaires.
3) **Création des méta-caractéristiques :** Grâce à la *validation croisée*, chaque modèle de base est entraîné sur une partie qu'il n'a pas vu pendant son apprentissage. Ces nouvelles prédictions sont ensuite rassemblées en une nouvelle matrice appelée *"ensemble de méta-caractéristiques (ou méta-features)"* qui servira d'entrée au méta-modèle.
4) **Entraînement du méta-modèle (Niveau 1) :** Choix d'un simple algorithme à entraîner sur les méta-données, tel que la régression linéaire ou logistique.
5) **Prédiction finale :** Utilisation des modèles de base pour la prédiction sur l'ensemble de test. Puis, le méta-modèle fusionne ces résultats et les pondère afin de produire une prédiction finale optimisée.

3.4.8.3 Avantages et inconvénients du stacking

Le *stacking* présente plusieurs points forts qui font de lui une technique puissante pour améliorer la performance prédictive. Toutefois, il comporte aussi quelques points faibles à ne pas négliger [10] comme indiqué dans le tableau 3.5.

Avantages	**Inconvénients**
• Il améliore la performance globale en combinant les prédictions de plusieurs modèles. • Il exploite les forces de chaque modèle afin d'atténuer leurs faiblesses. • Il élimine la nécessité d'avoir une connaissance approfondie de l'algorithme le plus approprié. • Il identifie dynamiquement les meilleurs modèles pour chaque ensemble de données.	• L'entraînement de plusieurs modèles prend plus de temps qu'un seul modèle. • Mise en œuvre complexe à cause du vaste choix de paramètres. • Énorme consommation de ressources. • Problèmes de généralisation entre entraînement et test (obligation d'utiliser la validation croisée).

TABLE 3.5 – Avantages et inconvénients du stacking [10].

3.4.8.4 Choix des algorithmes

Dans notre travail, nous avons utilisé le *stacking* pour combiner deux (02) algorithmes qui sont : *Naive Bayes* (NB) [18, 154] et *K-Nearest Neighbors* (KNN) qui atteignent respectivement **93,33%** et **90%** d'exactitude, en utilisant la *régression logistique* (LR).

Notre choix ne s'est pas fait au hasard. L'application du *stacking* exige que les algorithmes choisis fonctionnent de façon différente, et puissants, ce qui est notre cas. En effet, cette condition a été respectée car KNN est basé sur la distance, tandis que NB utilise une distribution probabiliste.

Pour la prédiction finale, il est généralement conseillé d'utiliser la régression linéaire ou logistique. Et nous avons opté pour cette dernière.

Remarque : Le choix du paramètre k=4 de l'algorithme KNN a été effectué en utilisant *la méthode de coude*.

3.4.8.5 Fonctionnement du modèle proposé

Après la phase du prétraitement, vient celle de l'exploitation des données.

D'abord, et après la division des données (ensemble de test et ensemble d'entraînement), la validation croisée (k-fold CV) est appliquée sur l'ensemble d'entraînement pour garantir que le modèle apprend à chaque fois sur des données qu'il n'a pas rencontrées précédemment afin d'éviter les problèmes de sur-apprentissage.

Ensuite, entraîner indépendamment *NB* et *KNN* sur les données d'entraînement.

Puis, les résultats de prédiction des deux (02) algorithmes précédents seront utilisés comme entrée au méta-modèle, qui est *LR*.

Enfin, *KNN* et *NB* seront utilisés pour la prédiction sur l'ensemble de test. Alors, la *régression logistique (LR)* fusionne ces résultats pour en sortir avec une prédiction finale.

Pour résumer, nous avons effectué deux (02) entraînements indépendants *(KNN, et NB)*. Les résultats de prédiction de ces deux (02) algorithmes ont été utilisés comme entrées du méta-modèle (*LR* dans notre cas) afin d'arriver à une prédiction finale plus efficace.

La figure 3.18 représente la logique de fonctionnement du modèle proposé.

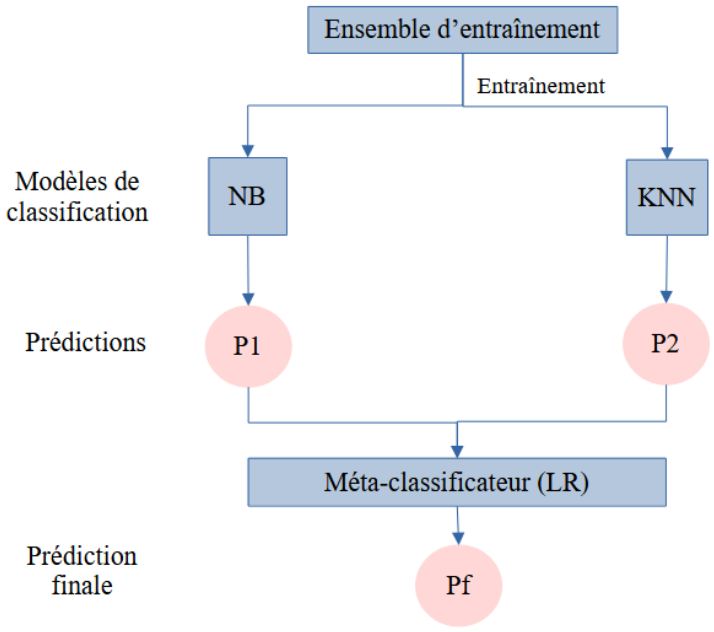


FIGURE 3.18 – Fonctionnement du modèle.

3.5 Évaluation des résultats et discussion

L'étape qui suit la construction et l'entraînement du modèle correspond à l'évaluation de ses performances afin de vérifier sa capacité de généralisation sur des données différentes de celles utilisées pour l'apprentissage. Cela vise à simuler son fonctionnement dans la réalité.

Dans cette section, nous présenterons en premier lieu la matrice de confusion ainsi que les métriques d'évaluation de notre modèle de classification obtenues après l'implémentation. Par la suite, nous les utiliserons pour analyser et discuter ces résultats. Cela en vue de valider l'approche proposée dans le cadre de prédiction des maladies cardiovasculaires.

3.5.1 Évaluation et comparaison des résultats obtenus

Après l'étape de l'entraînement du modèle, nous avons analysé les résultats de l'évaluation obtenus en présence et en absence des techniques exposées précédemment (*SMOTE-NC* et *MRMR*). Ceci dans le cadre de classification des patients atteints d'une maladie cardiovasculaire ou non.

Les matrices de confusion obtenues des modèles testés sont données dans le tableau 3.6.

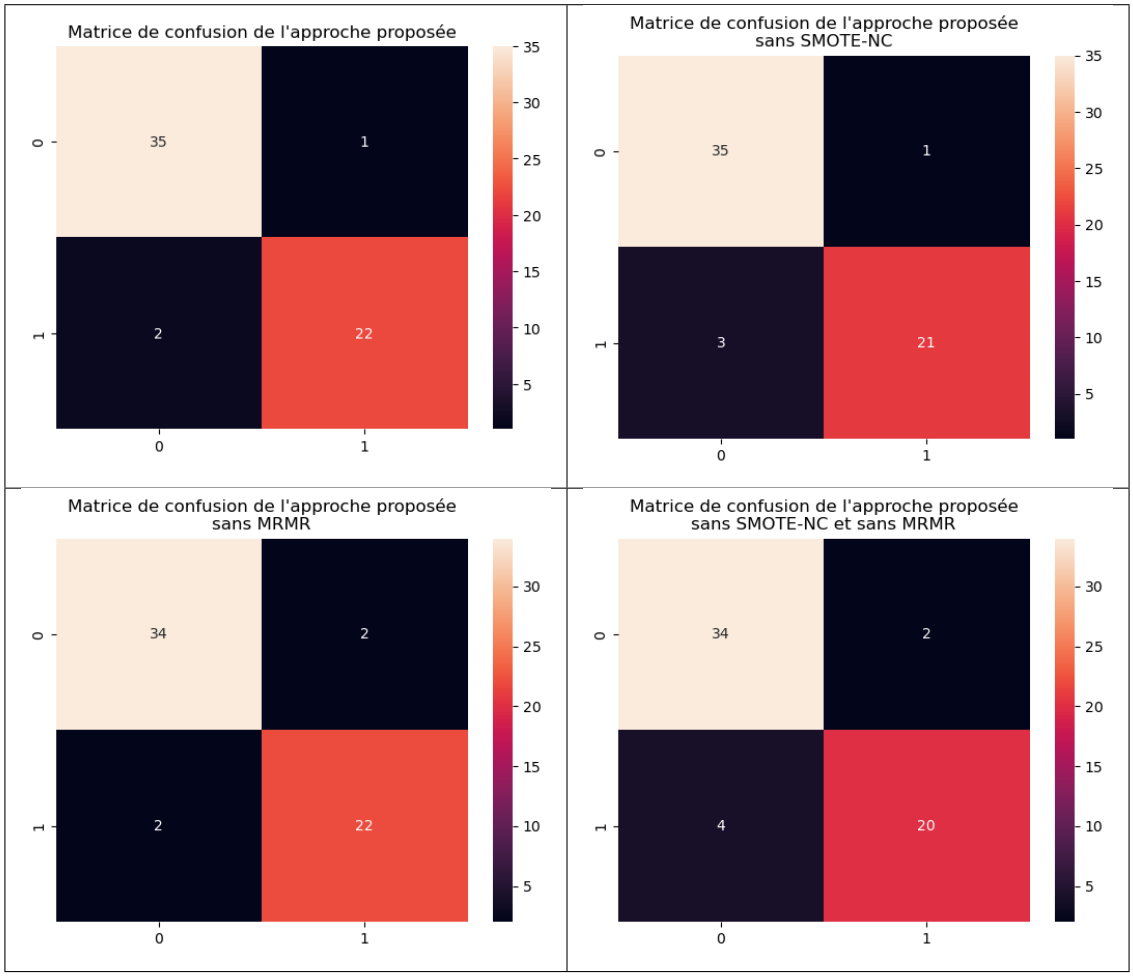


TABLE 3.6 – Matrices de confusion obtenues.

Le tableau 3.7 ci-dessous présente les résultats des rapports de classification ainsi que les AUC et les courbes ROC de chaque modèle.

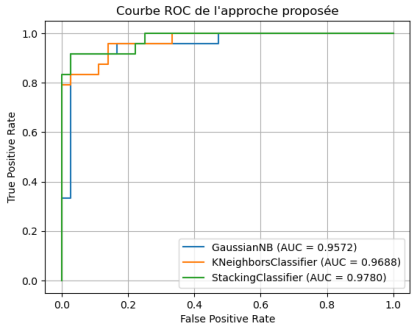
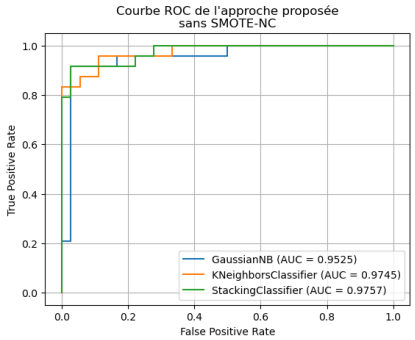
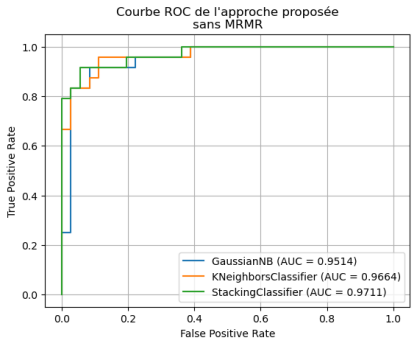
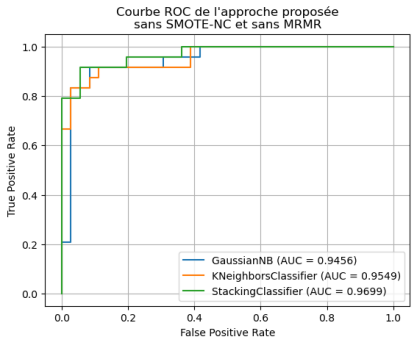
	Rapport de classification	**AUC et la courbe ROC**
Approche proposée	StackingClassifier 0.95 precision recall f1-score support 0 0.95 0.97 0.96 36 1 0.96 0.92 0.94 24 accuracy 0.95 60 macro avg 0.95 0.94 0.95 60 weighted avg 0.95 0.95 0.95 60	
Approche proposée sans SMOTE-NC	StackingClassifier 0.9333333333333333 precision recall f1-score support 0 0.92 0.97 0.95 36 1 0.95 0.88 0.91 24 accuracy 0.93 60 macro avg 0.94 0.92 0.93 60 weighted avg 0.93 0.93 0.93 60	
Approche proposée sans MRMR	StackingClassifier 0.9333333333333333 precision recall f1-score support 0 0.94 0.94 0.94 36 1 0.92 0.92 0.92 24 accuracy 0.93 60 macro avg 0.93 0.93 0.93 60 weighted avg 0.93 0.93 0.93 60	
Approche proposée sans SMOTE-NC et sans MRMR	StackingClassifier 0.9 precision recall f1-score support 0 0.89 0.94 0.92 36 1 0.91 0.83 0.87 24 accuracy 0.90 60 macro avg 0.90 0.89 0.89 60 weighted avg 0.90 0.90 0.90 60	

TABLE 3.7 – Résultats obtenus

Ci-après 3.19, un graphique à barre permettant de visualiser les résultats des métriques d'évaluation des quatre (04) modèles présentés dans le tableau 3.7. Il contient des indications sur la précision, le rappel et le f1-score qui correspondent à la classe 1 (patient malade), et nous permet de bien évaluer la capacité du modèle à atteindre le but clinique de détecter les patients malades et de minimiser le risque d'erreur.

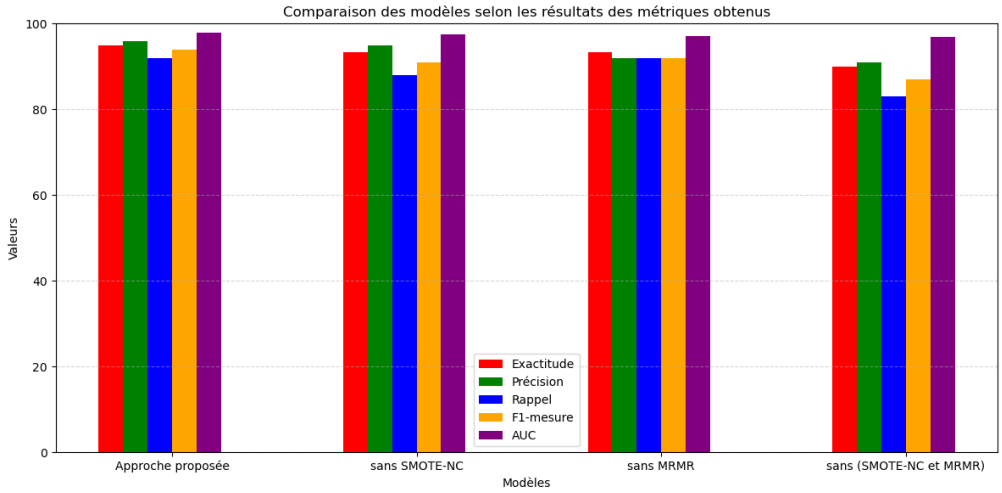


FIGURE 3.19 – Graphique à barre des résultats des métriques d'évaluation de l'approche proposée (avec SMOTE-NC et MRMR), sans SMOTE-NC, sans MRMR, et enfin sans SMOTE-NC ni MRMR.

3.5.1.1 Discussion

En analysant les résultats obtenus, nous observons que le modèle proposé (*SMOTE-NC*, *MRMR* et *stacking* (*KNN* et *NB* en utilisant *LR*)) présente les meilleures performances parmi les trois (03) autres modèles testés sans la gestion de déséquilibre des classes ainsi que sans la sélection des caractéristiques les plus importantes.

En premier lieu, cette approche nous a aidé à avoir des classes équitablement réparties en utilisant *SMOTE-NC*, ce qui fait de l'exactitude une mesure pertinente et représentative des performances globales [89]. Ainsi, le modèle a augmenté de **90%** à **95%** d'exactitude, cela a un poids significatif pour un ensemble de données médicales du fait qu'une petite amélioration peut avoir des conséquences réelles sur la santé des patients.

En second lieu, *MRMR* a joué un rôle vraiment utile en éliminant le bruit présent dans l'ensemble de données initial grâce à la sélection des caractéristiques essentielles.

Vu qu'en médecine les données sont souvent complexes et sensibles, un rappel élevé est essentiel pour les modèles de diagnostic. Il vise à minimiser les cas des patients malades non détectés (faux négatifs). Simplement dit, le modèle proposé a la capacité de détecter **92%** des malades contrairement

au dernier modèle présenté dans le tableau 3.7 qui révèle que **83%** des personnes souffrantes de maladies cardiovasculaires. C'est-à-dire, une amélioration de **9%** a été observée.

Nous observons également que le modèle proposé distingue efficacement les différentes classes atteignant une AUC de **97,80%**.

En conséquence, l'utilisation de ces techniques (*SMOTE-NC* et *MRMR*) ensemble donne un effet combiné plus performant, et nous a aidé à atteindre des résultats plus cohérents que ceux obtenus en utilisant chacune individuellement, ce qui en fait un bon compromis dans le cadre de cette étude.

3.6 Étude comparative avec des méthodes de l'état de l'art

Parmi les travaux présentés dans l'état de l'art utilisant l'ensemble de données *"Cleveland"*, il existe ceux qui se sont basés soit sur sa version originale contenant 74 caractéristiques [50], soit sur une classification multi-classes [68, 79, 109], ou sur une classification binaire des maladies cardiovasculaires.

Dans cette section, nous avons comparé notre approche qu'à celles appuyées sur cette dernière.

Les résultats (exactitudes) ainsi que la technique utilisée sont résumés dans le tableau 3.8.

	Papier	**Technique**	**Exactitude**
ML	Mistura Muibideen et Rajesh Prasad, 2020 [124].	Réseaux Bayésiens.	85%
	J.Nageswara Rao et al., 2021 [62].	Random Forest (RF).	90,16%
	Nadikatla Chandra sekhar et Samineni Peddakri shna, 2023 [34].	SVE avec (RF, KNN, LR, NB, GB et AB).	93,44% pour "Cleveland".
	Ahmed Sami Jaddoa, 2023 [94].	DT avec SMOTE.	91,4%
	Dhadkan Shrestha, 2024 [150].	XGBoost.	90%
DL	Liaqat Ali et al., 2019 [64].	χ^2 + 2-DNN.	93,33%
	P. Ramprakash et al., 2020 [70].	2-DNN + chi carré.	94%
	Snehal B. Gavande et Prof. Pramila M. Chawan, 2022 [85].	DNN.	95%
	Raniya Rone Sarra et al., 2023 [74].	ANN amélioré.	93,44%
	Abdulwahab Ali Almazroi et al., 2023 [47].	DNN.	82,5% pour "Cleveland".
	Approche proposée.	**Stacking (NB et KNN, avec LR) + SMOTE-NC + MRMR**.	**95%**

TABLE 3.8 – Comparaison de l'approche proposée avec d'autres méthodes de la littérature.

D'après le tableau 3.8, notre approche représente un résultat significatif en termes d'exactitude, comparée aux autres techniques proposées dans les travaux précédents, et ce que ce soit par rapport à celles utilisant le *ML* ou le *DL*. Nous observons également qu'elle atteint une exactitude similaire à celle de **Snehal B. Gavande et Prof. Pramila M. Chawan** [85] (**95%**). De ce fait, nous avons utilisé d'autres critères complémentaires pour approfondir la comparaison entre ces deux (02) techniques comme le montre le tableau 3.9.

Critère	**Snehal B. Gavande et Prof. Pramila M. Chawan [85]**	**Approche proposée**
Précision.	Classe 0 : 0,94 Classe 1 : 0,96	Classe 0 : **0,95** Classe 1 : 0,96
Rappel.	Classe 0 : 0,97 Classe 1 : 0,92	Classe 0 : 0,97 Classe 1 : 0,92
F1-mesure.	Classe 0 : 0,96 Classe 1 : 0,94	Classe 0 : 0,96 Classe 1 : 0,94
Sélection des caractéristiques.	Oui	Oui
Gestion du déséquilibre des classes.	Non mentionnée.	**Oui**
Validation croisée	Non	**Oui**

TABLE 3.9 – Comparaison de l'approche proposée avec la méthode de Snehal B. Gavande et Prof. Pramila M. Chawan, 2022 [85].

Suite à l'analyse de ces résultats, nous constatons que notre modèle est plus performant en termes de précision sur la classe "0" en plus de la validation croisée et la gestion du déséquilibre des classes qui ont été prises en considération.

3.7 Conclusion

Dans ce chapitre, nous avons commencé par aborder notre problématique. Puis, nous avons présenté l'idée générale de l'approche proposée basée sur la gestion du déséquilibre des classes, la sélection des caractéristiques importantes, et la méta-classification. Ceci après avoir bien préparé les données pour l'apprentissage.

Nous avons également mis en évidence les outils, le langage, ainsi que les bibliothèques utilisés lors du développement du modèle proposé.

En sachant que notre étude était dans le cadre de la prédiction des maladies cardiovasculaires afin d'arriver à des classifications plus précises qu'avant, nous avons choisi d'utiliser l'ensemble de données *"Cleveland"*.

Comme dernier point, nous avons évalué et comparé les résultats obtenus avec et sans l'approche proposée, ainsi qu'à d'autres approches analysées dans l'état de l'art afin de valider la nouvelle idée.

Conclusion générale

Au cours des dernières années, les maladies cardiovasculaires ont effrayé le monde par leur éruption rapide et le danger qu'elles provoquent. De plus, les médecins ont eu des difficultés à détecter tôt cette pathologie, surtout à cause de la quantité et de la qualité des facteurs à suivre, ce qui déclenche l'inquiétude de beaucoup de gens sur leur santé et leur vie en général.

Face à tous ces problèmes, l'évolution technologique en informatique a confirmé son importance, en facilitant cette tâche complexe utilisant l'IA et le ML en particulier. Ceci donne la possibilité de concevoir des modèles prédictifs précieux en termes d'aide à la décision médicale.

L'objectif principal de notre travail est de mettre en œuvre une nouvelle solution informatique plus efficace afin de contribuer à minimiser les pertes, et d'atténuer les différents obstacles rencontrés.

Comme point de départ de notre recherche, nous avons introduit les généralités essentielles de notre thème sur les maladies cardiovasculaires et l'IA. Ces informations nous ont aidé à bien capter les bases scientifiques de ce système circulatoire, en particulier les raisons les plus influentes sur l'apparition de ces maladies de toutes ses formes et ses signes. Nous avons également exploré les concepts fondamentaux en *IA*, *ML* et *DL*, ainsi que les différents algorithmes basiques qui désignent le cœur de la prédiction. Cette étude avait pour but de comprendre le sujet dès le départ et d'éclairer les principes.

Ensuite, nous avons découvert assez d'idées et diverses techniques du fait que nous avons présenté quelques travaux de recherche existants sur la prévision des maladies cardiovasculaires. En synthétisant les fondements clés, nous sommes arrivées à lister les avantages et les défis des modèles précédents et d'aller vers l'axe de recherche d'une nouvelle proposition pour améliorer les performances prédictives.

En outre, nous avons proposé une approche, et nous l'avons testée sur le jeu de données *"Cleveland"*. Après avoir analysé, nettoyé, et divisé ces données, nous avons opté pour la gestion du déséquilibre des classes de l'ensemble d'apprentissage à l'aide de *SMOTE-NC*, suivi d'une normalisation des données numériques, puis d'une sélection des caractéristiques les plus significatives grâce à *MRMR*. Enfin, nous avons entraîné un modèle d'apprentissage ensembliste avancé, qui s'agit de *"StackingClassifier"* qui a combiné les prédictions de *NB* et *KNN* en utilisant *LR*.

Une fois que l'implémentation de notre méthode est terminée, nous avons évalué ses performances avec des métriques calculables à partir de la matrice de confusion, notamment l'exactitude, la précision, le rappel, et la F1-mesure en plus de l'AUC et de la courbe ROC.

A cela, s'ajoute l'évaluation de trois (03) autres expérimentations : *méthode proposée sans SMOTE-NC*, *sans MRMR* et enfin *sans SMOTE-NC ni MRMR*.

De ce fait, notre approche s'est avérée la plus performante même par rapport à de nombreuses méthodes présentées dans l'état de l'art, utilisant le même ensemble de données avec une classification binaire. Ceci, nous a permis de valider l'importance des techniques combinées ainsi que l'efficacité de notre proposition.

Plusieurs travaux de la littérature possèdent des lacunes, comme le manque de gestion de déséquilibre des classes. Bien que notre méthode ait démontré de bons résultats en répondant à quelques-uns de ces défis, elle présente également quelques *limites*, comme la restriction du jeu de données utilisé. Ceci ouvre des *perspectives* à de nouvelles recherches utilisant des ensembles de données volumineux avec d'autres techniques plus avancées. Ces techniques peuvent être en matière de gestion du déséquilibre, de sélection des caractéristiques, ou d'optimisation des hyperparamètres avec des algorithmes plus adéquats aux jeux de données massifs.

Nous recommandons également d'utiliser des méthodes permettant d'expliquer les résultats des modèles de *ML* comme *SHAP*, ce qui soutient la confiance clinique, surtout dans ce secteur à risque, afin d'intégrer un modèle plus performant dans un système que les professionnels de santé peuvent utiliser pour prendre des décisions plus précises.

Nous souhaitons aussi tester notre approche dans un autre domaine moins sensible et d'évaluer ses résultats.

Pour conclure, ce projet était une très bonne expérience durant laquelle nous avons pu mettre en pratique nos modestes connaissances acquises pendant notre formation. Nous avons également appris beaucoup de choses malgré les difficultés que nous avons rencontrées. C'était un travail que nous avons réalisé avec patience et nous espérons que cela a apporté un avantage et qu'il sera un nouveau départ pour les futures recherches en informatique médicale et en prédiction des maladies cardiovasculaires en particulier.

« Le secret d'avancer, c'est de commencer » Mark Twain.

Bibliographie

[1] Cardiopathie congénitale. URL : http://www.docteurclic.com/galerie-photos/image_3197_m.jpg, (consulté le 26 février 2025).

[2] Cerebrovascular disease is an hemorrhagic stroke stock photo. URL : https://www.istockphoto.com/photo/gm1744130850-543090829, (consulté le 20 février 2025).

[3] Circulation sanguine. URL : https://www.larousse.fr/encyclopedie/divers/circulation_sanguine/34108, (consulté le 23 février 2025).

[4] Healthy heart blog. URL : https://www.qardio.com/healthy-heart-blog/5-tips-managing-hypertension/, (consulté le 25 février 2025).

[5] Heart muscle disease. URL : https://www.vectorstock.com/royalty-free-vector/heart-muscle-disease-medicine-vector-32863792, (consulté le 1 mars 2025).

[6] Le système circulatoire et son anatomie. URL : https://www.alloprof.qc.ca/fr/eleves/bv/sciences/le-systeme-circulatoire-et-son-anatomie-s1270, (consulté le 23 février 2025).

[7] Naive bayes overview. URL : https://databasecamp.de/wp-content/uploads/naive-bayes-overview.png, (consulté le 25 mai 2025).

[8] Risques associés aux troubles du sommeil : La maladie coronarienne. URL : https://www.info-somnolence.fr/risques-associes/cardio-vasculaire/maladie-coronarienne/, (consulté le 20 février 2025).

[9] Spyder : The scientific python development environment — documentation. URL : https://docs.spyder-ide.org/3/index.html, (consulté le 25 avril 2025).

[10] Stacking in machine learning. URL : https://www.geeksforgeeks.org/machine-learning/stacking-in-machine-learning/, (consulté le 18 juin 2025).

[11] Typical structure of random forest. URL : https://www.researchgate.net/figure/Typical-structure-of-Random-Forest_fig2_344437372, (consulté le 1 mars 2025).

[12] Tout savoir sur l'insuffisance cardiaque, 2019. URL : https://kafnewspro.blogspot.com/2019/08/tout-savoir-sur-linsuffisance-cardiaque.html, (consulté 25 février 2025).

[13] Introduction à l'apprentissage automatique, 2022. URL : https://sti.eduscol.education.fr/sites/eduscol.education.fr.sti/files/ressources/pedagogiques/14512/14512-introduction-lapprentissage-automatique-ensps.pdf, (consulté le 1 mars 2025).

[14] Logistic regression in machine learning, 2023. URL : https://www.almabetter.com/bytes/tutorials/data-science/logistic-regression, (consulté le 2 mars 2025).

[15] Understanding of lstm networks, 2023. URL : https://www.geeksforgeeks.org/understanding-of-lstm-networks/, (consulté le 17 juin 2025).

[16] Underfitting, optimalfitting and overfitting, 2024. URL : https://www.reddit.com/r/Btechtards/comments/19bwfp8/underfitting_optimalfitting_and_overfitting/?rdt=54526, (consulté le 1 mars 2025).

[17] Mohammed Ahmed and Idress Husien. Heart disease prediction using hybrid machine learning : A brief review. *Journal of Robotics and Control (JRC)*, 5(3) :884–892, 2024. URL : https://journal.umy.ac.id/index.php/jrc/article/view/21606.

[18] Ibnu Akil and Indra Chaidir. Classification of heart disease diagnoses using gaussian naïve bayes. *Komputasi : Jurnal Ilmiah Ilmu Komputer dan Matematika*, 21(2) :31–36, 2024. URL : https://scholar.google.com/scholar?q=Classification+of+Heart+Disease+Diagnoses+Using+Gaussian+Naïve+Bayes.

[19] Ibraheem M Alkhawaldeh et al. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, 13(5) :373–378, 2023. URL : https://www.wjgnet.com/2222-0682/full/v13/i5/373.htm.

[20] Anaconda. Getting started with anaconda. URL : https://www.anaconda.com/docs/getting-started/getting-started, (consulté le 25 avril 2025).

[21] Guy P. Armstrong. Présentation des maladies des valvules cardiaques, 2023. URL : https://www.msdmanuals.com/fr/accueil/troubles-cardiaques-et-vasculaires/maladies-des-valvules-cardiaques/pr%C3%A9sentation-des-maladies-des-valvules-cardiaques, (consulté le 28 février 2025).

[22] Jan Carlo T. Arroyo and Allemar Jhone P. Delima. An optimized neural network using genetic algorithm for cardiovascular disease prediction. *Journal of Advances in Information Technology*, 13(1), 2022. URL : https://www.researchgate.net/publication/357792868.

[23] American Heart Association. Qu'est-ce que le syndrome métabolique ? URL : https://www.heart.org/en/health-topics/metabolic-syndrome/about-metabolic-syndrome?, (consulté le 29 février 2025).

[24] Chirurgiens Cardiaques Associés. La chirurgie valvulaire, 2019. URL : https://www.chirurgie-cardiaque-caen.fr/Valvulopathie-mitrale-et-tricuspide, (consulté le 28 février 2025).

[25] I. Azizi, K. Echihabi, and T. Palpanas. Graph-based vector search : An experimental evaluation of the state-of-the-art. *Proceedings of the ACM on Management of Data*, 3(1) :Article 43, 1–31, 2025. URL : https://doi.org/10.1145/3709693.

[26] Giovanni Ballarin. Ridge regularized estimation of var models for inference. *Journal of Time Series Analysis*, 46(3) :235–257, 2025. URL : https://doi.org/10.1111/jtsa.12737.

[27] Dr Gerard BERTHIER. Anatomie et fonctionnement de l'appareil cardiovasculaire, 2019. URL : https://anticoag-pass-s2d.fr/anatomie-et-fonctionnement-de-lappareil-cardiovasculaire/, (consulté le 23 février 2025).

[28] Lomash Bhuva. Mastering stacking in machine learning : The ultimate guide with code, 2025. URL : https://medium.com/@lomashbhuva/mastering-stacking-in-machine-learning-the-ultimate-guide-with-code-da74bf3c329e, (consulté le 20 juin 2025).

[29] Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14 :106, 2013. URL : https://www.biomedcentral.com/1471-2105/14/106.

[30] Andriy Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019. Chapitre 5, page 13. URL : https://themlbook.com/wiki/doku.php.

[31] Adobe Business. Machine learning definition, models, and applications, 2025. URL :https://business.adobe.com/blog/basics/what-is-machine-learning#how-machine-learning-works, (consulté le 24 février 2025).

[32] Jie Cai et al. Feature selection in machine learning : A new perspective. *Neurocomputing*, 300 :70–79, 2018. URL : https://www.sciencedirect.com/science/article/abs/pii/S0925231218302911.

[33] Stéphane Caron. Une introduction aux arbres de décision. *Stéphane Caron*, 31, 2011. URL : https://scaron.info/doc/intro-arbres-decision/intro.pdf.

[34] Nadikatla Chandrasekhar and Samineni Peddakrishna. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, 11(4) :1210, 2023. URL : https://www.mdpi.com/2227-9717/11/4/1210.

[35] Agence Française d'Adoption. Les malformations cardiaques. URL : https://www.diplomatie.gouv.fr/IMG/pdf/les_malformations_cardiaques_cle821567.pdf.

[36] Daniella. Convolutional neural network : operation, advantages and applications in ai, 2024. URL : https://www.innovatiana.com/post/convolutional-neural-network, (consulté le 27 février 2025).

[37] Kaushik Das. How recurrent neural network (rnn) works, 2020. URL : sur https://dataaspirant.com/how-recurrent-neural-network-rnn-works/, (consulté le 17 juin 2025).

[38] Data Science Team. Smote, 2021. URL : https://datascience.eu/fr/programmation-informatique/smote/, (consulté le 25 avril 2025).

[39] Ministère de l'Économie des Finances et de la Souveraineté industrielle et numérique. Intelligence artificielle, un levier de croissance pour votre entreprise, 2025. URL : https://www.economie.gouv.fr/entreprises/intelligence-artificielle, (consulté le 16 juin 2025).

[40] H Dereppe et al. Recommandations relatives à la prévention des maladies cardiovasculaires en pratique clinique. groupe de travail belge de prévention des maladies cardiovasculaires. *Revue médicale de Bruxelles*, 30(1) :37, 2009. URL : https://www.researchgate.net/publication/238738984.

[41] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 2004. URL : https://www.worldscientific.com/doi/abs/10.1142/S0219720005001004.

[42] Société Alzheimer du Canada. Maladies apparentées - maladie cérébrovasculaire, 2018. URL : https://alzheimer.ca/sites/default/files/documents/maladies-apparentees_maladie-cerebro-vasculaire.pdf, (consulté le 25 février 2025).

[43] Elliot. Another feature selection algorithm : Mrmr, 2021. URL : https://elliot-weissberg.medium.com/another-feature-selection-algorithm-mrmr-3827b6b19e33, (consulté le 11 Avril 2025).

[44] Feature engine developers. Mrmr - minimum redundancy maximum relevance. URL : https://feature-engine.trainindata.com/en/1.8.x/user_guide/selection/MRMR.html, (consulté le 12 Avril 2025).

[45] EPA. Les risques environnementaux pèsent lourdement sur le cœur, 2007. URL : https://www.epa.gov/sites/default/files/2015-09/documents/ehwhh_french_2007_08.pdf.

[46] A. Vaswani et al. Attention is all you need. *arXiv preprint*, arXiv :1706.03762, 2017. URL : https://arxiv.org/pdf/1706.03762.

[47] Abdulwahab Ali Almazrio et al. A clinical decision support system for heart disease prediction using deep learning. *IEEE Access*, 11 :61646–61659, 2023. URL : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10148957.

[48] Abolfazl Younesi et al. A comprehensive survey of convolutions in deep learning : Applications, challenges, and future trends. *IEEE Access*, 12 :28638–28666, 2024. URL : https://doi.org/10.1109/ACCESS.2024.3376441.

[49] Amin Ul Haq et al. Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–4, 2019. URL : https://ieeexplore.ieee.org/abstract/document/9033683.

[50] Anna Karen Gárate-Escamila et al. Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, 19 :100330, 2020. URL : https://doi.org/10.1016/j.imu.2020.100330.

[51] Atharv Nikam et al. Cardiovascular disease prediction using machine learning models. In *2020 IEEE Pune section international conference (PuneCon)*, pages 22–27, 2020. URL : https://ieeexplore.ieee.org/document/9362367.

[52] Bruno Baudin et al. Données épidémiologiques des maladies cardiovasculaires et prise en charge des accidents cardiovasculaires. *Revue Francophone des Laboratoires*, 409, 2009. URL : https://doi.org/10.1016/S1773-035X(09)70198-4.

[53] Chustecki Margaret et al. Benefits and risks of ai in health care : Narrative review. *Interactive Journal of Medical Research*, 13(1) :e53616, 2024. URL : https://i-jmr.org/2024/1/e53616.

[54] D. Boldini et al. Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, 15 :73, 2023. URL : https://doi.org/10.1186/s13321-023-00743-7.

[55] D. Yu et al. Dynamic coverage control based on k-means. *IEEE Transactions on Industrial Electronics*, 69(5) :5333–5341, 2021. URL : https://doi.org/10.1109/TIE.2021.3080205.

[56] H. Zhang et al. β-dqn : Improving deep q-learning by evolving the behavior. *arXiv preprint*, arXiv :2501.00913, 2025. URL : https://arxiv.org/abs/2501.00913.

[57] Harshit Jindal et al. Heart disease prediction using machine learning algorithms. In *IOP conference series : materials science and engineering*, volume 1022, page 012072, 2021. URL : https://doi.org/10.1088/1757-899X/1022/1/012072.

[58] He Haibo et al. Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328, 2008. URL : https://ieeexplore.ieee.org/abstract/document/4633969.

[59] J. Guo et al. Incremental self-training for semi-supervised learning. *arXiv preprint*, abs/2404.12398, 2024. URL : https://arxiv.org/html/2404.12398v1.

[60] Jimin Liu et al. Predictive classifier for cardiovascular disease based on stacking model fusion. *Processes*, 10(4) :749, 2022. URL : https://www.mdpi.com/2227-9717/10/4/749.

[61] Jirong Zhang et al. Artificial intelligence applied in cardiovascular disease : a bibliometric and visual analysis. *Frontiers in cardiovascular medicine*, 11 :1323918, 2024. URI : https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2024.1323918/full.

[62] J.Nageswara Rao et al. Cardiovascular disease prediction using machine learning techniques. *Turkish Journal of Physiotherapy and Rehabilitation*, 32 :6875–6880. URL : https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&q=Cardiovascular+Disease+Prediction+Using+Machine+Learning+Techniques+J.Nageswara+Rao&btnG=.

[63] L. Grillotti et al. Quality-diversity actor-critic : Learning high-performing and diverse behaviors via value and successor features critics. *arXiv preprint*, arXiv :2403.09930, 2024. URL : https://arxiv.org/abs/2403.09930.

[64] Liaqat Ali et al. An automated diagnostic system for heart disease prediction based on x2 statistical model and optimally configured deep neural network. *Ieee Access*, 7 :34938–34945, 2019. URL : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8666632.

[65] Loveleen Kumar et al. Deep learning based healthcare method for effective heart disease prediction. *EAI Endorsed Transactions on Pervasive Health and Technology*, 9 :1–6, 2023. URL : https://www.researchgate.net/publication/375148796_Deep_Learning_Based_Healthcare_Method_for_Effective_Heart_Disease_Prediction.

[66] M. Hahsler et al. dbscan : Fast density-based clustering with r. *Journal of Statistical Software*, 91(1) :1–30, 2019. URL : https://doi.org/10.18637/jss.v091.i01.

[67] María Teresa García-Ordás et al. Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimedia Tools and Applications*, 82(20) :31759–31773, 2023. URL : https://link.springer.com/article/10.1007/s11042-023-14817-z.

[68] Mohd Ashraf et al. Improved heart disease prediction using deep neural network. *Asian Journal of Computer Science and Technology*, 8(2) :49–54, 2024. URL : https://www.researchgate.net/publication/353926742_Improved_Heart_Disease_Prediction_Using_Deep_Neural_Network.

[69] Nitesh V. Chawla et al. SMOTE : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16 :321–357, 2002. URL : https://www.jair.org/index.php/jair/article/view/10302/24590.

[70] P. Ramprakash et al. Heart disease prediction using deep neural network. In *2020 international conference on inventive computation technologies (ICICT)*, pages 666–670, 2020. URL : https://ieeexplore.ieee.org/document/9112443.

[71] Pablo Bermejo et al. Incremental wrapper-based subset selection with replacement : An advantageous alternative to sequential forward selection. In *2009 IEEE symposium on computational intelligence and data mining*, pages 367–374, 2009. URL : https://ieeexplore.ieee.org/abstract/document/4938673.

[72] Q. Liu et al. Enhancing ppo with trajectory-aware hybrid policies. *arXiv preprint*, arXiv :2502.15968, 2025. URL : https://arxiv.org/abs/2502.15968.

[73] Qandeel Asghar et al. Optimized deep learning framework for early detection of heart disease. *Journal of Computing & Biomedical Informatics*, 8(01), 2024. URL : https://jcbi.org/index.php/Main/article/view/737.

[74] Raniya Rone Sarra et al. Enhanced accuracy for heart disease prediction using artificial neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(1) :375–383, 2023. URL : https://www.researchgate.net/publication/364949647.

[75] Roweida Mohammed et al. Machine learning with oversampling and undersampling techniques : Overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248, 2020. URL : https://www.researchgate.net/publication/340978368.

[76] S. Wang et al. Diabetes risk analysis based on machine learning lasso regression model. *Journal of Theory and Practice of Engineering Science*, 4(1) :58–64, 2024. URL : https://centuryscipub.com/index.php/jtpes/article/view/429.

[77] Saba Bashir et al. Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*, pages 619–623, 2019. URL : https://ieeexplore.ieee.org/document/8667106.

[78] Segun Akinola et al. Enhancing cardiovascular disease prediction : A hybrid machine learning approach integrating oversampling and adaptive boosting techniques. *AIMS Medical Science AIMS Medical Science*, 11(2) :58–71, 2024. URL : https://doi.org/10.3934/medsci.2024005.

[79] Senthilkumar Mohan et al. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7 :81542–81554, 2019. URL : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8740989.

[80] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006. URL : https://www.researchgate.net/publication/222511520.

[81] FFC. Le fonctionnement du cœur, 2025. URL : https://www.fedecardio.org/je-m-informe/le-fonctionnement-du-coeur/, (consulté le 24 février 2025).

[82] FFC. Le système cardiovasculaire, 2025. URL : https://www.fedecardio.org/je-m-informe/le-systeme-cardiovasculaire/, (consulté le 24 février 2025).

[83] Vascular Disease Foundation. Lifesaving tips on pad - flyer, 2012. URL : https://vasculardisease.org/flyers/lifesaving-tips-on-pad-flyer.pdf.

[84] Freepik. Enfermedad arterial periférica. URL : https://www.freepik.es/vector-premium/enfermedad-arterial-periferica_29601591.htm, (consulté le 24 février 2025).

[85] S Gavande and P Chawan. Prediction of heart disease using neural network. *Int Res J Eng Technol (IRJET)*, 9, 2022. URL : https://www.irjet.net/archives/V9/i9/IRJET-V9I9171.pdf.

[86] Julie Giorgetta. Capillaire sanguin : définition, rôle, schéma, fragilité, 2021. URL : https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2686177-capillaire-sanguin, (consulté le 16 juin 2025).

[87] Google. Questions fréquentes. URL : https://research.google.com/colaboratory/faq.html?hl=fr, (consulté le 22 avril 2025).

[88] Alisa J. Hamilton et al. Machine learning and artificial intelligence : applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology*, 1(1) :e28, 2021. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9495400/.

[89] Haibo He and Eduardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9) :1263–1284, 2009. URL : https://ieeexplore.ieee.org/document/5128907.

[90] Apollo Hospitals. Heart disease : Types, symptoms, and treatment. URL : https://www.apollohospitals.com/fr/health-library/heart-disease-types-symptoms-and-treatment/, (consulté le 1 mars 2025).

[91] IBM. What is machine learning ?, 2021. URL : https://www.ibm.com/think/topics/machine-learning?, (consulté le 24 février 2025).

[92] IBM. What is deep learning ?, 2024. URL : https://www.ibm.com/think/topics/deep-learning?, (consulté le 27 février 2025).

[93] Imbalanced-learn. imbalanced-learn documentation, 2024. URL : https://imbalanced-learn.org/stable/, (consulté le 28 avril 2025).

[94] Ahmed Sami Jaddoa. Heart disease prediction system using (smote technique) balanced dataset and decision tree classifier. In *AIP Conference Proceedings*, volume 2834, 2023. URL : https://doi.org/10.1063/5.0161558.

[95] S Jarvis and S Saman. Cardiac system 1 : anatomy and physiology. *Nursing Times*, 114(2) :34–37, 2018. URL : https://emap-moon-prod.s3.amazonaws.com/wp-content/uploads/sites/3/2018/01/310118-Cardiac-system-1-anatomy-and-physiology.pdf.

[96] Dr. Khacha. Insuffisance cardiaque congestive de l'adulte. URL : https://medecine.univ-batna2.dz/sites/default/files/medecine/files/ic_congestive_dr_khacha_1.pdf.

[97] M. J. Khan, S. H. Ahmed, and G. Sukthankar. Smart sampling : Self-attention and bootstrapping for improved ensembled q-learning. *arXiv preprint*, arXiv :2405.08252, 2024. URL : https://arxiv.org/abs/2405.08252.

[98] Richard E. Korf. Linear-space best-first search. *Artificial intelligence*, 62(1) :41–78, 1993. URL : https://doi.org/10.1016/0004-3702(93)90045-D.

[99] Niklas Lang. Long short-term memory networks (lstm) – simply explained !, 2022. URL : https://databasecamp.de/en/ml/lstms, (consulté le 18 juin 2025).

[100] Stahle Lars and Svante Wold. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4) :259–272, 1989. URL : https://doi.org/10.1016/0169-7439(89)80095-4.

[101] François Lehn. Les 16 symptômes de l'apnée du sommeil et comment les reconnaître, 2024. URL : https://www.pressesante.com/les-16-symptomes-de-lapnee-du-sommeil-et-comment-les-reconnaitre/, (consulté le 29 février 2025).

[102] François Lehn. Les signes et symptômes des principales maladies cardiovasculaires, 2024. URL : https://www.pressesante.com/les-signes-et-symptomes-des-principales-maladies-cardiovasculaires/, (consulté le 1 mars 2025).

[103] Bo Li and Benjamin Haibe-Kains. pymrmre : Parallelized minimum redundancy, maximum relevance (mrmr) ensemble feature selections, 2021. URL : https://pypi.org/project/pymrmre/, (consulté le 22 avril 2025).

[104] Y.-F. Li and Z.-H. Zhou. Improving semi-supervised support vector machines through unlabeled instances selection. *arXiv preprint*, arXiv :1005.1545, 2010. URL : https://arxiv.org/abs/1005.1545.

[105] Wei-Chao Lin et al. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409 :17–26, 2017. URL : https://www.sciencedirect.com/science/article/abs/pii/S0020025517307235.

[106] G. C. Linderman and S. Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2) :313–332, 2019. URL : https://doi.org/10.1137/18M1216134.

[107] LinkedIn Community. Quelles sont les techniques de sélection de caractéristiques les plus efficaces pour la classification ?, 2023. URL : https://fr.linkedin.com/advice/1/what-most-effective-feature-selection-techniques-classification-0kmbc, (consulté le 12 Avril 2025).

[108] Jean-Sébastien Louis. Schéma simplifié de la circulation sanguine systémique et pulmonaire. URL : https://www.researchgate.net/figure/fig6_344463950, (consulté le 24 février 2025).

[109] G. Magesh and P. Swarnalatha. Optimal feature selection through a cluster-based dt learning (cdtl) in heart disease prediction. *Evolutionary intelligence*, 14(2) :583–593, 2020. URL : https://link.springer.com/article/10.1007/s12065-019-00336-0.

[110] Santé Maghreb. Les maladies cardiovasculaires, première cause de mortalité en algérie, 2021. URL : https://www.santemaghreb.com/sites_pays/actus.asp?id=29471&rep=algerie, (consulté le 22 avril 2025).

[111] C. Malzer and M. Baum. Constraint-based hierarchical cluster selection in automotive radar data. *Sensors*, 21(10) :3410, 2021. URL : https://doi.org/10.3390/s21103410.

[112] MarketMuse. What is a multilayer perceptron (mlp). URL : https://blog.marketmuse.com/glossary/multilayer-percpetron-definition/, (consulté le 17 juin 2025).

[113] Matplotlib. Matplotlib — visualization with python. URL : https://matplotlib.org/, (consulté le 05 avril 2025).

[114] Matplotlib. Pyplot tutorial. URL : https://matplotlib.org/stable/tutorials/pyplot.html, (consulté le 05 avril 2025).

[115] Medtronic. Fibrillation atriale. URL : https://www.medtronic.com/fr-fr/patients/pathologies/fibrillation-atriale.html, (consulté le 29 février 2025).

[116] Rick Merritt. What is a transformer model ?, 2022. URL : https://blogs.nvidia.com/blog/what-is-a-transformer-model/, (consulté le 18 juin 2025).

[117] Pr : MESGHOUNI.K. Maladie thromboembolique veineuse, 2024. URL : https://facmed.univ-constantine3.dz/wp-content/uploads/2023/11/MALADIE-THROMBO.pdf.

[118] Satwik Mishra. Handling imbalanced data : Smote vs. random undersampling. *Int. Res. J. Eng. Technol*, 4(8) :317–320, 2017. URL : https://scholar.google.com/scholar?q=Handling+imbalanced+data%3A+SMOTE+vs.+random+undersampling.

[119] L. Brent Mitchell. Présentation des troubles du rythme cardiaque, 2024. URL : https://www.msdmanuals.com/fr/accueil/troubles-cardiaques-et-vasculaires/troubles-du-rythme-cardiaque/pr%C3%A9sentation-des-troubles-du-rythme-cardiaque, (consulté le 25 février 2025).

[120] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, New York, USA, 1 edition, 1997.

[121] TOUAH Souheyla MOKHTAR Hanane, REBIHI Fatima Zohra. Etude de facteurs de risque de la maladie cardiovasculaire dans la population de tiaret. Master's thesis, Université de Tiaret, 2019. URL : http://dspace.univ-tiaret.dz/bitstream/123456789/4634/1/TH.M.SNV.FR.2019.170.pdf.

[122] Dr Pascal Motreff. Facteurs de risques cardiovasculaires, 2005. URL : https://www.afdn.org/sites/www.afdn.org/files/medias/documents/id-facteurs-risques-cardiovasculaire.pdf.

[123] Muhila. Support vector machine (svm) algorithm – machine learning everything you need to know, 2021. URL : https://www.acte.in/support-vector-machine-algorithm-machine-learning-article, (consulté le 1 mars 2025).

[124] Mistura Muibideen and Rajesh Prasad. A fast algorithm for heart disease prediction using bayesian network model, 2020. URL : https://arxiv.org/pdf/2012.09429.

[125] National Heart, Lung, and Blood Institute. What is coronary heart disease ? URL : https://www.nhlbi.nih.gov/health/coronary-heart-disease, (consulté le 27 mai 2025).

[126] Santé Chez Nous. Maladie artérielle périphérique. URL : https://santecheznous.com/condition/getcondition/maladie-arterielle-peripherique, (consulté le 24 février 2025).

[127] NumPy. Numpy : the absolute basics for beginners. URL : https://numpy.org/doc/2.2/user/absolute_beginners.html, (consulté le 05 avril 2025).

[128] Codes of Interest. Difference between artificial intelligence, machine learning, and deep learning, 2016. URL : https://www.codesofinterest.com/2016/11/difference-artificial-intelligence-machine-learning-deep-learning.html, (consulté le 24 février 2025).

[129] OMS. Maladies cardiovasculaires, 2021. URL : https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29, (Consulté le 23 février 2025).

[130] OMS. Ethics and governance of artificial intelligence for health : Guidance on large multi-modal models, 2025. URL : https://www.who.int/publications/i/item/9789240084759, (consulté le 23 mai 2025).

[131] OMS. Hypertension - fiche d'information, 2025. URL : https://www.who.int/fr/news-room/fact-sheets/detail/hypertension, (consulté le 20 février 2025).

[132] Pandas. pandas documentation, 2025. URL : https://pandas.pydata.org/docs/, (consulté le 05 avril 2025).

[133] Michael Steinbach Pang-Ning Tan and Vipin Kumar. *Introduction to Data Mining*. Pearson New International Edition, first edition, 2014. isbn : 978-1-292-02615-2.

[134] Joel Paul. A comprehensive study of advanced machine learning algorithms for predicting heart disease using the cleveland dataset. 2024. URL : https://www.researchgate.net/publication/385920159.

[135] Bohdan Pavlyshenko. Using stacking approaches for machine learning models. *IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 255–258, 2018. URL : https://ieeexplore.ieee.org/abstract/document/8478522.

[136] Van Hiep Phung and Eun Joo Rhee. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9(21) :4500, 2019. URL : https://doi.org/10.3390/app9214500, (16 juin 2025).

[137] Swathi Pothuganti. Review on over-fitting and under-fitting problems in machine learning and solutions. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng*, 7(9) :3692–3695, 2018. URL : https://doi.org/10.1109/ACCESS.2024.3376441.

[138] Agir pour le Cœur des Femmes. La maladie veineuse thrombo-embolique : diagnostic et prise en charge. URL : https://www.agirpourlecoeurdesfemmes.com/anticiper/media/La-maladie-veineuse-thrombo-embolique-diagnostic-et-prise-en-charge, (consulté le 1 mars 2025).

[139] Python. What is python ? executive summary, 2025. URL : https://www.python.org/doc/essays/blurb/, (consulté le 25 avril 2025).

[140] Raimath. Arythmie cardiaque : quels sont les aliments à éviter ?, 2022. URL : https://www.informationhospitaliere.com/arythmie-cardiaque-quels-sont-les-aliments-a-eviter, (consulté le 25 février 2025).

[141] Raphaël Richard. Sélection de caractéristiques, 2019. URL : https://24pm.com/117-definitions/358-selection-de-caracteristiques, (consulté le 12 avril 2025).

[142] Ph.D Robert Detrano, M.D. Cleveland clinic heart disease dataset. URL : https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset, (consulté le 25 mars 2025).

[143] Narges Roustaei. Application and interpretation of linear-regression analysis. *Medical Hypothesis Discovery & Innovation in Ophthalmology*, 13(3) :151–159, 2024. URL : https://doi.org/10.51329/mehdiophthal1506.

[144] Guillaume Saint-Cirgue. *Apprendre le machine learning en une semaine*. éditeur, 2024.

[145] Madison Schott. K-nearest neighbors (knn) algorithm for machine learning, 2019. URL : https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26, (consulté le 1 mars 2025.

[146] Scikit-learn. Getting started. URL : https://scikit-learn.org/stable/getting_started.html, (consulté le 05 avril 2025).

[147] Scikit-learn developers. *Naive Bayes*. URL : https://scikit-learn.org/stable/modules/naive_bayes.html, (onsulté le 25 mai 2025).

[148] Seaborn. An introduction to seaborn. URL : https://seaborn.pydata.org/tutorial/introduction.html, (consulté le 22 avril 2025).

[149] Sumit Sharma and Mahesh Parmar. Heart diseases prediction using deep learning neural network model. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(3) :2244–2248, 2020. URL : https://www.ijitee.org/portfolio-item/c9009019320/.

[150] Dhadkan Shrestha. Advanced machine learning techniques for predicting heart disease : A comparative analysis using the cleveland heart disease dataset. *Applied Medical Informatics*, 46(3), 2024. URL : https://www.researchgate.net/publication/385938995.

[151] Coursera Staff. What is a recurrent neural network ?, 2025. URL : https://www.coursera.org/articles/what-is-a-recurrent-neural-network, (consulté le 17 juin 2025).

[152] StudyRaid. Fondamentaux du machine learning normalisation et standardisation. URL : https://app.studyraid.com/fr/read/2715/55444/normalisation-et-standardisation, (consulté le 6 avril 2025).

[153] Tisha Suboc. Revue générale des cardiomyopathies, 2024. URL : https://www.msdmanuals.com/fr/professional/troubles-cardiovasculaires/cardiomyopathies/revue-g%C3%A9n%C3%A9rale-des-cardiomyopathies, (consulté le 24 février 2025).

[154] S Suriya, NH Madhumitha, and PG Scholar. Heart failure prediction using gaussian naïve bayes algorithm. *Journal of Information Technology and Digital World*, 5(2) :125–143,

2023. URL : https://scholar.google.com/scholar?q=Heart+Failure+Prediction+using+Gaussian+Naïve+Bayes+Algorithm.

[155] M Darshan Teja and G Mokesh Rayalu. Optimizing heart disease diagnosis with advanced machine learning models : a comparison of predictive performance. *BMC Cardiovascular Disorders*, 25(1) :212, 2025. URL : https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-025-04627-6.

[156] UNESCO. Artificial intelligence : between myth and reality, 2018. URL : https://courier.unesco.org/en/articles/artificial-intelligence-between-myth-and-reality, (consulté le 16 juin 2025).

[157] Dr Anne-Christine Della Valle. Cœur : anatomie, rôle, opération, 2023. URL : https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2526844-coeur-anatomie-role-operation-maladies-cardiovasculaires/, (consulté le 23 février 2025).