République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Echahid Cheikh Larbi Tebessi- Tebessa

Faculté des sciences exactes, sciences de la nature et de la vie
Département d'informatique / de mathématique
Laboratoire de domiciliation : laboratoire des mathématiques, informatique et systèmes (LAMIS)

THÈSE

Présentée et Soutenue Publiquement en Vue de l'Obtention du Diplôme de Doctorat en Troisième Cycle

Par :
Boughougal Safa

Domaine : mathématiques et Informatique **Filière :** informatique
Spécialité : Système d'information

Intitulée

Approche Intelligente pour la prédiction de la maladie rénale chez les diabétiques

Soutenue le : 22 Janvier 2026 Devant le Jury composé de :

Nom et Prénom	Grade		
Mr Gattal Abdeljalil	Pr	Univ. De Larbi Tébessi	Président
Mr Laouar Mohamed Ridda	Pr	Univ. De Larbi Tébessi	Encadrant
Mr Siam Abderrahim	Pr	Univ. De Abbes Laghrour	Co-encadrant
Mr Khediri Abderrazak	MCA	Univ. De De Larbi Tébessi	Examinateur
Mr Maarouk Toufik Messaoud	Pr	Univ. De Abbes Laghrour	Examinateur

Année Universitaire : 2024/2025

Résumé

La maladie rénale chronique (MRC) est l'une des affections silencieuses les plus redoutables, progressant souvent sans symptômes jusqu'à un stade irréversible. Cette discrétion clinique, conjuguée à une prise en charge tardive, en fait un véritable défi de santé publique. Dans ce contexte, l'intelligence artificielle (IA) offre des opportunités inédites pour anticiper son apparition, en exploitant des données médicales hétérogènes et parfois incomplètes.

Cette thèse propose une approche intelligente et pragmatique pour la prédiction précoce de la MRC, s'appuyant sur des techniques avancées d'apprentissage automatique. Elle vise à concevoir des modèles de classification robustes, capables de détecter efficacement les cas à risque, malgré les contraintes courantes des données médicales : petit volume, déséquilibre des classes, hétérogénéité des variables (numériques/catégorielles), et valeurs manquantes.

Notre travail s'articule autour de quatre contributions complémentaires :

- *Une analyse comparative approfondie de méthodes de sélection de caractéristiques couplées à divers algorithmes d'apprentissage supervisé, dans le but d'identifier les combinaisons les plus performantes pour prédire la MRC.*
- *Une validation expérimentale rigoureuse sur deux ensembles de données cliniques indépendants, permettant d'évaluer la stabilité et la capacité de généralisation des modèles proposés.*
- *Le développement d'une stratégie intégrée, centrée sur l'amélioration de la qualité des données, leur transformation clinique pertinente, et l'enrichissement par des variables à forte valeur ajoutée, comme le taux de filtration glomérulaire estimé (eGFR), calculé et intégré afin d'affiner la précision du diagnostic automatisé.*
- *Enfin, l'enrichissement du jeu de données par des techniques d'augmentation compatibles avec les variables médicales, couplé à une sélection récursive des variables les plus discriminantes, pour renforcer la représentativité des données et la fiabilité des modèles générés.*

Les résultats obtenus démontrent des performances prédictives significatives, consolidées par une cohérence clinique accrue. Ce travail ouvre la voie à la mise en place d'un système intelligent d'aide à la décision médicale, particulièrement pertinent pour le dépistage anticipé de la MRC.

Mots-clés : *Maladie rénale chronique, prédiction précoce, intelligence artificielle, apprentissage automatique, données cliniques, eGFR, sélection de caractéristiques, augmentation de données, équilibrage des classes, modèles supervisés, enrichissement des données.*

Abstarct

Chronic Kidney Disease (CKD) is one of the most serious silent conditions, often progressing undetected until reaching an irreversible stage. Its asymptomatic nature, combined with delayed diagnosis, makes it a major public health concern. In this context, Artificial Intelligence (AI) offers unprecedented opportunities for early detection by leveraging heterogeneous and often incomplete medical data.

This thesis proposes a smart and pragmatic approach for the early prediction of CKD, based on advanced machine learning techniques. The aim is to design robust classification models capable of effectively identifying at-risk cases, despite common constraints in clinical datasets such as small size, class imbalance, variable heterogeneity (numerical/categorical), and missing values.

Our work is structured around four complementary contributions :

- *A comparative analysis of several feature selection techniques combined with supervised learning models to identify the best-performing combinations for CKD prediction.*
- *A rigorous validation using two independent clinical datasets, to evaluate the robustness and generalization capacity of the proposed models.*
- *The development of an integrated strategy, focused on improving data quality, transforming features to reflect clinical relevance, and enriching the dataset with high-value medical indicators such as the estimated Glomerular Filtration Rate (eGFR), which is calculated and added to refine automated diagnosis.*
- *Finally, the expansion of the dataset using augmentation techniques suitable for medical variables, combined with recursive feature selection to enhance the dataset's representativeness and the reliability of the generated models.*

The results demonstrate significant predictive performance, strengthened by enhanced clinical relevance. This research lays the groundwork for a smart clinical decision-support system, especially useful for early CKD.

Keywords : *Chronic Kidney Disease, early prediction, artificial intelligence, machine learning, clinical data, eGFR, feature selection, data augmentation, class balancing, supervised models, data enrichment.*

مــلــخص

تُعدّ المرض الكلوي المزمن من أخطر الأمراض الصامتة، حيث يتطور تدريجيًا دون أعراض واضحة إلى أن يصل إلى مرحلة لا رجعة فيها. إن غياب الأعراض في المراحل المبكرة، إلى جانب التأخر في التشخيص، يجعل من هذا المرض تحديًا كبيرًا للصحة العامة. في هذا السياق، تفتح الذكاء الاصطناعي آفاقًا واعدة للتنبؤ المبكر بالمرض، من خلال استغلال بيانات طبية غير متجانسة وغالبًا غير مكتملة.

تهدف هذه الأطروحة إلى تقديم مقاربة ذكية وعملية للتنبؤ المبكر بمرض الكلى المزمن، بالاعتماد على تقنيات متقدمة للتعلم الآلي. يتمثل الهدف الأساسي في تصميم نماذج تصنيف قوية، قادرة على اكتشاف الحالات المعرضة للخطر بدقة، رغم القيود الشائعة في البيانات الطبية، مثل الحجم المحدود، وعدم توازن الفئات، وتنوع المتغيرات (رقمية/فئوية)، ووجود القيم المفقودة.

ينقسم عملنا إلى أربع مساهمات متكاملة:

- إجراء تحليل مقارن بين تقنيات متعددة لاختيار الخصائص، إلى جانب نماذج تعلم خاضع للإشراف، بهدف تحديد أفضل التوليفات أداءً في التنبؤ بالمرض.
- تحقيق تجريبي دقيق باستخدام مجموعتي بيانات طبيتين مستقلتين لتقييم قوة النماذج وقدرتها على التعميم.
- تطوير استراتيجية متكاملة تركز على تحسين جودة البيانات، وتحويلها بطريقة تتناسب مع السياق الطبي، وتعزيزها بمؤشرات طبية ذات قيمة عالية، مثل معدل الترشيح الكبيبي المقدر (DFG) الذي تم حسابه وإدماجه في البيانات لتعزيز دقة التشخيص.
- وأخيرًا، زيادة حجم قاعدة البيانات باستخدام تقنيات ملائمة للبيانات الطبية، إلى جانب اختيار تكراري للخصائص الأكثر تميزًا، بما يعزز تمثيلية العينات ويزيد من موثوقية النماذج المُولَّدة.

أظهرت النتائج المحصلة كفاءة تنبؤية عالية، مدعومة بوجاهة سريرية قوية. وتضع هذه الدراسة الأساس لإنشاء نظام ذكي لدعم القرار الطبي، يُعدّ أداة فعّالة للكشف المبكر عن المرض، خاصة لدى الفئات المعرضة للخطر مثل مرضى السكري.

الكلمات المفتاحية: مرض الكلى المزمن، التنبؤ المبكر، الذكاء الاصطناعي، التعلم الآلي، البيانات الطبية، DFG، اختيار الخصائص، زيادة البيانات، توازن الفئات، النماذج الخاضعة للإشراف، إثراء البيانات.

Remercîment

Louange à Allah, le Tout-Puissant, le Clément, qui m'a accordé la patience, la force et la persévérance nécessaires pour mener à bien ce travail. Sans Sa miséricorde, ce parcours n'aurait pu être accompli.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse, ***Pr. Mohamed Ridda LAOUAR****, pour la confiance qu'il m'a témoignée, pour sa disponibilité constante, ses conseils éclairés, ses encouragements sincères et sa rigueur scientifique. Son encadrement a été précieux à chaque étape de cette recherche.*

Mes sincères remerciements vont également à ***Pr Siam Abderrahim****, mon co-directeur de thèse, pour son soutien indéfectible, ses encouragements motivants et son accompagnement bienveillant tout au long de ce travail.*

Je souhaite également remercier ***Pr. Sean B. EOM****, de Southeast Missouri State University, pour sa précieuse collaboration, la pertinence de ses conseils et son ouverture dans le cadre de nos travaux communs.*

Mes remerciements s'étendent à ***l'ensembles des enseignants et enseignants-chercheurs*** *que j'ai eus tout au long de mon parcours, du premier au second cycle, qui ont contribué à ma formation et m'ont transmis le goût du savoir. Qu'Allah les récompense pour leur noble mission.*

Je remercie chaleureusement ***les membres du laboratoire LAMIS****, pour leur accueil, leur disponibilité, et pour l'environnement de travail stimulant et amical qu'ils ont su créer.*

Je rends un hommage sincère à ma famille, en particulier à ***mes parents****, pour leur amour inconditionnel, leurs prières, leurs sacrifices et leur soutien moral inestimable. À mon* ***frère et mes sœurs****, pour leur présence constante et leurs encouragements.*

Enfin, une pensée affectueuse à ***mes amis et collègues****, pour leur soutien moral, leur bonne humeur et tous les moments partagés. À toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de cette thèse, je dis :* ***merci du fond du cœur.***

Dédicace

À mes chers parents,
Pour votre amour inestimable, vos prières, vos sacrifices et votre soutien indéfectible. Ce travail est le prolongement de votre patience et de votre confiance. Qu'Allah vous comble de Sa miséricorde.

À mon oncle ***Djamel*** *et ma tante* ***Narimène****,*
Pour leur présence bienveillante, leur aide précieuse et leur soutien sincère à chaque étape de ce parcours. Merci du fond du cœur.

À mon frère et mes sœurs,
Pour leur affection, leurs encouragements constants et leur foi en moi.

À mes enseignants,
Qui m'ont transmis les clés du savoir et l'esprit critique nécessaire à cette aventure scientifique.

A my rzouga
Pour ses remarques précieuses, ses observations pertinentes, ses conseils sincères et son encouragement constant. Sa présence et son soutien ont énormément compté pour moi.

À mes copines fidèles,
Pour leur écoute, leur bonne humeur, et leur présence apaisante durant les moments difficiles.

À moi-même,
Pour avoir persévéré malgré les épreuves, douté sans abandonner, et rêvé sans renoncer.

Table des matières

Table des matières .....

Liste de Tableaux.....

Liste de Figures .....

Introduction générale.....

Chapitre 1:

Contexte de la maladie rénale chronique

1.1 Introduction 1

1.2 Les reins..... 1

1.3 La maladie rénale chronique (MRC)..... 2

1.4 Prévalence et tendances de la maladie rénale à l'échelle mondiale:..... 3

1.4.1 Prévalence de la MRC en Algérie 4

1.5 Les Causes et les facteurs risques de la maladie rénale chronique..... 5

1.5.1 Les Causes de la maladie rénale chronique 5

1.5.1.1 Le diabète 5

1.5.1.2 L'hypertension artérielle 6

1.5.1.3 Glomérulonéphrite..... 6

1.5.1.4 Les maladies génétiques 7

1.5.2Les facteurs risques de la maladie rénale chronique 7

1.5.2.1 Le diabète: 7

1.5.2.2 L'Hypertension artérielle : 7

1.5.2.3 L'obésité et ses effets nocifs sur la fonction rénale..... 8

1.5.2.4 Le tabagisme..... 8

1.5.2.5 L'âge..... 8

1.5.2.6 A Les antécédents familiaux et les prédispositions génétiques: 8
1.5.3 Impact du diabète sur la maladie rénale chronique 8
1.5.4 Importance du dépistage précoce chez les patients diabétiques 10
1.6 Diagnostique de la MRC 10
1.6.1 Biomarqueurs de la fonction rénale 11
1.6.1.1 Débit de filtration glomérulaire (DFG) 11
1.6.1.2 L'albuminurie 11
1.6.2 Stade et classification de la maladie rénale chronique 12
1.6.3 Méthodes de calcul du débit de filtration glomérulaire (DFG) 13
1.6.3.1 Formule CKD-EPI (chronic kidney disease- Epidemiology collaboration) 13
1.6.3.2 Formule Cockroft et Gault (CG) 14
1.6.3.3 Formule MDRD (Modification of Diet in Renal Disease Study) 14
1.7 Dépistage de la maladie rénale chronique 14
1.7.1 Excrétion urinaire d'albumine 14
1.7.2 Physiopathologie 14
1.7.3 Indications et réalisation du dosage 15
1.8 Conclusion 16

Chapitre 2:

L'intelligence artificielle (IA)

2.1 Introduction 18
2.2 L'intelligence artificielle (IA) 18
2.3 Apprentissage automatique 19
2.3.1 Apprentissage supervisé 20
2.3.2 Apprentissage non supervisé 21
2.3.3 Apprentissage par renforcement 22
2.4 Les principales phases du processus de machine learning 22

2.4.1 Collecte des données 23
2.4.2 Préparation des données 23
2.4.3 Choix du modèle 24
2.4.4 Entraînement du modèle 24
2.4.5 Évaluation du modèle 24
2.4.6 Ajustement des hyperparamètres 24
2.4.7 Prédiction 24
2.5 L'apprentissage profond: 24
2.6 Différence entre Machine Learning et Deep learning 25
2.7 Applications du Machine Learning et Deep learning dans le domaine de la santé (médecine) 26
2.8 Exemples d'IA dans les systèmes de santé 27
2.9 Rôle de l'IA dans la médecine personnalisée et préventive: 27
2.10 Conclusion 28

Chapitre 3:

L'IA pour la prédiction de la MRC : Revue de la littérature

3.1 Introduction 31
3.2 Stratégies récentes basées sur l'IA pour la prédiction de la maladie rénale chronique 31
3.3 La Médecine prédictive et la prédiction précoce de la MRC 32
3.4 Les méthodes traditionnelles pour la prédiction des maladies chronique 32
3.5 Intégration de l'apprentissage automatique dans la prédiction de la MRC 33
3.5.1 Collecte et structuration des données dans le domaine de la santé 33
3.5.1.1 Les types de données dans le domaine médicale 33
3.5.2 Les outils de préparation et de prétraitement des données 35
3.5.2.1 Nettoyage des données 35
3.5.3 Augmentation des données 40
3.5.4 Extraction et sélection des caractéristiques 42

3.5.4.1 Algorithme de sélection et de réduction minimale absolue (LASSO) 42
3.5.4.2 Élimination des caractéristiques récursives (RFE) 43
3.5.4.3 RELIEF 43
3.5.4.4 Analyse en Composantes Principales (PCA) 44
3.5.4.5 Boruta 44
3.5.4.6 Autres techniques de sélection des caractéristiques 45
3.5.5 Les modèles de ML utilisés dans la littérature 45
3.5.5.1 Classificateurs traditionnels 45
3.5.5.2 Méthodes d'ensembles 47
3.5.5.3 Réduction de dimension et régression 48
3.5.6 Les modèles de DL utilisés dans la littérature 49
3.5.6.1 Artificial Neural Network 49
3.5.6.2 Réseau de neurones convolutifs (CNN) 49
3.5.7 Évaluation des performances des modèles IA 50
3.6 Analyse des travaux récents sur la prédiction de la MRC 51
3.7 Analyse critique des approches existantes et opportunités d'amélioration 55
3.8 Conclusion 56

Chapitre 4 :

Contributions et Evaluation (Partie 2)

4.1 Introduction 58
4.2 Contribution 1 : Efficient Chronic Kidney Disease Prediction: A Comparative Analysis Using Feature Selection and Machine Learning models 58
4.2.1 Description de l'ensembles de données : 58
4.2.2 Visualisation de la distribution de l'ensembles de données : 60
4.2.3 Corrélation entre les variables 63
4.2.4 Prétraitement des données : 64
4.2.5 Méthode de sélection des caractéristiques : 66

4.2.6 Les algorithmes d'apprentissage automatique utilisés : 67

4.2.7 Modèle hybride proposé : 68

4.2.8 Résultats expérimentaux de la contribution 1 : 68

4.2.9 Discussion des résultats de la contribution 1 71

4.3 Contribution 2 : Prédiction précoce de la néphropathie chronique à l'aide de la sélection de caractéristiques et de l'apprentissage supervisé 72

4.3.1 Méthode de sélection des caractéristiques utilisé dans la contribution 2 : 73

4.3.2 Modélisation 74

4.3.3 Résultats expérimentaux de la contribution 2 : 75

4.3.4 Discussion 77

4.4 Conclusion 78

Chapitre 5 :

Contributions et Evaluation (Partie 2)

5.1 Introduction 81

5.2 Conception de l'approche proposé (Contribution 3): An innovative approach for predictive modeling and staging of chronic kidney disease 81

5.3 Ensembles de données : 82

5.4 Prétraitement des données 85

5.5 Sélection des caractéristiques 87

5.6 Modélisation 89

5.7 Résultats expérimentaux et discussion 89

5.7.1 Identification et répartition des stades de la MRC 89

5.7.2 Comparaison des résultats obtenus pour chaque modèle sur le premier jeu de données 90

5.7.2.1 Comparaison des modèles selon l'exactitude avec PCA 91

5.7.2.2 Comparaison des modèles selon la précision avec PCA 91

5.7.2.3 Comparaison des modèles selon le rappel avec PCA 92

5.7.2.4 Comparaison des modèles selon le score F1 avec PCA 93

5.7.2.5 Analyse de l'AUC du modèle hybride pour la prédiction de la MRC (PCA) 94

5.7.2.6 Comparaison des modèles selon l'exactitude avec LASSO 95

5.7.2.7 Comparaison des modèles selon la précision avec LASSO 96

5.7.2.8 Comparaison des modèles selon le rappel avec LASSO 96

5.7.2.9 Comparaison des modèles selon le F1-Score avec LASSO 97

5.7.2.10 Analyse comparative des performances de l'Accuracy avec les sélections des caractéristiques LASSO et PCA 98

5.7.3 Comparaison des performances de l'Accuracy des modèles sur le jeu de données de test 99

5.7.4 Comparaison des performances de notre approche avec des travaux connexes 100

5.8 Contribution 4 : Approche d'augmentation intelligente des données 101

5.8.1 Augmentation des données 102

5.8.2 Distribution des caractéristiques après l'augmentation des données 102

5.8.3 Validation de l'intégrité et de la cohérence des données augmentées 106

5.8.4 Résultats expérimentaux et discussion 106

5.8.4.1. Analyse des courbes ROC et des valeurs d'AUC 108

5.8.4.2. Évaluation des modèles sur l'ensembles detest 111

5.8.5 Comparaison des résultats obtenus avec d'autres travaux connexes 112

5.9 Conclusion 114

Conclusion générale et perspectives 116

Références bibliographiques 119

Liste de Tableaux

Tableau 1.1: Classification des stades de la MRC [58] 12

Tableau 2.1: Tableau comparatif entre la machine learning et le deep learning 25

Tableau 4.1: Description des attributs de l'ensembles de données et leurs typologies 59

Tableau 4.2: Résultats de performance de chaque modèle 69

Tableau 4.3: Description de l'ensembles de données de test 72

Tableau 4.4: Évaluation des modèles avec et sans sélection des caractéristiques (LASSO) ... 75

Tableau 4.5: Comparaison de la performance de nos modèles avec des modèles issus de travaux connexes 78

Tableau 5.1: Comparaison des fonctionnalités entre les ensembles de données d'entrainement et de test 83

Tableau 5.2: Comparaison de l'Accuracy de notre étude avec les travaux existants 100

Tableau 5.3: Évaluation de différents classificateurs pour différents ratios de division basés sur une validation croisée k-fold 105

Tableau 5.4: Comparaison de l'approche proposée avec d'autres travaux 113

Liste de Figures

Figure 1.1: Anatomie du rein [8] 2

Figure 1.2: Prévalence de la MRC dans le monde [17] 3

Figure 1.3: PRÉVALENCE DE LA MRC DANS LE MONDE, PAR SEXE [18] 4

Figure 1.4: Répartition géographique des cas de MRC au stade V dans le sud-est de l'Algérie, classés par wilaya de résidence [19] 5

Figure 1.5: Schéma comparatif d'un rein sain et d'un rein atteint de maladie rénale chronique chez un patient diabétique [53] 10

Figure 1.6: Pronostic de la maladie rénale chronique basé sur les catégories de débit de filtration glomérulaire et d'albuminurie [55] 13

Figure 2.1: Le processus de Machine Learning 23

Figure 3.1: Matrice de confusion illustrant les principaux indicateurs de performance 51

Figure 4.1: Distribution de la classe cible (Classification) 60

Figure 4.2: Distribution des variables l'Age et la glycémie 61

Figure 4.3: Distribution des variables Pression artérielle et Urée sanguine 61

Figure 4.4: Distribution des variables l'hémoglobine et Volume globulaire 62

Figure 4.5: Distribution des variables potassium et globules rouges 62

Figure 4.6: Distribution de Créatinine sérique et sodium 63

Figure 4.7: Distribution des variables globule blanche 63

Figure 4.8: Carte de corrélation des variables numériques de l'ensembles de données 64

Figure 4.9: Les caractéristiques importantes sélectionnés par RFE et leurs scores 67

Figure 4.10: Résultats de l'accuracy de chaque modèle 69

Figure 4.11: Analyse des résultats de la prédiction de la maladie rénale chronique à l'aide de RF et AdaBoost 70

Figure 4.12: Analyse des résultats de la prédiction de la maladie rénale chronique à l'aide de Bagging et le modèle hybride 70

Figure 4.13: Comparaison des résultats de la contribution 1 avec des travaux connexes 71

Figure 4.14: Les caractéristiques importantes sélectionnées par la méthode LASSO et leurs coefficients 74

Figure 4.15: Performances des modèles utilisant la sélection de fonctionnalités LASSO 76

Figure 4.16: Courbe d'apprentissage pour KNN et Gradient Boosting 76

Figure 4.17: Courbe ROC des modèles KNN et Gradient Boosting 77

Figure 5.1: Architecture de notre approche 82

Figure 5.2: Aperçu du dataset d'entraînement après ajout du DFG 83

Figure 5.3: Distribution des caractéristiques numériques avant l'imputation 86

Figure 5.4: Distribution des caractéristiques numériques après l'imputation 86

Figure 5.5: Distribution des variables catégorielles 87

Figure 5.6: Les caractéristiques importantes sélectionnés par LASSO 88

Figure 5.7: Variance expliquée par les composantes principales 88

Figure 5.8: Distribution des stades de l'MRC 90

Figure 5.9: Distribution du taux de DFGe par stade de l'MRC 90

Figure 5.10: Résultats de l'Accuracy obtenus à l'aide de PCA 91

Figure 5.11: Résultats de la précision obtenus à l'aide de PCA 92

Figure 5.12: Résultats de la précision obtenus à l'aide de PCA 93

Figure 5.13: Résultats de F1-Score avec PCA 94

Figure 5.14: Analyse ROC du modèle hybride pour la prédiction de la MRC 95

Figure 5.15: Résultats des scores de l'Accuracy à l'aide de LASSO 96

Figure 5.16: Résultats des scores de précision à l'aide de LASSO 96

Figure 5.17: Résultats des scores de Recall à l'aide de LASSO 97

Figure 5.18: Résultats des scores de F1-Score à l'aide de LASSO 98

Figure 5.19: Résultats de l'Accuracy des modèles avec LASSO et PCA 99

Figure 5.20: Résultats de l'Accuracy de nos modèles sur l'ensembles de données de test 100

Figure 5.21: Distribution de la classe cible après SMOTE 105

Figure 5.22: Distribution des caractéristiques apprès augmentation 103

Figure 5.23: Distribution de la classe cible avant après augmentation 104

Figure 5.24: Vérification de l'intégrité des données avant et après augmentation 104

Figure 5.25: Courbe ROC des modèles RF, DT, GNB et AdaBoost 106

Figure 5.26: Performance en Accuracy des classificateurs selon plusieurs ratios de validation croisée k-fold 107

Figure 5.27: Performance en Précision des classificateurs selon plusieurs ratios de validation croisée k-fold 107

Figure 5.28: Performance en Recall des classificateurs selon plusieurs ratios de validation croisée k-fold 108

Figure 5.29: Performance en F1-Score des classificateurs selon plusieurs ratios de validation croisée k-fold 108

Figure 5.30: Evaluation des performances des modèles sur le datatset de test 112

Figure 5.31: Matrices de confusion des modèles sur le datatset de test 112

Liste des abréviations

IA : Intelligence Artificielle

MRC : Maladie Rénale Chronique

DFG : Débit de Filtration Glomérulaire

OMS : Organisation Mondiale de la Santé

IRT : Insuffisance Rénale Terminale

TRR : Traitement de Remplacement Rénal

MRD : Maladie Rénale Diabétique

ND : Néphropathie Diabétique

CKD-EPI : Chronic Kidney Disease Epidemiology Collaboration

CG : Cockcroft–Gault

MDRD : Modification of Diet in Renal Disease

HAS : Haute Autorité de Santé

DME : Dossier Médical Électronique

ML : Machine Learning

KNN : K-Nearest Neighbors

SVM : Support Vector Machine

PCA : Principal Component Analysis

AUC : Area Under the Curve

SMOTE : Synthetic Minority Over-sampling Technique

ADASYS : Adaptive Synthetic Sampling

LASSO : Least Absolute Shrinkage and Selection Operator

RFE : Recursive Feature Elimination

RF : Random Forest

AdaBoost : Adaptive Boosting

LDA : Linear Discriminant Analysis

ANN : Artificial Neural Network

BPN : Backpropagation Neural Network

CNN : Convolutional Neural Network

GB : Gradient Boosting

DT : Decision Tree

LR : Logistic Regression

ROC : Receiver Operating Characteristic

GAN : Generative Adversarial Network

VAE : Variational Autoencoder

SMOTENC : Synthetic Minority Over-sampling Technique for Nominal and Conti

ROS : Random Over-Sampling

Introduction Générale

Introduction générale

L'intelligence artificielle (IA) représente aujourd'hui une révolution technologique majeure, dont les répercussions s'étendent bien au-delà des domaines classiques de l'industrie et de l'informatique. Depuis ses origines dans les années 1950, l'IA s'est progressivement imposée comme un levier incontournable dans la résolution de problèmes complexes, en s'appuyant sur des algorithmes capables d'apprendre, de raisonner et de prendre des décisions. Ces capacités ont transformé la manière dont les données sont exploitées, notamment dans les secteurs à forte intensité informationnelle, comme la médecine. Grâce aux avancées considérables en apprentissage automatique (Machine Learning) et en apprentissage profond (Deep Learning), il est désormais possible d'extraire des connaissances pertinentes à partir de ensembles de données considérables, hétérogènes et souvent bruités, ce qui ouvre des perspectives inédites en matière de diagnostic, de prédiction et de suivi des maladies.

Dans le domaine médical, l'IA joue un rôle de plus en plus central, en venant compléter l'expertise des professionnels de santé. Elle permet d'automatiser des tâches complexes, de détecter précocement certaines pathologies à partir de signaux faibles, et de personnaliser les traitements en fonction des profils cliniques des patients. Ces technologies intelligentes offrent la possibilité de passer d'une médecine réactive, entrée sur la maladie, à une médecine prédictive et préventive, orientée vers l'anticipation et l'amélioration continue de la qualité des soins. Ainsi, l'intégration de l'IA dans les systèmes de santé est aujourd'hui au cœur de la transition vers la médecine dite "4P" : prédictive, préventive, personnalisée et participative.

Parmi les maladies chroniques les plus préoccupantes à l'échelle mondiale, la maladie rénale chronique (MRC) occupe une place singulière. Cette pathologie silencieuse, évolutive et irréversible, se caractérise par une altération progressive de la fonction rénale, pouvant aboutir à une insuffisance rénale terminale nécessitant une dialyse ou une transplantation. D'après les estimations de l'Organisation mondiale de la santé, environ 10 % de la population mondiale est atteinte de MRC, avec un taux de prévalence en constante augmentation en raison du vieillissement démographique et de l'essor des pathologies associées telles que le diabète et l'hypertension artérielle. Le diabète sucré, est reconnu comme la cause principale de la maladie rénale chronique (MRC) dans le monde. L'hyperglycémie chronique endommage progressivement les vaisseaux sanguins des reins, entraînant une détérioration lente mais irréversible de leur fonction. Ainsi, une grande proportion de patients diabétiques développent une atteinte rénale, souvent silencieuse dans ses premiers stades. En Algérie, bien que les données épidémiologiques restent précises limitées, les néphrologues s'accordent à dire que la MRC constitue un enjeu majeur de santé publique, notamment en raison du manque de dépistage précoce, du faible accès aux soins spécialisés, et des coûts élevés des traitements substitutifs.

La prédiction tardive de la MRC est particulièrement problématique, car les symptômes cliniques n'apparaissent souvent qu'à un stade avancé de la maladie, lorsque les lésions rénales sont déjà sévères. Dans ce contexte, le recours à des outils de prédiction automatisés basés sur l'IA permet de mieux identifier les patients à risque, d'orienter rapidement les décisions

médicales et de mettre en place des mesures de prévention individualisées. Ces outils intelligents reposent sur la capacité des algorithmes à analyser des données hétérogènes (biologiques, cliniques, historiques) et à détecter des schémas prédictifs invisibles aux approches traditionnelles. Ils peuvent, à terme, s'intégrer aux systèmes d'aide à la décision clinique pour soutenir les praticiens dans leurs diagnostics et améliorer les trajectoires de soins.

Cependant, malgré ces promesses, la mise en œuvre effective de ces approches dans le cadre de la prédiction de la MRC soulève de nombreux défis scientifiques et techniques. D'abord, les données médicales disponibles sont souvent limitées en, comme c'est le cas du jeu de données issu du référentiel UCI Machine Learning, qui ne contient que 400 enregistrements. Deuxièmement, ces données présentent un déséquilibre des classes important (entre patients atteints et non atteints), ce qui peut biaiser l'apprentissage des modèles et réduire leur performance. Troisièmement, la présence de valeurs manquantes, fréquentes dans les bases médicales, complique le traitement et la modélisation. Quatrièmement, les variables sont d'hétérogène : certaines sont continués (taux de créatinine, hémoglobine), d'autres catégorielles (présence ou absence d'anémie, d'œdème, d'hypertension). Cette complexité nécessite l'adoption de méthodes de prétraitement adaptées et de techniques d'imputation rigoureuses

Un autre enjeu fondamental dans la construction de modèles prédictifs fiables pour la prédiction de la maladie rénale chronique (MRC) concerne le choix judicieux des variables cliniques et biologiques à intégrer. Toutes les caractéristiques présentes dans un jeu de données médicales ne présentent pas nécessairement une valeur informative égale, et certaines peuvent même s'avérer redondantes, non pertinentes ou nuisibles à la performance du modèle. L'inclusion de variables non discriminantes peut conduire à un phénomène de surapprentissage, où le modèle s'adapte excessivement aux spécificités du jeu de données d'entrée. Isoler un sous-ensemble optimal de caractéristiques pertinentes, à la fois informatives, non corrélées de manière excessive, et représentatives des mécanismes physiopathologiques sous-jacents à la MRC. Cette démarche vise à renforcer à la fois la robustesse, la simplicité et l'interprétabilité des modèles générés. En affinant le choix des variables, on améliore non seulement la précision des prédictions, mais également la lisibilité des résultats pour les professionnels de santé, ce qui constitue un critère décisif pour une adoption concrète des modèles en pratique clinique.

Enfin, un aspect fondamental, mais encore trop souvent négligé dans la littérature scientifique, réside dans l'enrichissement clinique des données utilisées pour entraîner les modèles prédictifs. La majorité des travaux existants s'appuient uniquement sur des variables biologiques ou cliniques de base (créatinine, albumine, pression artérielle, etc.) sans intégrer des marqueurs plus spécifiques, pourtant essentiels pour mieux représenter l'état fonctionnel du rein. Parmi ceux-ci, le taux de filtration glomérulaire augmentant, ou DFG (eGFR pour Estimation Glomerular Filtration Rate), occupe une place centrale. Ce biomarqueur, recommandé par les sociétés internationales de néphrologie, est aujourd'hui reconnu comme un critère essentiel pour améliorer la prédiction de la MRC et en estimer les stades de progression. Il offre une vision plus complète de l'évolution de la maladie qu'aucun autre paramètre isolé. En fournissant une estimation standardisée de la capacité des reins à filtrer le sang, le DFG permet aux modèles prédictifs non seulement de détecter plus précocement les cas de MRC, mais également d'anticiper la gravité potentielle. Malgré son importance médicale avérée, ce marqueur reste

étonnamment absent de nombreuses études dans le domaine de l'intelligence artificielle appliquée à la santé.

L'absence de cette variable dans la majorité des bases de données utilisées dans les études antérieures constitue une lacune majeure, car elle prive les modèles d'une information à très forte valeur ajoutée, cruciale pour améliorer la précision et la pertinence médicale des prédictions. Le simple ajout du DFG permet non seulement de mieux prédire les cas de MRC, mais aussi d'en préciser la sévérité, ce qui est essentiel dans une logique de médecine préventive. Intégrer cet indicateur dans une démarche d'intelligence artificielle, c'est donc rapprocher l'algorithme de la logique médicale réelle, en renforçant la robustesse et l'interprétation du modèle.

C'est dans cette optique que notre travail s'inscrit : nous avons choisi de calculer et d'ajouter le DFG à notre jeu de données initiales, afin d'enrichir le profil clinique de chaque patient et d'augmenter la capacité des algorithmes à discriminer correctement les individus atteints de MRC. À notre connaissance, aucun des travaux publiés à ce jour n'intègrent cette variable dans une démarche automatisée de prédiction, ce qui confère à notre approche une originalité méthodologique et une pertinence clinique renforcée.

Ainsi, cette thèse vise à proposer une approche rigoureuse, méthodologiquement fondée et médicalement pertinente pour la prédiction de la maladie rénale chronique, en mobilisant les outils de l'intelligence artificielle de manière pragmatique. À travers quatre contributions complémentaires, nous abordons successivement le prétraitement des données, la sélection des caractéristiques, la comparaison de modèles supervisés, l'enrichissement du jeu de données avec le DFG, et l'augmentation du volume d'apprentissage à travers des techniques de l'IA. L'objectif ultime est de construire des modèles prédictifs qui soient non seulement performants sur le plan algorithmique, mais aussi généralisables, interprétables, et exploitables dans un contexte clinique réel.

L'objectif principal de cette thèse est de développer une approche intelligente et médicalement pertinente pour la prédiction précoce de la maladie rénale chronique (MRC), en s'appuyant sur les techniques d'intelligence artificielle. Il s'agit de concevoir, expérimenter et valider des modèles de classification robustes, capables de détecter efficacement les cas de MRC à partir de données cliniques et biologiques, même lorsque celles-ci présentent des contraintes typiques du contexte médical réel, telles que la faible volumétrie, l'hétérogénéité des variables, le déséquilibre des classes ou encore la présence de valeurs manquantes.

Dans cette optique, nos objectifs spécifiques sont les suivants :

- Assurer une qualité optimale des données d'entrée, en appliquant des techniques de prétraitement appropriées, incluant une imputation rigoureuse des valeurs manquantes et une gestion cohérente des variables de nature mixte (numériques et catégorielles), afin de garantir la fiabilité et la stabilité des prédictions ;
- Sélectionner un sous-ensemble pertinent de variables cliniques, de manière à optimiser la performance des modèles tout en limitant leur complexité, grâce à une démarche de sélection fondée sur des critères à la fois statistiques et médicaux ;

- Enrichir le jeu de données existantes avec des informations à forte valeur ajoutée, en intégrant notamment débit de filtration glomérulaire estimé (eDFG), un biomarqueur fondamental de la fonction rénale, encore trop peu pris en compte dans les approches de prédiction automatique ;
- Augmenter la taille et la représentativité de l'ensembles de données, par l'emploi de techniques de suréchantillonnage compatibles avec les variables catégorielles, afin de corriger le déséquilibre des classes et d'améliorer la capacité de généralisation des modèles.

Contributions de la thèse :

- **Contribution 1** : Proposition d'une approche comparative pour la prédiction précoce de la MRC en utilisant plusieurs techniques de sélection de caractéristiques (LASSO, RELIEF, RFE) et divers modèles d'apprentissage automatique (Random Forest, AdaBoost, Bagging, modèle hybride), appliqués à un jeu de données issu du référentiel UCI, après un prétraitement incluant soigneusement l'imputation des valeurs manquantes.
- **Contribution 2** : Développement d'une approche prédictive robuste validée sur deux ensembles de données distinctes (400 et 202 échantillons), permettant de renforcer la généralisation des modèles et d'évaluer la stabilité des prédictions sur des données indépendantes, avec une sélection de variables cliniques et biologiques liées aux facteurs de risque de la MRC.
- **Contribution 3** : Conception d'une approche méthodologiquement structurée, pour la prédiction de la MRC. Cette approche inclut le calcul du taux de filtration glomérulaire (DFGe) à partir des variables biologiques disponibles, conformément aux formules médicales validées, puis l'ajout systématique de cette variable au dataset initial afin d'enrichir le profil clinique des patients. L'objectif est d'améliorer la cohérence statistique, la validité clinique et la capacité de détection des modèles, tout en assurant la reproductibilité et l'applicabilité dans un contexte médical réel.
- **Contribution 4** : Enrichissement du jeu de données initiales, à la fois en volume et en contenu, par la génération contrôlée de nouveaux échantillons permettant de rééquilibrer la distribution des classes, et par l'intégration d'une caractéristique clinique à forte valeur ajoutée calculée à partir des données disponibles, renforçant ainsi la représentativité médicale de chaque instance. Le jeu de données final, élargi à 1000 observations, a ensuite été exploité pour l'évaluation croisée de plusieurs modèles de classification afin d'assurer la robustesse et la pertinence des résultats

Structure de la thèse :

- **Chapitre 1** : Présentation du contexte médical et épidémiologique de la maladie rénale chronique.
- **Chapitre 2** : Introduction aux fondements de l'intelligence artificielle et des techniques d'apprentissage appliquées à la santé.
- **Chapitre 3** : Revue de la littérature sur les travaux liés à la prédiction de la MRC à l'aide de l'IA.
- **Chapitre 4** : Présentation de la première partie des contributions, incluant les approches classiques de sélection de variables, les modèles testés et les résultats obtenus.
- **Chapitre 5** : Seconde partie des contributions, portant sur l'enrichissement des données avec le DFG, l'augmentation du dataset, l'application de méthodes avancées de sélection de caractéristiques et la comparaison expérimentale des modèles.

- **Conclusion générale** : Synthèse des apports de la thèse, bilan méthodologique, perspectives d'amélioration et ouverture vers des applications cliniques réelles.

Chapitre 1 :

Contexte de la maladie rénale chronique

1.1 Introduction

La maladie rénale chronique (MRC) est un problème de santé publique majeur selon l'Organisation Mondiale de la Santé (OMS), touchant environ 10 % de la population mondiale, avec des millions de décès chaque année en raison de l'absence d'un accès abordable aux traitements. Cette pathologie se caractérise par une altération progressive et irréversible de la fonction rénale, entraînant des complications graves pouvant aboutir à une insuffisance rénale terminale (IRT), nécessitant alors une dialyse ou une transplantation rénale pour assurer la survie des patients. L'une des principales difficultés liées à la MRC réside dans son caractère silencieux aux premiers stades, rendant son diagnostic précoce complexe. En l'absence de symptômes clairs, la maladie évolue souvent de manière asymptomatique jusqu'à un stade avancé, où les dommages rénaux deviennent irréversibles. Par conséquent, une prédiction précoce et une prise en charge efficace sont essentielles pour ralentir la progression de la maladie et améliorer la qualité de vie des patients.

Dans ce chapitre, nous aborderons la maladie rénale chronique (MRC) en commençant par le rôle et la physiologie du rein. Nous définirons ensuite la MRC, ses causes et facteurs de risque, en mettant l'accent sur l'impact du diabète. Nous présenterons également les stades, la classification et les méthodes d'évaluation de la fonction rénale. Enfin, nous discuterons du dépistage, des complications associées et des traitements disponibles, comme la dialyse et la transplantation rénale.

1.2 Les reins

Les reins sont des organes vitaux du système urinaire, situés de part et d'autre de la colonne vertébrale, sous les côtes, en position rétropéritonéale. Ils sont de forme haricot, mesurant environ 10 à 12 cm de long, 5 à 7 cm de large et 2 à 3 cm d'épaisseur, avec un poids variant entre 125 et 170 g chez l'adulte [1]-[2]. Chaque rein est composé de deux parties principales : le parenchyme rénal, qui assure la filtration du sang, et le système collecteur, responsable de l'élimination de l'urine [3].

Le rein joue un rôle fondamental dans le maintien de l'homéostasie corporelle. Sa fonction principale est d'éliminer les déchets métaboliques, tels que l'urée, la créatinine et l'acide urique, qui, en s'accumulant, deviennent toxiques pour l'organisme [4]-[5]. Chaque jour, il filtre 120 à 150 litres de sang pour produire 1 à 2 litres d'urine, permettant ainsi l'évacuation des toxines et la régulation des fluides corporels [6]. Ce processus de filtration se déroule dans les néphrons, les unités fonctionnelles du rein, qui sont environ un million par rein. Chaque néphron est composé d'un glomérule, où le plasma est filtré, et d'un tubule rénal, où l'eau et certains éléments essentiels sont réabsorbés avant que le sang épuré ne retourne à la circulation sanguine [7].

En plus de son rôle d'épuration, le rein est un organe endocrinien qui produit plusieurs hormones essentielles. Il sécrète notamment l'érythropoïétine (EPO), qui stimule la production de globules rouges, et le calcitriol, la forme active de la vitamine D, qui participe à l'absorption du calcium et à la minéralisation osseuse [2]. Il joue également un rôle clé dans la régulation de la pression artérielle par la sécrétion de rénine, une enzyme impliquée dans le système rénine-angiotensine-aldostérone (RAA), ainsi que dans l'équilibre acido-basique par l'excrétion des ions hydrogène (H^+) et la rétention des ions bicarbonate (HCO_3^-) [3]-[4].

Ainsi, les reins ne sont pas seulement des organes d'excrétion, mais également des régulateurs essentiels de l'organisme, assurant un équilibre optimal des fluides, des électrolytes et des hormones pour le bon fonctionnement du corps humain.

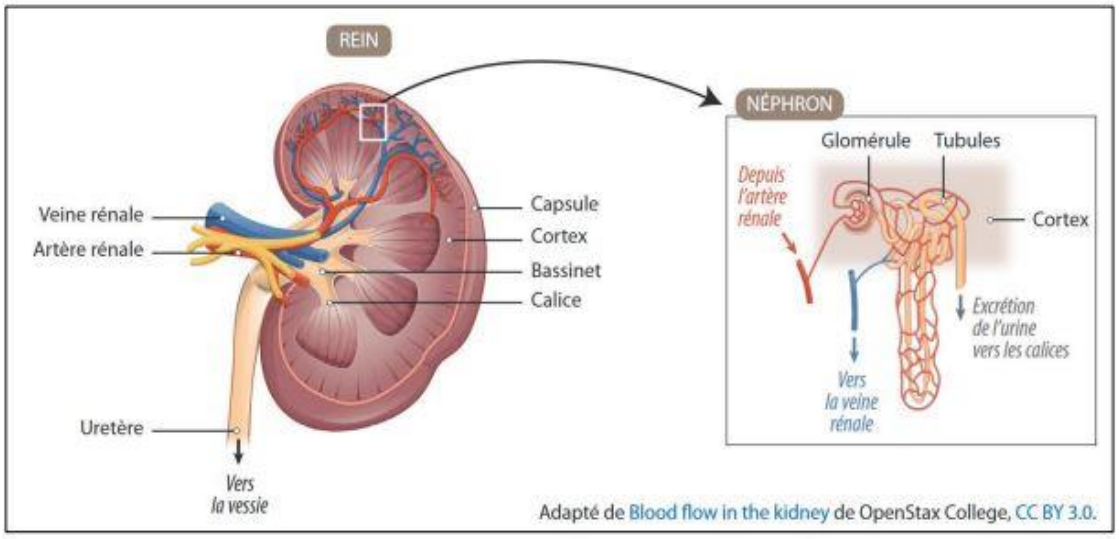


Figure 1.1: Anatomie du rein [8]

1.3 La maladie rénale chronique (MRC)

La maladie rénale chronique (MRC) est une pathologie progressive caractérisée par une détérioration lente et irréversible de la fonction rénale sur une période d'au moins trois mois [9]. Cette affection constitue un enjeu majeur de santé publique mondiale, affectant environ 10 % de la population, soit plus de 800 millions de personnes [9]-[10].

La MRC est souvent asymptomatique à ses premiers stades, ce qui explique pourquoi près de 90 % des patients ignorent leur maladie jusqu'à ce qu'ils aient déjà perdu 25 % de leur fonction rénale [11]-[12].

Les reins jouent un rôle essentiel dans l'organisme en filtrant le sang pour éliminer les déchets métaboliques et l'excès d'eau, produisant ainsi l'urine [13]. Lorsque les reins sont endommagés, ces déchets s'accumulent dans le corps, entraînant des déséquilibres hydriques, électrolytiques et hormonaux. La MRC étant une maladie qui évolue sur le long terme, elle est classée en plusieurs stades selon le débit de filtration glomérulaire (DFG). Aux premiers stades, la fonction rénale est altérée mais encore suffisante pour assurer l'homéostasie. Aux stades avancés, la filtration devient insuffisante, nécessitant un traitement de suppléance comme la dialyse ou la greffe rénale. Elle est souvent diagnostiquée tardivement, ce qui complique sa prise en charge et augmente le risque de progression vers l'insuffisance rénale chronique terminale nécessitant une dialyse ou une greffe rénale [14].

Plusieurs facteurs augmentent le risque de développer une MRC, notamment le diabète, l'hypertension artérielle et les maladies cardiovasculaires, qui sont les principales causes de cette pathologie [13]-[15]. En outre, des éléments tels que l'âge, le sexe et certaines prédispositions génétiques influencent également la susceptibilité à la maladie[16].

En raison de l'absence de symptômes évidents aux stades précoces, de nombreux patients ignorent leur état jusqu'à ce que la maladie atteigne un niveau avancé [16]. La prévention de la

MRC repose donc en grande partie sur le contrôle des maladies sous-jacentes, en particulier le diabète et l'hypertension, afin de limiter les dommages rénaux à long terme et d'améliorer la qualité de vie des patients [13]-[16].

1.4 Prévalence et tendances de la maladie rénale à l'échelle mondiale :

La prévalence de la MRC varie selon les régions du monde. En Chine, environ 10,8 % de la population adulte est touchée, tandis qu'aux États-Unis, cette proportion est estimée entre 10 % et 15 %. Au Mexique, la prévalence atteint 14,7 % de la population adulte. D'après une analyse mondiale du Global Burden of Disease (GBD), en 2017, le nombre total de cas d'insuffisance rénale chronique (tous stades confondus) était estimé à 697,5 millions, correspondant à une prévalence mondiale de 9,1 %. Par ailleurs, la mortalité liée à la MRC est en augmentation, avec 1,2 million de décès en 2017, chiffre ayant progressé à 1,43 million en 2019.En France, au 31 décembre 2021, on recensait 92 535 patients sous traitement de suppléance rénale, dont 55 % en dialyse et 45 % avec une greffe fonctionnelle. L'hémodialyse reste la technique la plus courante (94 % des cas), et 29 % des patients ont dû débuter leur dialyse en situation d'urgence. Aux États-Unis, environ 1 adulte sur 7 est atteint de MRC, et 9 patients sur 10 ignorent qu'ils en sont atteints.
(Source : Réseau Epidémiologie, Information, Néphrologie ; Fondation du rein)



Figure 1.2: Prévalence de la MRC dans le monde [17]

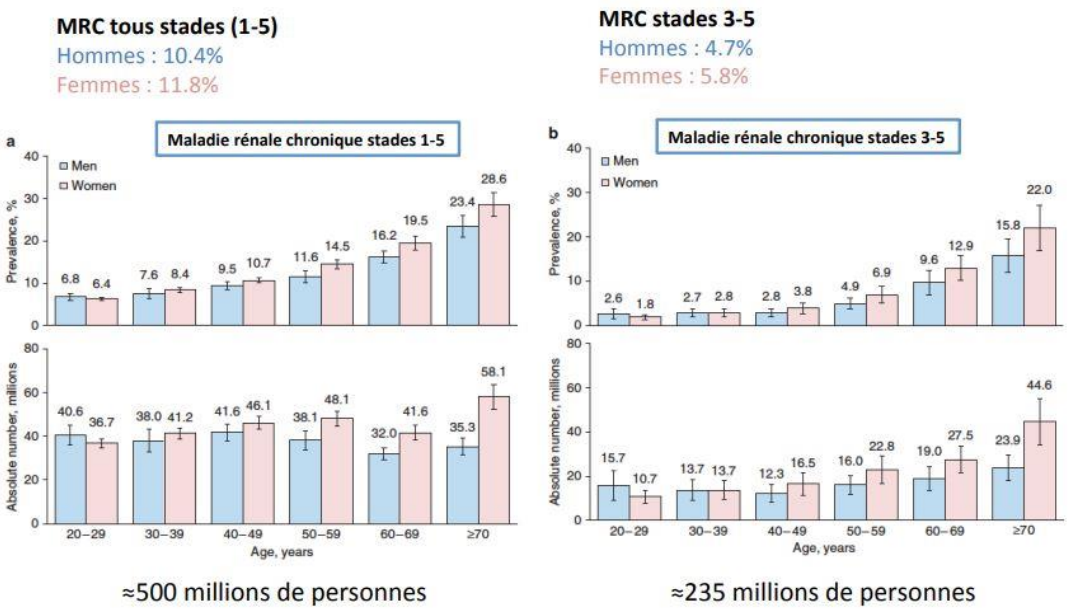


Figure 1.3: PRÉVALENCE DE LA MRC DANS LE MONDE, PAR SEXE [18]

1.4.1 Prévalence de la MRC en Algérie

En Algérie, la prévalence de la maladie rénale chronique (MRC) est en constante augmentation, en particulier au stade V, où l'insuffisance rénale terminale devient un problème majeur de santé publique. Une étude menée dans le Sud-Est du pays a révélé que la MRC au stade V touche une population relativement jeune, avec un âge moyen de 50,17 ans, et une prédominance masculine de 59,10 % [19]. Cette pathologie nécessite une prise en charge urgente dans 91,2 % des cas, et la majorité des patients hémodialysés chroniques (97,1 %) utilisent une fistule artérioveineuse comme abord vasculaire permanent. Bien que les patients de cette étude présentent moins de comorbidités et de handicaps que dans les pays développés, la mortalité reste élevée, soulignant la nécessité d'un registre national pour mieux comprendre l'épidémiologie de la MRC en Algérie et améliorer la prévention ainsi que la prise en charge des patients à travers le pays [19].

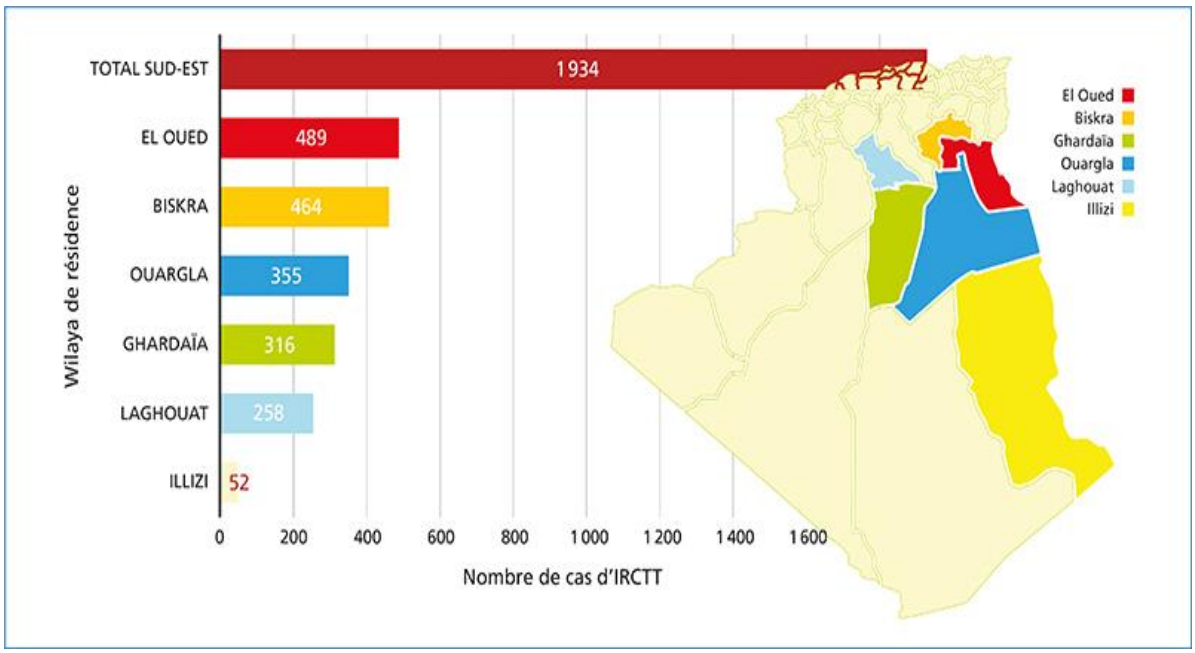


Figure 1.4: Répartition géographique des cas de MRC au stade V dans le sud-est de l'Algérie, classés par wilaya de résidence [19]

1.5 Les Causes et les facteurs risques de la maladie rénale chronique

La maladie rénale chronique (MRC) constitue un problème de santé publique majeur en raison de sa prévalence croissante et de son impact considérable sur la qualité de vie des patients. Caractérisée par une dégradation progressive et irréversible de la fonction rénale, pouvant évoluer vers une insuffisance rénale terminale nécessitant la dialyse ou la transplantation, elle est souvent diagnostiquée tardivement en raison de son évolution silencieuse aux premiers stades. L'identification des causes et des facteurs de risque est donc cruciale, non seulement pour prévenir la progression de la maladie, mais également pour en prédire l'évolution. Les données les plus fiables concernant la répartition des causes de la MRC proviennent des cas incidents de traitement de remplacement rénal (TRR), c'est-à-dire des patients qui débutent ce traitement pour la première fois. Bien que le diabète sucré soit la cause prédominante dans de nombreux contextes, une proportion importante de patients atteints d'insuffisance rénale terminale présente une étiologie indéterminée, malgré les progrès réalisés en matière de connaissances et de techniques diagnostiques [20]. Une meilleure compréhension de ces éléments est indispensable pour optimiser le dépistage précoce et adapter les stratégies de prise en charge aux besoins spécifiques des patients.

1.5.1 Les Causes de la maladie rénale chronique

1.5.1.1 Le diabète

Le diabète sucré constitue la principale cause de la maladie rénale chronique (MRC) dans la plupart des pays, qu'ils soient développés ou en voie de développement. Environ 30 à 40 % des patients atteints de diabète développeront une forme de MRC au cours de leur vie [21] . L'hyperglycémie chronique, caractéristique du diabète, provoque une inflammation et une fibrose glomérulaire, entraînant une altération progressive de la fonction rénale. Ces mécanismes expliquent pourquoi le diabète est la principale cause de MRC. Un suivi régulier

de la fonction rénale, notamment par la surveillance annuelle du taux de filtration glomérulaire (TFG) et du rapport albuminurie/créatinine urinaire (UACR), est essentiel. En présence de microalbuminurie ou de protéinurie, une gestion rigoureuse de la glycémie, associée à une prise en charge précoce de l'hypertension par l'utilisation d'inhibiteurs de l'enzyme de conversion de l'angiotensine (IEC) ou de bloqueurs des récepteurs de l'angiotensine (BRA), demeure l'intervention thérapeutique la plus efficace pour freiner l'évolution de la MRC [22][23].

1.5.1.2 L'hypertension artérielle

L'hypertension artérielle est à la fois une cause majeure et une conséquence fréquente de la maladie rénale chronique (MRC), jouant un rôle clé dans son apparition et sa progression. Elle représente la deuxième cause d'insuffisance rénale terminale, avec un risque accru en fonction de la sévérité de l'hypertension. Une pression artérielle systolique élevée (supérieure à 200 mm Hg) multiplie considérablement le risque d'insuffisance rénale terminale par rapport à une pression normale (120 mm Hg) [24]. Cette élévation prolongée de la pression entraîne une surcharge mécanique sur les glomérules rénaux, accélérant leur destruction et exacerbant la détérioration de la fonction rénale. Par ailleurs, l'hypertension favorise des complications vasculaires, telles que l'athérosclérose, pouvant provoquer des occlusions et des embolies qui aggravent encore la perfusion rénale. En plus d'être un facteur causal, l'hypertension est une manifestation quasi systématique de la MRC. Des études épidémiologiques ont démontré son rôle en tant que facteur de risque indépendant d'évolution vers l'insuffisance rénale terminale[24]. Toutefois, le lien de causalité directe entre l'hypertension et la MRC reste débattu, certaines formes d'hypertension pouvant résulter d'une pathologie rénale sous-jacente plutôt que l'inverse. Ainsi, tout patient nouvellement diagnostiqué hypertendu devrait bénéficier d'un dépistage systématique de la fonction rénale afin d'identifier précocement une éventuelle atteinte. La prise en charge repose sur un contrôle strict de la pression artérielle, notamment par l'administration d'antihypertenseurs néphroprotecteurs, permettant de ralentir la progression de la maladie et d'améliorer le pronostic rénal des patients [24]- [25] .

1.5.1.3 Glomérulonéphrite

La glomérulonéphrite désigne un ensembles de maladies rénales, primaires ou secondaires, caractérisées par une inflammation des glomérules, les unités fonctionnelles des reins responsables de la filtration du sang. Cette atteinte peut être d'origine auto-immune, infectieuse ou métabolique, et sa progression varie en fonction de la cause sous-jacente. L'inflammation des glomérules perturbe la filtration sanguine, entraînant une perte de protéines (protéinurie) et de sang (hématurie) dans les urines, ce qui peut accélérer la progression vers l'insuffisance rénale chronique [26]. Une prédiction précoce est essentielle pour limiter l'évolution vers l'insuffisance rénale chronique (IRC). Le diagnostic repose principalement sur la présence d'anomalies urinaires, notamment une protéinurie et une hématurie, détectables grâce à un test urinaire par bandelette. Tout patient présentant une insuffisance rénale aiguë (IRA) ou une maladie rénale chronique (MRC) devrait bénéficier de ce dépistage afin d'identifier une éventuelle glomérulonéphrite sous-jacente. Dans les cas complexes ou lorsque l'étiologie demeure incertaine, une biopsie rénale est souvent nécessaire pour confirmer le diagnostic, préciser le type de glomérulonéphrite et adapter la prise en charge thérapeutique en conséquence[23].

1.5.1.4 Les maladies génétiques

Les maladies rénales d'origine génétique représentent une cause importante de la maladie rénale chronique (MRC). Parmi elles, la polykystose rénale autosomique dominante (PKRAD) est la plus courante. Cette affection héréditaire se transmet selon un mode autosomique dominant et se manifeste généralement entre la trentaine et la quarantaine [27]. Elle ne se limite pas aux reins, car elle peut entraîner des complications extrarénales, notamment la formation de kystes dans d'autres organes comme le foie et la rate, ainsi que des anomalies vasculaires telles que les anévrismes cérébraux et certaines atteintes des valves cardiaques [28]. Le dépistage familial repose sur l'échographie, permettant d'identifier précocement les kystes rénaux et d'anticiper l'évolution de la maladie [23]. Le syndrome d'Alport constitue une autre pathologie génétique affectant les reins, associée à une surdité neurosensorielle progressive. Son mode de transmission est lié au chromosome X dans la majorité des cas (80 %), tandis que 15 % des formes sont héritées selon un mode autosomique récessif et 5 % selon un mode autosomique dominant. Outre ces maladies bien identifiées, il existe d'autres affections génétiques rares pouvant être à l'origine d'une MRC. Par conséquent, une évaluation approfondie des antécédents familiaux est essentielle pour poser un diagnostic précis et optimiser la prise en charge des patients. Les tests génétiques jouent un rôle crucial dans le diagnostic précoce de ces maladies, permettant une prise en charge personnalisée et un suivi adapté pour ralentir la progression de la maladie et limiter ses complications [23].

1.5.2Les facteurs risques de la maladie rénale chronique

La maladie rénale chronique (MRC) est une pathologie multifactorielle dont le développement et la progression sont influencés par plusieurs facteurs de risque. Ces facteurs peuvent être modifiables (comme l'hypertension et le diabète) ou non modifiables (comme l'âge et les prédispositions génétiques). Une compréhension approfondie de ces facteurs est essentielle pour la prévention, le diagnostic précoce et la prise en charge efficace de la MRC.

1.5.2.1 Le diabète :

Le diabète sucré est l'un des principaux facteurs de risque de la maladie rénale chronique. Environ 30 à 40 % des patients diabétiques développent une néphropathie diabétique, une complication rénale grave qui peut évoluer vers une insuffisance rénale terminale [21]. L'hyperglycémie chronique entraîne une inflammation et une fibrose glomérulaire, altérant progressivement la fonction rénale. Un contrôle strict de la glycémie, associé à une surveillance régulière de la fonction rénale (taux de filtration glomérulaire et albuminurie), est crucial pour prévenir ou ralentir la progression de la MRC chez les patients diabétiques[29].

1.5.2.2 L'Hypertension artérielle :

L'hypertension artérielle est un facteur de risque majeur pour la MRC, à la fois comme cause et comme conséquence de la maladie. Une pression artérielle élevée exerce une surcharge mécanique sur les glomérules rénaux, entraînant leur destruction progressive et une diminution de la fonction rénale [30]. Des études montrent qu'une pression artérielle systolique supérieure à 140 mm Hg augmente significativement le risque de développer une insuffisance rénale terminale[31]. Le contrôle de l'hypertension, notamment par l'utilisation d'inhibiteurs de

l'enzyme de conversion (IEC) ou de bloqueurs des récepteurs de l'angiotensine (BRA), est essentiel pour ralentir la progression de la MRC [32].

1.5.2.3 L'Obésité et ses Effets Nocifs sur la Fonction Rénale

L'obésité est un facteur de risque indépendant pour la MRC, principalement en raison de son association avec le diabète de type 2, l'hypertension et le syndrome métabolique. L'excès de tissu adipeux favorise une inflammation systémique et une résistance à l'insuline, qui contribuent à la détérioration de la fonction rénale [33]. De plus, l'obésité augmente le risque de protéinurie et de glomérulosclérose segmentaire et focale, une lésion rénale spécifique [34]. La perte de poids et une alimentation équilibrée sont des stratégies clés pour réduire ce risque.

1.5.2.4 Le tabagisme

Le tabagisme est un facteur de risque modifiable qui contribue à la progression de la MRC. Les substances toxiques présentes dans la fumée de cigarette endommagent les vaisseaux sanguins rénaux, réduisant le flux sanguin vers les reins et aggravant l'hypertension. Des études ont montré que les fumeurs ont un risque accru de développer une insuffisance rénale terminale par rapport aux non-fumeurs [35] . L'arrêt du tabac est donc une mesure préventive essentielle pour protéger la fonction rénale.

1.5.2.5 L'âge

L'âge est un facteur de risque non modifiable pour la MRC. Avec le vieillissement, la fonction rénale diminue naturellement en raison de la perte progressive de néphrons et de la réduction du débit de filtration glomérulaire [36]. Les personnes âgées sont également plus susceptibles de présenter des comorbidités telles que l'hypertension et le diabète, qui exacerbent le risque de MRC [37].Une surveillance régulière de la fonction rénale est donc recommandée chez les personnes âgées.

1.5.2.6 A Les antécédents familiaux et les prédispositions génétiques :

Les antécédents familiaux de MRC et les prédispositions génétiques jouent un rôle important dans le développement de la maladie. Des affections héréditaires comme la polykystose rénale autosomique dominante (PKRAD) et le syndrome d'Alport sont des causes bien documentées de MRC [27]. Une évaluation des antécédents familiaux et des tests génétiques peuvent aider à identifier les patients à risque et à mettre en place des mesures préventives précoces[38].

1.5.3 Impact du diabète sur la maladie rénale chronique

La maladie rénale diabétique (MRD) comprend la néphropathie diabétique (ND) typique ainsi que d'autres formes d'atteintes rénales [39]. La ND est un syndrome clinique affectant 25 à 40 % des patients diabétiques, caractérisé par une albuminurie (>300 mg/jour, macroalbuminurie A3), confirmée sur au moins deux mesures espacées de trois mois, associée à une diminution du débit de filtration glomérulaire estimé (DFGe) et souvent à une hypertension artérielle. La microalbuminurie (A2, <300 mg/24 h) représente un stade précoce de la ND et est aussi un marqueur de risque de complications microangiopathiques et microangiopathiques, en particulier chez les patients atteints de diabète de type 2. Cependant, de nombreux patients

diabétiques, notamment atteints de DT2, présentent une baisse du DFG sans albuminurie ni rétinopathie [39].

Le diabète est aujourd'hui l'une des principales causes de la maladie rénale chronique (MRC), avec une prévalence atteignant environ 30 à 40 % des patients diabétiques [10.53738]. L'hyperglycémie chronique, caractéristique du diabète, entraîne une cascade de mécanismes pathologiques, notamment une inflammation glomérulaire, un stress oxydatif accru et une activation du système rénine-angiotensine-aldostérone (SRAA). Ces processus contribuent à la fibrose rénale et à la perte progressive de la fonction rénale, conduisant à la néphropathie diabétique [21].En plus de l'hyperglycémie, des facteurs de risque associés tels que l'hypertension artérielle et l'obésité aggravent la progression de la MRC chez les patients diabétiques. L'hypertension, en particulier, augmente la pression intra glomérulaire, accélérant la destruction des néphrons [30].Des biomarqueurs émergents, tels que les marqueurs d'inflammation (comme la TNF-α) et de fibrose (comme le TGF-β), sont à l'étude pour détecter précocement la néphropathie diabétique avant l'apparition de l'albuminurie ou de la baisse du DFGe [40].

Malgré les progrès dans la prise en charge du diabète (contrôle glycémique, surveillance tensionnelle, traitements ciblant le système rénine-angiotensine), l'incidence de l'insuffisance rénale terminale (IRT) chez ces patients reste élevée. La progression de la ND vers l'IRT est une cause majeure du recours à la dialyse, avec plus de 20 % des patients dialysés étant diabétiques [10.53738]. La MRC chez les patients diabétiques représente un fardeau socio-économique important, avec des coûts élevés liés à la dialyse et à la transplantation rénale. En 2020, les dépenses liées à la dialyse pour les patients diabétiques ont dépassé 50 milliards de dollars aux États-Unis [41].

L'augmentation mondiale de la prévalence du diabète, qui pourrait concerner plus d'1,3 milliard de personnes d'ici 2050, accentue ce phénomène [42]-[43]. Environ 24 % des patients diabétiques développent une MRC [21], un chiffre qui peut atteindre 40 % lorsque la prédiction repose à la fois sur la baisse du DFGe et la présence d'albuminurie[44]-[45]. Par ailleurs, 15 % des patients diabétiques présentent déjà une MRC au moment du diagnostic du diabète[46], et cette proportion peut dépasser 43 % après plus de 20 ans d'évolution [47].L'absence de microalbuminurie ne permet pas d'exclure totalement une MRC [48], ce qui renforce la nécessité d'une surveillance rigoureuse et multidimensionnelle.

Enfin, la ND est fortement associée aux maladies cardiovasculaires [49], un risque qui augmente avec la détérioration de la fonction rénale. Ces constatations renforcent l'importance de la prévention et de la surveillance chez les patients diabétiques à risque. Il est donc essentiel de suivre simultanément l'albuminurie et le DFGe pour améliorer la prise en charge et ralentir la progression vers l'IRT. Les inhibiteurs du SGLT2 et les agonistes des récepteurs du GLP-1 ont démontré des bénéfices rénaux significatifs chez les patients diabétiques, réduisant la progression de la MRC et le risque d'insuffisance rénale terminale [50]-[51]-[52].

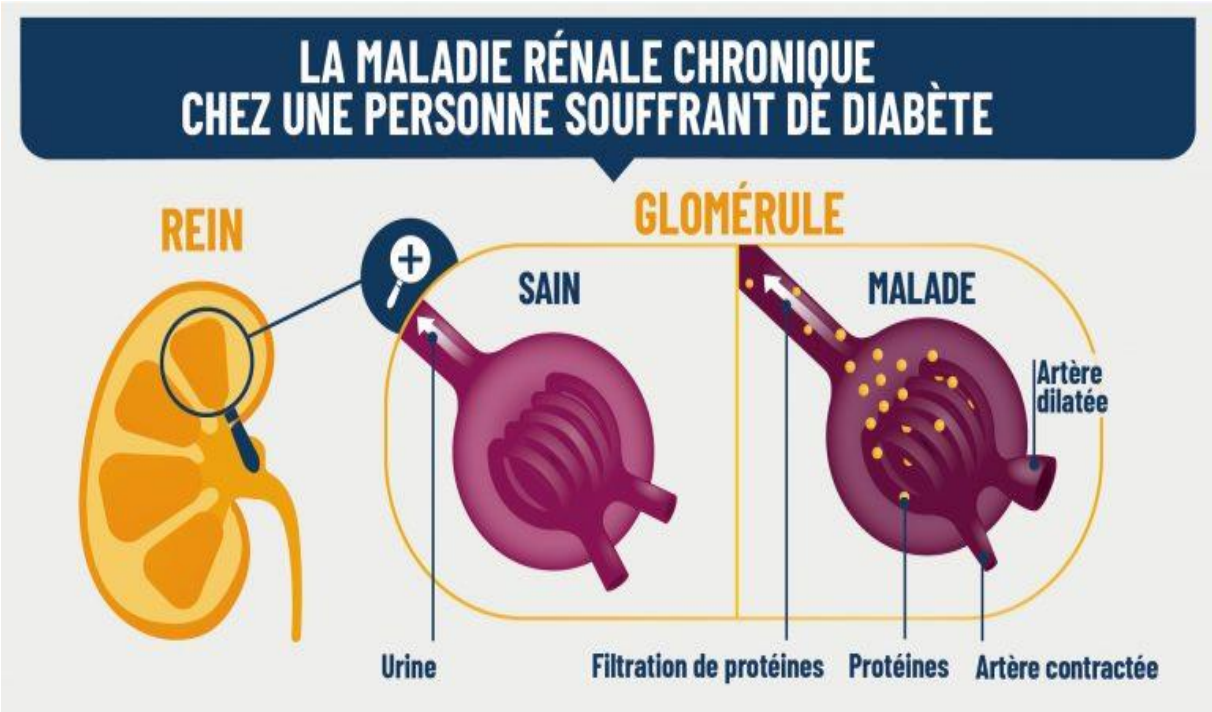


Figure 1.5: Schéma comparatif d'un rein sain et d'un rein atteint de maladie rénale chronique chez un patient diabétique [53]

1.5.4 Importance du dépistage précoce chez les patients diabétiques

Le dépistage précoce de la maladie rénale chronique (MRC) chez les patients diabétiques est essentiel pour prévenir ou ralentir la progression vers l'insuffisance rénale terminale (IRT). La microalbuminurie (30 à 300 mg/jour) et la baisse du débit de filtration glomérulaire estimé (DFGe) sont des marqueurs clés pour identifier les patients à risque [21] . Une prédiction précoce permet une intervention thérapeutique rapide, notamment par le contrôle strict de la glycémie, la gestion de l'hypertension artérielle et l'utilisation d'inhibiteurs du système rénine-angiotensine-aldostérone (SRAA) [54]. Des études ont montré que le dépistage régulier de l'albuminurie et du DFGe réduit significativement le risque de progression de la MRC et améliore le pronostic rénal des patients diabétiques [45] . En outre, l'identification précoce de la MRC permet de prévenir les complications cardiovasculaires, qui sont fréquentes chez les patients diabétiques atteints de néphropathie [49]. Enfin, des biomarqueurs émergents, tels que les marqueurs d'inflammation et de fibrose, sont à l'étude pour améliorer la prédiction précoce de la MRC avant l'apparition de l'albuminurie ou de la baisse du DFGe [40]. Un dépistage systématique et précoce chez les patients diabétiques est une stratégie clé pour réduire le fardeau de la MRC et améliorer la qualité de vie des patients.

1.6 Diagnostique de la MRC

Le diagnostic de la maladie rénale chronique (MRC) repose sur l'évaluation de la fonction rénale, principalement à travers le débit de filtration glomérulaire (DFG) et la présence de marqueurs de lésions rénales, tels que l'albuminurie. La combinaison de ces deux paramètres permet non seulement de poser le diagnostic de MRC, mais aussi de stratifier le risque de complications et de guider la prise en charge thérapeutique [44]-[55].

1.6.1 Biomarqueurs de la fonction rénale

Les biomarqueurs jouent un rôle central dans l'évaluation de la fonction rénale, permettant de diagnostiquer, de surveiller et de prédire la progression de la maladie rénale chronique (MRC). Parmi ces biomarqueurs, la créatinine sérique est le plus couramment utilisé en pratique clinique. La créatinine est un produit de dégradation musculaire qui est filtré par les glomérules rénaux et excrété dans l'urine. Cependant, son utilisation présente des limites importantes. En effet, la concentration de créatinine sérique peut être influencée par des facteurs non rénaux, tels que l'âge, le sexe, la masse musculaire, l'alimentation et certains médicaments. Par exemple, les personnes âgées ou les patients avec une faible masse musculaire peuvent avoir une concentration de créatinine sérique plus basse, ce qui peut conduire à une surestimation du débit de filtration glomérulaire (DFG) [45].

1.6.1.1 Débit de filtration glomérulaire (DFG)

Le DFG est un indicateur clé de la fonction rénale, mesurant le volume de sang filtré par les glomérules rénaux par unité de temps. Il est généralement estimé à partir de la créatinine sérique, en utilisant des équations validées comme la formule CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) ou MDRD (Modification of Diet in Renal Disease). Ces équations prennent en compte des facteurs tels que l'âge, le sexe et l'origine ethnique pour améliorer la précision de l'estimation du DFG. Cependant, comme la créatinine sérique, le DFG peut être affecté par des variations de masse musculaire et d'autres facteurs non rénaux, ce qui souligne la nécessité de compléter son évaluation par d'autres biomarqueurs [56]-[57].

1.6.1.2 L’albuminurie

Un autre biomarqueur essentiel est l'albuminurie, définie par la présence de protéines d'albumine dans l'urine. L'albuminurie reflète une atteinte glomérulaire et est un marqueur précoce de la néphropathie diabétique et d'autres formes de MRC. Elle est également un prédicteur fort de la progression de la MRC et du risque cardiovasculaire. Les patients présentant une albuminurie persistante (≥ 30 mg/g de créatinine) ont un risque accru de développer une insuffisance rénale terminale (IRT) et des complications cardiovasculaires, indépendamment de leur DFG [55]. La surveillance régulière de l'albuminurie est donc recommandée chez les patients à risque, notamment ceux atteints de diabète ou d'hypertension artérielle. En plus de ces biomarqueurs classiques, des biomarqueurs émergents sont à l'étude pour améliorer la précision du diagnostic et du pronostic de la MRC. Parmi eux, la cystatine C, une protéine produite par toutes les cellules nucléées et filtrée par les reins, présente l'avantage d'être moins influencée par la masse musculaire que la créatinine. La cystatine C est particulièrement utile pour estimer le DFG chez les patients présentant des variations de masse musculaire, tels que les personnes âgées ou les patients atteints de maladies chroniques [57]. D'autres biomarqueurs, tels que les marqueurs d'inflammation (par exemple, la TNF-α) et de fibrose rénale, sont également en cours d'évaluation. Ces biomarqueurs pourraient permettre une prédiction plus précoce de la MRC et une meilleure évaluation du risque de progression[40].

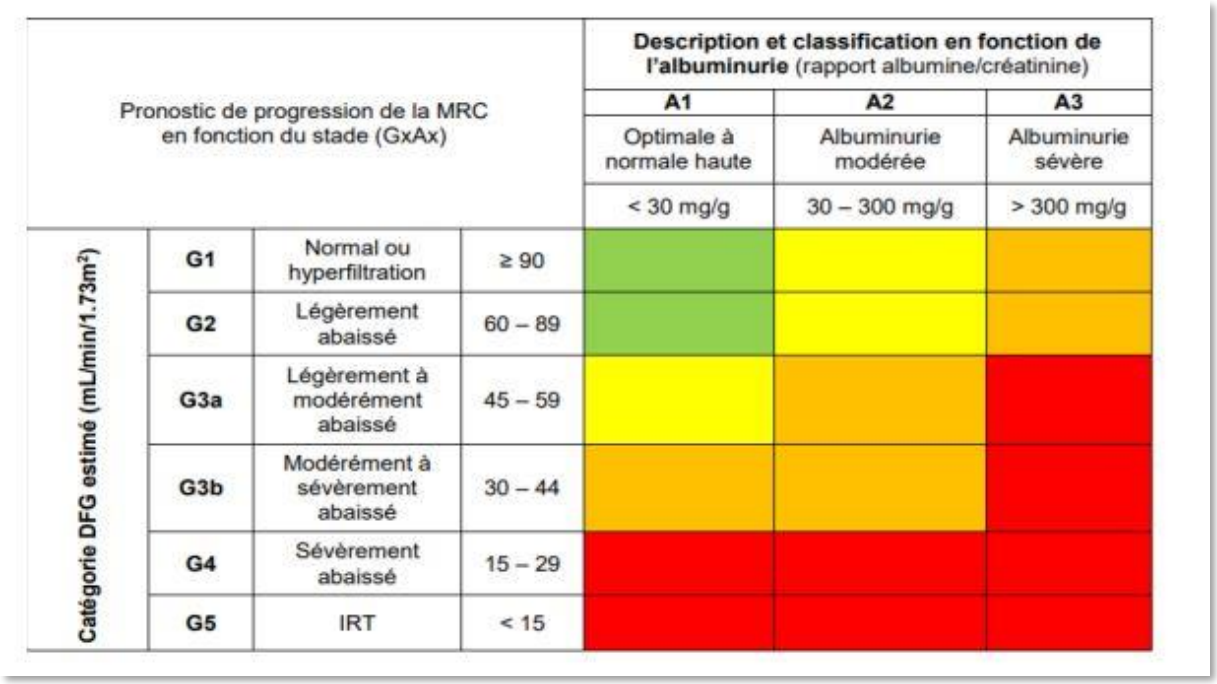
1.6.2 Stade et classification de la maladie rénale chronique

La maladie rénale chronique (MRC) est une affection progressive caractérisée par une diminution durable de la fonction rénale. Elle est classée en cinq stades, déterminés principalement par le débit de filtration glomérulaire (DFG) et la présence de marqueurs de lésions rénales, tels que l'albuminurie. Cette classification permet d'évaluer la gravité de la maladie et d'adapter la prise en charge médicale en conséquence [55].

Tableau 1.1: Classification des stades de la MRC [58]

Stade de la MRC	DFG (mL/min/1,73m2à)	Description
1	≥ 90	Maladie rénale chronique avec DFG normal ou augmenté
2	60-89	Maladie rénale chronique avec DFG légèrement diminué
3	Stade 3 A :45-59	Insuffisance rénale Modéré
	Stade 3 B :30-44	
4	15-29	Insuffisance rénale sévère
5	< 15	Insuffisance rénale terminale

Cette stratification est essentielle pour anticiper les complications potentielles et déterminer les interventions appropriées à chaque stade. Par exemple, aux stades précoces (1 et 2), l'accent est mis sur le traitement des facteurs de risque et la surveillance régulière de la fonction rénale. Aux stades 3 et 4, une prise en charge plus intensive est nécessaire pour ralentir la progression de la maladie et gérer les complications métaboliques. Au stade 5, une thérapie de suppléance rénale, telle que la dialyse ou la transplantation, devient souvent indispensable [59].Il est important de noter que la progression de la MRC n'est pas inéluctable. Des interventions précoces, telles que le contrôle strict de la pression artérielle, la gestion de l'albuminurie et l'adoption de mesures hygiéno-diététiques, peuvent ralentir, voire stopper, l'évolution de la maladie.



Pronostic de progression de la MRC en fonction du stade (GxAx)				Description et classification en fonction de l'albuminurie (rapport albumine/créatinine)		
				A1	A2	A3
				Optimale à normale haute	Albuminurie modérée	Albuminurie sévère
				< 30 mg/g	30 – 300 mg/g	> 300 mg/g
Catégorie DFG estimé (mL/min/1.73m²)	G1	Normal ou hyperfiltration	≥ 90			
	G2	Légèrement abaissé	60 – 89			
	G3a	Légèrement à modérément abaissé	45 – 59			
	G3b	Modérément à sévèrement abaissé	30 – 44			
	G4	Sévèrement abaissé	15 – 29			
	G5	IRT	< 15			

Figure 1.6: Pronostic de la maladie rénale chronique basé sur les catégories de débit de filtration glomérulaire et d'albuminurie [55]

Les patients classés dans la catégorie à faible risque (case verte) doivent bénéficier d'un suivi annuel. Ceux appartenant à la catégorie à risque modéré (case jaune) nécessitent un contrôle semestriel. Pour les patients à risque élevé (case orange), un suivi trimestriel est recommandé. Enfin, les patients à très haut risque (case rouge), en particulier ceux proches du stade de dialyse, doivent être suivis au minimum quatre fois par an, voire mensuellement si leur état clinique l'exige.

1.6.3 Méthodes de calcul du débit de filtration glomérulaire (DFG)

Le débit de filtration glomérulaire (DFG) est le paramètre clé pour évaluer la fonction rénale. Plusieurs formules ont été développées pour estimer le DFG à partir de la créatinine sérique et d'autres variables.

1.6.3.1 Formule CKD-EPI (chronic kidney disease- Epidemiology collaboration)

La formule CKD-EPI est actuellement la méthode de référence pour estimer le DFG. Elle a été développée pour améliorer la précision de l'estimation du DFG en (mL/min/1,73m2), en particulier chez les patients avec un DFG modérément réduit. La formule CKD-EPI prend en compte la créatinine sérique, l'âge, le sexe et l'origine ethnique [56]-[60].

$$DFG = 141 \times \min\left(\frac{Scr}{k}\right)^{\alpha} \times max\left(\frac{Scr}{k}\right)^{-1,209} \times 0,993^{age} \times 1,018\ (si\ femme) \times 1,159\ (si\ sujet\ noir)$$

$$Avec\ k = 62\ pour\ les\ femmes\ et\ k = 80\ pour\ les\ hommes\ et\colon\ -0,329\ pour\ les\ femmes\ et$$

$$-0,411\ pour\ les\ hommes; a\colon\ -0,329\ pour\ les\ femmes\ et\ -0,411\ pour\ les\ hommes$$

1.6.3.2 Formule Cockroft et Gault (CG)

La formule de Cockroft et Gault est l'une des premières méthodes utilisées pour estimer le DFG. Elle prend en compte la créatinine sérique, l'âge, le poids et le sexe. Cependant, elle surestime souvent le DFG chez les patients obèses ou âgés, et elle n'est pas ajustée à la surface corporelle, ce qui limite son utilisation dans la pratique clinique actuelle [61].

$$ClCr\left(\frac{mL}{min}\right) = k * \left(\frac{(140 - age) * poids(kg)}{Créatininémie}\right) \; avec: k = 1.04 \; pour \; les \; femmes \; et \; k = 1.23 \; pour \; les \; hommes$$

1.6.3.3 Formule MDRD (Modification of Diet in Renal Disease Study)

La formule MDRD simplifiée, introduite en 1999, estime le DFG en utilisant la créatinine sérique, l'âge, le sexe et l'origine ethnique. Elle a amélioré la précision de l'estimation du DFG par rapport à la formule de Cockcroft et Gault, notamment chez les patients atteints de MRC. Cependant, elle tend à sous-estimer le DFG chez les individus avec une fonction rénale normale ou légèrement diminuée, ce qui a conduit au développement de la formule CKD-EPI pour une meilleure précision dans ces plages de DFG[62].

$$DFG = 186 \times (créatinémie)^{-1,154} \times (age)^{-0,203} \times 1 \; (si \; homme) \times 0,742 \; (si \; femme) \times 1,21 \; si \; sujet \; noir$$

1.7 Dépistage de la maladie rénale chronique

Le dépistage de la maladie rénale chronique (MRC) est une étape cruciale pour identifier précocement les patients à risque et ralentir la progression de la maladie. Il repose principalement sur l'évaluation de l'excrétion urinaire d'albumine et la mesure de la fonction rénale à l'aide du débit de filtration glomérulaire (DFG). Une prédiction précoce permet d'instaurer des mesures préventives et thérapeutiques adaptées afin de limiter les complications et d'améliorer la qualité de vie des patients.

1.7.1 Excrétion urinaire d'albumine

La microalbuminurie est définie par une excrétion urinaire d'albumine comprise entre 30 et 300 mg/24h ou entre 30 et 300 µg/mg de créatinine sur un échantillon d'urine isolé, est un indicateur d'atteinte glomérulaire. Elle joue un rôle clé dans le diagnostic, le suivi thérapeutique et l'évaluation du pronostic, notamment chez les patients atteints de diabète de type 1 ou 2[63]-[55].Selon les recommandations de la Haute Autorité de Santé (HAS) [64], les patients présentant un risque accru de maladie rénale chronique (MRC) doivent bénéficier d'une évaluation du débit de filtration glomérulaire (DFG) ainsi que d'un dépistage des marqueurs d'atteinte rénale. La présence d'albuminurie, même avec un DFG normal, permet de poser un diagnostic précoce de MRC, rendant ce dépistage essentiel [63].

1.7.2 Physiopathologie

L'excrétion de l'albumine par les reins est influencée par plusieurs facteurs, notamment la taille et la charge de la molécule, la pression artérielle et la réabsorption tubulaire. Normalement, une

faible quantité d'albumine traverse les glomérules et est presque totalement réabsorbée. Une présence anormale d'albumine dans l'urine peut être liée à une altération de la structure glomérulaire, à l'hypertension artérielle ou à des variations des résistances vasculaires, comme celles observées dans le diabète en raison de l'hyperglycémie [65].

Dans le diabète de type 1, la microalbuminurie est un indicateur précoce d'atteinte rénale, reflétant des lésions microvasculaires. Dans le diabète de type 2, elle est principalement associée à un risque cardiovasculaire accru et, dans une moindre mesure, à une atteinte rénale. Chez les patients hypertendus, elle témoigne d'une altération endothéliale et constitue un marqueur de risque cardiovasculaire [63].

L'évolution physiopathologique implique une atteinte progressive des petits vaisseaux rénaux. L'hyperglycémie chronique induit une glycosylation des protéines glomérulaires, favorisant l'expansion des cellules mésangiales et des lésions endothéliales, entraînant un épaississement de la membrane basale glomérulaire. Ce processus conduit à la glomérulosclérose, à l'hyalinose des artérioles et à l'artériosclérose [21].

L'évolution de la néphropathie diabétique est progressive :

- **Stade précoce** : augmentation transitoire du DFG et de la taille des reins.
- **Stade 2** : phase asymptomatique qui peut durer de 2 à 5 ans selon le contrôle glycémique.
- **Stade 3** : apparition de la microalbuminurie (30-300 mg/24h) et possible élévation de la pression artérielle, sur une période de 5 à 10 ans.
- **Stade 4** : aggravation de la protéinurie (>300 mg/24h), avec un syndrome néphrotique dans 10 % des cas (>3 g/24h) et une hypertension associée dans 75 % des cas, évoluant sur 10 à 20 ans.
- **Stade 5** : insuffisance rénale terminale nécessitant une prise en charge par dialyse ou transplantation après environ 20 ans d'évolution.

La néphropathie diabétique reste longtemps silencieuse, mais en l'absence de traitement, elle peut conduire à une hypertension artérielle et à des œdèmes des membres inférieurs [63].

1.7.3 Indications et réalisation du dosage

Le dépistage de la microalbuminurie est recommandé pour les patients à risque de néphropathie, en particulier les diabétiques, hypertendus et ceux présentant une insuffisance cardiaque, une maladie auto-immune ou une dysglobulinémie monoclonale. Différentes méthodes permettent de détecter l'albuminurie :

1. Bandelette urinaire : détecte l'albumine uniquement si elle dépasse 50-100 mg/L.
2. Recueil des urines sur 24 heures : référence historique, mais nécessitant une bonne adhésion du patient. Un résultat supérieur à 30 mg/24h est considéré comme positif.
3. Rapport minuté : technique mesurant l'excrétion urinaire par minute, mais exigeant des conditions strictes de réalisation.

4. Dosage sur échantillon isolé : méthode simple, mais influencée par l'hydratation du patient. Les urines du matin, plus concentrées, sont préférables.

5. Rapport albuminurie/créatininurie : recommandé par l'HAS [64] et largement utilisé dans les pays anglo-saxons. Ce rapport exprimé en mg/mmol ou mg/g doit être inférieur à 2 mg/mmol ou 30 mg/g.

Il est essentiel de choisir le bon moment pour réaliser ces tests afin d'éviter des faux positifs liés à des facteurs transitoires tels que l'exercice physique, les menstruations, une dyslipidémie sévère, une infection, une décompensation aiguë du diabète ou une insuffisance cardiaque [63].

1.8 Conclusion

Ce premier chapitre a permis de jeter les bases de la compréhension de la maladie rénale chronique (MRC), en explorant à la fois l'anatomie et la physiologie du rein et en définissant précisément la MRC, ses causes et ses facteurs de risque majeurs, notamment le rôle prépondérant du diabète. L'analyse de la prévalence et des tendances, tant à l'échelle mondiale qu'en Algérie, souligne l'importance cruciale du dépistage précoce et d'une gestion adaptée afin de freiner l'évolution de cette pathologie silencieuse. Les différents biomarqueurs et méthodes de diagnostic, tels que le débit de filtration glomérulaire et l'albuminurie, se révèlent être des outils indispensables pour évaluer la fonction rénale et établir la gravité de la maladie. En mettant en lumière ces divers aspects, ce chapitre ouvre la voie à une approche intégrée et proactive dans la prise en charge de la MRC, rappelant que la prévention et le suivi régulier restent les clés pour améliorer la qualité de vie des patients et limiter l'impact de cette maladie sur la santé publique. Dans le chapitre suivant, nous nous pencherons sur l'intégration de l'intelligence artificielle dans la prédiction de la maladie rénale chronique, illustrant ainsi l'importance de l'innovation technologique dans le domaine de la santé.

Chapitre 2 :

L’intelligence artificielle (IA)

2.1 Introduction

Dans un contexte de digitalisation et d’innovations technologiques accélérées, l'intelligence artificielle (IA) s’impose comme l’un des domaines les plus prometteurs et transformateurs du XXIe siècle. Elle révolutionne des secteurs aussi variés que l'industrie, la finance, les transports et, de manière particulièrement significative, la santé. L'IA se définit comme la capacité des machines à imiter l'intelligence humaine en apprenant, en raisonnant et en s'adaptant à partir de données. Elle englobe un large éventail de technologies, allant des systèmes experts traditionnels aux approches modernes d'apprentissage automatique (Machine Learning) et d'apprentissage profond (Deep Learning), permettant ainsi de résoudre des problèmes complexes, d'automatiser des tâches répétitives et d'extraire des informations précieuses à partir de vastes quantités de données.

Dans le domaine médical, l'IA ouvre de nouvelles perspectives pour améliorer le diagnostic, le traitement et la prévention des maladies. Grâce à sa capacité à analyser des données massives et hétérogènes telles que des images médicales, des dossiers patients ou des données génomiques, l’IA offre des outils puissants pour identifier des schémas invisibles à l'œil nu, prédire l'évolution des pathologies et personnaliser les soins. Par exemple, des algorithmes d'IA sont déjà utilisés pour détecter des cancers à partir d'images radiologiques, prédire le risque de maladies cardiovasculaires ou optimiser les protocoles thérapeutiques pour des patients atteints de maladies chroniques.

Cependant, malgré ses immenses promesses, l'IA soulève également d'importants défis, notamment en termes de qualité des données, de biais algorithmiques, de confidentialité et d'éthique. Ces enjeux sont particulièrement critiques dans le domaine de la santé, où les décisions fondées sur l'IA peuvent impacter directement la vie des patients. Une compréhension approfondie des principes, des méthodes et des limites de l'IA s'avère donc indispensable pour exploiter pleinement son potentiel tout en garantissant une utilisation responsable et efficace.

Dans ce chapitre, nous explorerons les fondements de l'intelligence artificielle en nous concentrant sur les techniques d'apprentissage automatique et d'apprentissage profond. Nous aborderons également les étapes clés telles que le prétraitement des données, la sélection des caractéristiques et l'évaluation des modèles, qui sont essentielles pour le développement de systèmes d'IA robustes et fiables. Enfin, nous examinerons les applications concrètes de l'IA dans le domaine médical, en mettant en lumière son rôle dans l'amélioration des soins et en discutant des défis ainsi que des perspectives futures. Ce chapitre vise ainsi à fournir une base solide pour comprendre comment l'IA peut contribuer à relever les défis complexes de la médecine moderne, notamment dans la prise en charge de maladies chroniques comme la maladie rénale chronique (MRC).

2.2 L’intelligence artificielle (IA)

Au cours de la dernière décennie, l'intelligence artificielle (IA) [66] s'est imposée comme une technologie révolutionnaire, transformant de nombreux secteurs grâce à des techniques avancées telles que l'apprentissage automatique (ML), le traitement du langage naturel et la vision par ordinateur. Ces technologies, qui simulent des fonctions cognitives humaines

notamment la pensée, l'adaptation et l'apprentissage en profondeur [67]-[68], permettent d'automatiser des tâches complexes et d'améliorer la prise de décision dans divers domaines, tels que la logistique, l'industrie manufacturière, les transports, la santé, l'éducation et la gouvernance publique. En optimisant l'efficacité opérationnelle et en réduisant les coûts, l'IA contribue significativement à la croissance économique et au bien-être social [69].

Dans le domaine médical, l'IA occupe une place essentielle depuis les premières tentatives d'assistance informatique aux diagnostics dans les années 1950 [70]-[71]. L'évolution rapide de cette technologie, soutenue par une puissance de calcul accrue et l'explosion des données numériques [72], a permis de transformer la pratique médicale. Aujourd'hui, l'IA intègre et analyse d'importants volumes de données issues de diverses sources, telles que les images médicales, dossiers patients, données génomiques, etc. Afin d'identifier des patterns souvent invisibles à l'œil humain et ainsi améliorer la précision diagnostique ainsi que la prise de décision clinique [73].

Par ailleurs, l'IA se définit comme l'étude et la conception d'agents intelligents capables de percevoir leur environnement et d'optimiser leurs décisions [74]. Ces avancées ont conduit au développement de modèles d'apprentissage automatique capables d'intégrer divers facteurs de risque tels que l'âge, le sexe, les antécédents familiaux et les habitudes de vie pour générer des prédictions personnalisées dans le domaine de la maladie rénale chronique (MRC). Ces modèles assistent les cliniciens en fournissant des diagnostics précis, en évaluant le pronostic et en orientant les choix thérapeutiques, tout en ouvrant de nouvelles perspectives pour la prévention et la gestion de cette pathologie complexe [75].

2.3 Apprentissage automatique

L'apprentissage automatique (ML) est un domaine multidisciplinaire ancré dans les statistiques, l'algèbre, le traitement des données et l'analyse des connaissances, ce qui rend sa définition unique et complexe [76]. Il s'agit d'une branche spécifique de l'intelligence artificielle (IA) qui permet aux machines d'extraire des informations à partir de données d'entraînement sans intervention explicite, en identifiant des schémas et en adaptant ses décisions en conséquence [77].

Le ML repose sur la construction de modèles statistiques capables de prédire des résultats ou de classer des observations à partir de données observées. Les méthodes d'apprentissage supervisé, par exemple, consistent à entraîner un modèle sur un ensembles de données dont les résultats attendus sont déjà connus (gold standard). Ces modèles, souvent basés sur des techniques de régression pénalisée, sont ensuite appliqués à de nouvelles données pour générer des prédictions, avec une évaluation rigoureuse de leur performance [78]-[79].

Arthur Samuel (1959) définit l'apprentissage automatique comme la capacité des ordinateurs à apprendre sans être explicitement programmés. Cela implique l'utilisation d'attributs distinctifs pour identifier des modèles exploitables dans des situations données, permettant ainsi aux systèmes d'affiner leurs prédictions et d'améliorer leur précision avec le temps [80].

Le secteur de la santé est l'un des domaines où le ML présente un fort potentiel. Il est déjà utilisé dans l'analyse de données médicales, les diagnostics assistés par IA et l'optimisation des

traitements. En appliquant des techniques avancées d'analyse prédictive, le ML peut aider à évaluer un large éventail de données cliniques, réduire les coûts et accélérer le processus des essais médicaux [81]-[82]. Son intégration dans les systèmes de santé, notamment à travers les dossiers médicaux électroniques (DME) et les outils d'analyse big data, favorise une prise de décision plus intelligente et une automatisation améliorée des soins [83]-[84]-[85].

De plus, le ML joue un rôle clé dans la personnalisation des soins de santé. Grâce à l'analyse de données biomédicales et génomiques, il permet de proposer des thérapies adaptées aux caractéristiques spécifiques de chaque patient, optimisant ainsi l'efficacité des traitements et améliorant la qualité des soins [86]-[87]-[88].

2.3.1 Apprentissage supervisé

L'apprentissage supervisé est une méthode fondée sur l'utilisation d'un ensembles de données étiquetées, où chaque exemple inclut des variables explicatives (entrées) et une réponse attendue (étiquette) [89]. L'objectif est de construire un modèle prédictif capable d'associer correctement de nouvelles entrées à leurs sorties correspondantes. Le processus consiste à entraîner l'algorithme sur un sous-ensembles des données, puis à évaluer sa performance sur un ensembles de test indépendant, en minimisant les erreurs de prédiction [89]-[90].

Ce type d'apprentissage repose sur deux grandes catégories de problèmes :

- **La classification**, qui vise à affecter une observation à une classe prédéfinie (valeurs discrètes),
- **La régression**, qui consiste à estimer une valeur continue à partir de données historiques.

Comme l'indiquent [91] et [92], l'approche supervisée peut être vue comme une relation entre un "enseignant" et l'algorithme, où des paires d'exemples (X, Y) permettent d'apprendre une fonction cible capable de généraliser les prédictions à de nouvelles situations.

Plusieurs algorithmes sont classiquement utilisés en apprentissage supervisé, chacun ayant ses forces et ses limites [89]-[90]-[93] :

- **K-NN (k-nearest neighbors)** : méthode non paramétrique basée sur la proximité des données.
- **Régression linéaire** : technique statistique pour modéliser la relation linéaire entre une variable dépendante et une ou plusieurs variables indépendantes.
- **Régression logistique** : utilisée principalement pour la classification binaire, elle modélise la probabilité d'appartenance à une classe.
- **SVM (Support Vector Machines)** : algorithme efficace dans les espaces de grande dimension, qui cherche un hyperplan séparateur optimal.
- **Arbres de décision (Decision Trees)** : modèle de type arborescent facilitant l'interprétation des décisions.

- **Forêts aléatoires (Random Forest)** : ensembles d'arbres de décision pour améliorer la robustesse et réduire le surapprentissage.

- **Réseaux de neurones artificiels (Artificial Neural Networks)** : modèles inspirés du cerveau humain, organisés en couches denses (entrée, cachées, sortie) pour traiter des structures complexes de données.

- **Naïve Bayes** : basé sur le théorème de Bayes, ce modèle probabiliste suppose l'indépendance conditionnelle entre les variables.

L'apprentissage supervisé trouve des applications variées dans plusieurs domaines :

1. **En médecine**, pour la prédiction de maladies et la prédiction des issues thérapeutiques [94]–[95],

2. **En finance**, pour l'identification de fraudes ou l'anticipation des fluctuations de marché,

3. **En reconnaissance vocale**, pour améliorer la transcription automatique de la parole,

4. **En vision par ordinateur**, notamment pour la segmentation d'images selon des caractéristiques extraites.

   Malgré ses avantages, l'une des limites majeures de cette approche réside dans la nécessité de disposer de données étiquetées, dont la collecte peut être coûteuse et fastidieuse [90].

2.3.2 Apprentissage non supervisé

L'apprentissage non supervisé est une approche qui s'applique à des problèmes nécessitant l'analyse de grandes quantités de données non étiquetées. L'objectif principal est de découvrir des structures ou des modèles cachés dans ces données, sans l'intervention humaine pour guider le processus. Les tâches principales de l'apprentissage non supervisé comprennent le clustering, l'association et la réduction de dimensionnalité [89].

- **Clustering :** Il s'agit de regrouper des données similaires en différents clusters. Cette technique est utile pour la compression de données, la segmentation d'images et la reconnaissance de motifs dans des ensembles de données complexes [96].
- **Association :** Cette méthode sert à identifier les relations ou les règles entre les variables d'un ensembles de données. Par exemple, dans le commerce de détail, on peut découvrir les relations entre les articles achetés ensembles [96].
- **Réduction de dimensionnalité :** Cette technique réduit la taille de l'ensembles de données tout en conservant les informations cruciales. Cela est particulièrement utile lorsqu'il y a un grand nombre de caractéristiques, permettant une gestion plus facile des données complexes [96].

Parmi les algorithmes les plus courants d'apprentissage non supervisé, on retrouve le K-moyennes (k-means), qui partitionne les données en un nombre prédéfini de clusters [97], et le clustering hiérarchique (HCA) , qui crée une hiérarchie de clusters sans besoin de définir leur nombre [96]. L'Analyse en Composantes Principales (PCA) est également fréquemment

utilisée pour la réduction de la dimensionnalité, permettant de conserver les informations les plus importantes tout en éliminant les variables redondantes. L'algorithme Apriori, quant à lui, sert à l'exploration de données et à la recherche de règles d'association [96].

En termes d'avantages et d'inconvénients, l'apprentissage non supervisé présente plusieurs points positifs : il permet de découvrir des modèles inattendus et des relations cachées dans les données, tout en exigeant peu d'étiquetage manuel des données, ce qui récupère du temps et des ressources [98]. Cependant, il présente également des défis, comme la difficulté d'interpréter les résultats, la gestion des données de grande taille et la détermination du nombre optimal de clusters à utiliser. En outre, les algorithmes d'apprentissage non supervisé ne peuvent pas toujours évaluer la qualité des résultats de la même manière que les algorithmes supervisés [98].

Les applications de l'apprentissage non supervisé sont nombreuses et variées. Par exemple, dans le clustering, on peut utiliser cette technique pour segmenter des clients ou regrouper des documents en fonction de leur contenu. Dans la réduction de la dimensionnalité, elle est appliquée dans la compression d'images ou la détection d'anomalies. L'analyse de l'expression génétique, ainsi que la détection de fraudes ou d'intrusions, sont également des exemples d'applications où l'apprentissage non supervisé joue un rôle clé. En outre, il est également utilisé dans des systèmes de recommandation, la création de nouveaux points de données (modèles génératifs) et la découverte de règles d'association dans des ensembles de données [99]–[100].

2.3.3 Apprentissage par renforcement

L'apprentissage par renforcement est une approche où un système est constitué d'agents capables d'observer leur environnement, d'effectuer des actions et d'obtenir des récompenses ou des pénalités en fonction de leurs choix [96]. L'objectif est de permettre à l'agent d'apprendre de manière autonome et d'ajuster ses actions pour maximiser les récompenses sur le long terme. En d'autres termes, l'agent cherche à déterminer la stratégie optimale, également appelée politique, pour atteindre ses objectifs de manière efficace dans un environnement donné. Ce type d'apprentissage est particulièrement adapté aux situations où l'agent doit naviguer dans un environnement dynamique, en apprenant de ses expériences passées et en s'adaptant aux nouvelles informations disponibles.

Dans un cadre de machine learning, l'apprentissage par renforcement permet à l'agent de prendre des décisions séquentielles, en agissant dans un environnement complexe et potentiellement imprévisible [101]. L'agent utilise une méthode de trial and error (essais et erreurs), où chaque action entreprise entraîne soit une récompense soit une pénalité, ce qui lui permet d'ajuster son comportement au fur et à mesure. Cette approche est souvent comparée à un processus de compétition, où l'intelligence artificielle explore différentes stratégies pour accomplir des tâches et maximiser son efficacité.

2.4 Les principales phases du processus de machine learning :

Le processus d'apprentissage automatique peut être divisé en sept étapes principales, comme illustré dans la Figure 7. Afin de mieux en comprendre l'utilité, nous illustrerons chaque phase par des exemples concrets permettant de saisir la logique globale d'un pipeline de ML.

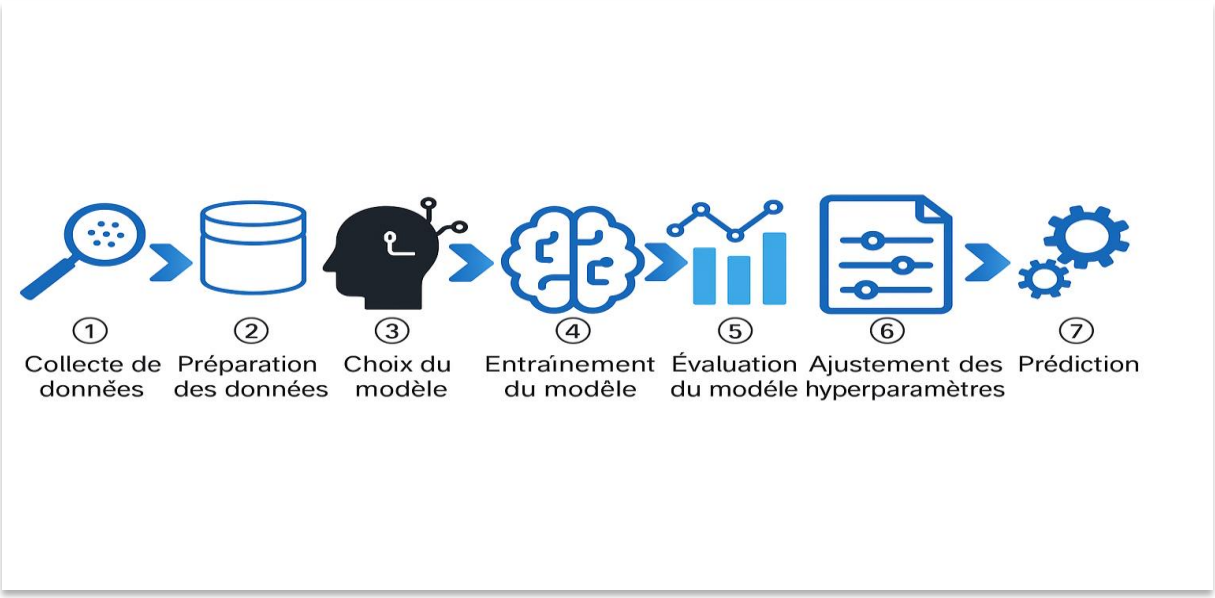


Figure 2.1: Le processus de Machine Learning

2.4.1 Collecte des données

Cette phase constitue le socle fondamental de l'apprentissage automatique. La qualité, la quantité et la pertinence des données collectées influencent directement la performance du modèle. Il est crucial d'éviter des erreurs telles que la sélection de caractéristiques inadéquates ou des restrictions excessives dans la structure des données, sous peine de compromettre totalement l'efficacité du modèle.

2.4.2 Préparation des données

Une fois les données collectées, elles doivent être préparées pour être utilisées dans le processus d'entraînement. Cela inclut :

- Nettoyage des données : suppression des doublons, gestion des valeurs manquantes (par suppression, imputation simple, imputation multiple ou modélisation prédictive) [102],
- Conversion de types et correction d'erreurs,
- Exploration et visualisation : permettant de détecter des relations utiles et des déséquilibres éventuels,
- Normalisation : transformation des données pour les adapter aux algorithmes, par exemple via :

– La normalisation Min-Max, mappant les valeurs dans une plage [0,1] ou [-1,1],
– Ou la normalisation Z-score, basée sur la moyenne et l'écart-type des données.

Division du dataset en deux sous-ensembles : 80 % pour l'entraînement, 20 % pour l'évaluation, afin d'éviter toute forme de surapprentissage.

Cette phase, appelée prétraitement des données, est essentielle pour garantir des prédictions fiables et robustes [102].

2.4.3 Choix du modèle

Une fois les données préparées, il est nécessaire de sélectionner un modèle adapté. Chaque famille de modèles (régression, arbres de décision, réseaux de neurones, etc.) possède ses forces selon le type de données ou de tâche : par exemple, certains modèles sont plus efficaces pour l'analyse de texte, d'autres pour le traitement d'images.

2.4.4 Entraînement du modèle

L'entraînement consiste à ajuster les paramètres du modèle à partir des données d'entraînement. Le modèle commence avec des poids aléatoires, puis apprend à ajuster ses prédictions en comparant ses sorties à la vérité terrain (labels). Ce processus itératif, souvent basé sur des algorithmes comme la descente de gradient, permet au modèle d'optimiser ses performances, tout comme un enfant apprend à faire du vélo par essais et erreurs.

2.4.5 Évaluation du modèle

Le modèle est ensuite testé sur un ensembles de données qu'il n'a jamais vu (données de validation ou de test) pour vérifier s'il généralise bien. Des métriques telles que l'exactitude, la précision, le rappel ou l'AUC sont souvent utilisées. Si les résultats ne sont pas satisfaisants, il peut être nécessaire de revenir aux étapes précédentes pour améliorer la qualité des données ou ajuster le modèle.

2.4.6 Ajustement des hyperparamètres

Les hyperparamètres (taux d'apprentissage, profondeur des arbres, nombre de couches dans un réseau, etc.) ne sont pas appris automatiquement, mais doivent être réglés manuellement ou via des techniques d'optimisation (GridSearch, RandomSearch, etc.). Leur réglage peut grandement influencer la précision et le temps d'entraînement du modèle.

2.4.7 Prédiction

Une fois le modèle entraîné et validé, il est prêt à être utilisé sur de nouvelles données. Il peut alors fournir des prédictions ou des classifications selon le problème traité, ce qui représente l'objectif final de tout processus d'apprentissage automatique.

2.5 L'apprentissage profond :

L'apprentissage profond, ou Deep Learning (DL), constitue une branche spécifique de l'apprentissage automatique, qui repose sur des algorithmes inspirés de la structure et du fonctionnement du cerveau humain, en particulier des réseaux de neurones artificiels. Le principe de l'apprentissage profond est basé sur l'utilisation de réseaux neuronaux multicouches, dont la profondeur permet de traiter des données complexes et volumineuses. Chaque couche dans un réseau de neurones profond sert à extraire des caractéristiques de plus en plus abstraites à partir des données brutes, ce qui permet au modèle d'effectuer des prédictions de plus en plus précises à mesure que l'information passe à travers ces couches [103]-[104].

L'un des points essentiels du deep learning est la capacité des réseaux neuronaux à apprendre directement des données, sans avoir besoin de caractéristiques manuellement extraites par un expert. Ce processus est particulièrement efficace pour des tâches complexes comme la reconnaissance d'images, la classification de textes ou la transcription de la parole. L'augmentation de la performance des modèles de deep learning est en grande partie attribuable à l'amélioration des algorithmes d'entraînement, comme la rétropropagation, et à l'utilisation de grandes quantités de données étiquetées, combinées avec des architectures de plus en plus sophistiquées [104].

Les réseaux de neurones profonds sont capables de traiter de vastes ensembles de données et sont utilisés dans des domaines variés tels que la reconnaissance vocale [105], l'analyse d'images [106], la surveillance de la santé [107] et la détection d'objets [108]. Ces architectures ont permis des avancées significatives dans des applications telles que la vision par ordinateur, la détection de maladies, et la modélisation de phénomènes complexes dans des domaines comme la biologie et les sciences de la santé [109]-[110].

2.6 Différence entre Machine Learning et Deep learning

Selon la nature de la problématique, le volume de données disponibles et les ressources informatiques, il peut être plus pertinent d'opter pour l'apprentissage automatique ou pour l'apprentissage profond. Alors que le Deep Learning repose sur des réseaux de neurones profonds nécessitant de grandes quantités de données et des capacités de calcul élevées, le Machine Learning englobe des méthodes plus classiques pouvant être efficaces même avec moins de données. Le tableau ci-dessous présente une comparaison entre ces deux approches[111].

Tableau 2.1: Tableau comparatif entre la machine learning et le deep learning

Critère	Apprentissage Automatique (Machine Learning)	Apprentissage Profond (Deep Learning)
Définition	Domaine général de l'apprentissage à partir des données	Sous-catégorie de l'apprentissage automatique basée sur les réseaux de neurones profonds
Architecture	Utilise divers algorithmes comme SVM, arbres de décision, etc.	Basée sur des réseaux de neurones à plusieurs couches
Représentation	Caractéristiques extraites manuellement	Apprentissage hiérarchique automatique des caractéristiques
Données requises	Peut fonctionner avec peu de données étiquetées	Nécessite de grandes quantités de données étiquetées
Performance	Bonne performance si les caractéristiques sont bien choisies	Très bonnes performances sur des tâches complexes

Interprétabilité	Modèles souvent plus transparents	Faible interprétabilité due à la complexité des modèles
Exigences matérielles	Fonctionne sur des CPU classiques	Nécessite des GPU pour un apprentissage efficace
Domaines d'application	Analyse prédictive, classification, régression, etc.	Vision par ordinateur, NLP, reconnaissance vocale, etc.

2.7 Applications du machine learning et Deep learning dans le domaine de la santé (médecine) :

L'intelligence artificielle, en particulier le Machine Learning (ML) et le Deep Learning (DL), transforme progressivement le paysage médical. Ces technologies permettent d'automatiser et d'optimiser plusieurs étapes du processus de soins, allant de la prévention au traitement personnalisé [112],[113],[114],[115].

L'une des principales applications du ML est la médecine de précision, qui vise avant tout à prédire l'apparition de maladies en fonction des caractéristiques cliniques et biologiques du patient, permettant ainsi une détection précoce et des interventions plus ciblées. Les algorithmes d'apprentissage automatique interviennent dans toutes les phases de la prise en charge médicale : prévention, dépistage, diagnostic, traitement et suivi.

Parmi les cas d'usage notables :

- **Diagnostic précoce et prédiction des maladies** : Grâce à l'analyse automatisée des données cliniques, le ML permet d'identifier les signes précoces de maladies complexes, améliorant ainsi la rapidité et la précision des diagnostics tout en précisant les coûts[112].
- **Découverte et personnalisation des médicaments** : Le ML est utilisé pour identifier de nouvelles molécules thérapeutiques, prédire leur efficacité et adapter les traitements à chaque patient, accélérant ainsi la recherche pharmaceutique.
- **Chirurgie assistée par l'IA** : Des algorithmes intelligents soutiennent les chirurgiens dans l'identification anatomique précise et l'exécution d'actes chirurgicaux robotisés.
- **Analyse et interprétation des données médicales** : L'apprentissage automatique aide à comprendre les schémas de progression des maladies, à anticiper les complications et à proposer des mesures préventives adaptées.
- **Médecine préventive et dépistage ciblé** : En évaluant les risques individuels à partir de données cliniques, les modèles prédictifs permettent aux professionnels de santé d'identifier les individus à haut risque et de recommander des interventions préventives pertinentes[112],[113],[114],[116].

Au cours des dernières années, l'intégration croissante de l'IA en médecine a permis des avancées significatives vers une médecine personnalisée, dynamique et fondée sur les données. Les systèmes intelligents contribuent non seulement à la prise de décision clinique mais

également à la rationalisation des processus hospitaliers, tout en améliorant les résultats pour les patients [115],[117].

2.8 Exemples d'IA dans les systèmes de santé

L'intelligence artificielle est désormais omniprésente dans les systèmes de santé modernes, avec des applications allant du diagnostic clinique à la gestion des infrastructures hospitalières. Dans le domaine de l'imagerie médicale, des algorithmes de Deep Learning sont capables de détecter automatiquement des anomalies sur des radiographies, des IRM ou des scanners, souvent avec une précision dépassant celle des radiologues humains, ce qui permet une prédiction plus rapide et plus fiable de maladies comme le cancer du sein ou les tumeurs cérébrales.

Des plateformes avancées telles qu'IBM Watson for Health utilise le traitement du langage naturel (NLP) pour analyser de vastes corpus de publications scientifiques, facilitant ainsi la prise de décision clinique en fournissant aux médecins des recommandations basées sur les dernières données probantes. En oncologie, l'IA est utilisée pour prédire l'évolution des tumeurs, simuler les réponses aux traitements et concevoir des plans thérapeutiques personnalisés en fonction du profil génomique du patient.

L'IA est également un outil de choix pour la prédiction des maladies chroniques telles que le diabète, l'hypertension ou la maladie rénale chronique (MRC). Grâce à l'analyse de grands volumes de données cliniques (dossiers médicaux électroniques, résultats de laboratoire, médicaux), les modèles prédictifs peuvent identifier des patients à risque avant même l'apparition des symptômes, permettant ainsi une intervention précoce et préventive.

Dans les services d'urgence, des systèmes d'IA analysent en temps réel les données cliniques pour trier les patients selon la gravité de leur état, optimisant ainsi la gestion des priorités et des ressources. L'IA joue également un rôle central dans la télémédecine et la surveillance à distance, notamment en exploitant les données de dispositifs connectés (bracelets de suivi, tensiomètres intelligents, etc.) pour surveiller en continu les signes vitaux et détecter des anomalies chez les patients atteints de maladies chroniques.

Enfin, dans les volets administratifs et logistiques, l'IA est utilisée pour automatiser la saisie des dossiers médicaux, optimiser les plannings opératoires, prédire les flux de patients, et améliorer la gestion des stocks hospitaliers, contribuant ainsi à une meilleure efficacité organisationnelle et à une réduction des coûts de santé.

2.9 Rôle de l'IA dans la médecine personnalisée et préventive :

L'intelligence artificielle (IA) joue un rôle de plus en plus central dans la transition des approches médicales traditionnelles, uniformes, vers des modèles de soins plus personnalisés et préventifs. Dans la médecine personnalisée, l'IA permet aux cliniciens d'adapter les traitements en fonction du profil génétique, du mode de vie et des antécédents médicaux propres à chaque patient. Par exemple, en oncologie, des algorithmes d'apprentissage automatique peuvent analyser des données génomiques pour prédire la réponse individuelle d'un patient à

une chimiothérapie spécifique, ce qui permet d'éviter des effets secondaires inutiles et d'améliorer l'efficacité du traitement. Une application concrète est celle du cancer du sein, où des outils basés sur l'IA permettent d'identifier les patientes porteuses de mutations génétiques particulières (comme BRCA1 ou BRCA2) et de recommander des thérapies ciblées appropriées.

Sur le plan préventif, l'IA contribue à la prédiction précoce des maladies, bien avant l'apparition des symptômes, rendant possibles des interventions à un stade opportun. Les dispositifs portables, combinés à des algorithmes d'IA, peuvent surveiller en continu les signes vitaux tels que la fréquence cardiaque, le sommeil ou le taux de glucose dans le sang. Ces systèmes sont capables de détecter des tendances anormales et d'alerter à la fois les patients et les professionnels de santé, évitant ainsi des complications graves. Par exemple, dans la gestion du diabète, certaines plateformes alimentées par l'IA peuvent prédire la probabilité d'un épisode d'hypoglycémie en analysant les données issues des capteurs de glucose, de l'alimentation et de l'activité physique du patient.

Par ailleurs, les outils d'analyse prédictive sont de plus en plus utilisés en santé publique pour identifier les populations à risque de développer des maladies chroniques telles que les maladies cardiovasculaires ou la maladie rénale chronique (MRC). En intégrant les dossiers médicaux électroniques avec des données démographiques et environnementales, les modèles d'IA permettent de classer les patients selon leur niveau de risque et de proposer des stratégies de prévention personnalisées. Cette approche proactive améliore non seulement la qualité des soins mais contribue aussi à réduire les coûts de santé à long terme en limitant le recours à des traitements lourds.

Finalement, l'IA soutient l'émergence d'un système de santé plus proactif, centré sur le patient et guidé par les données, où les interventions ne sont plus uniquement curatives, mais aussi anticipatrices, redéfinissant ainsi les contours de la médecine moderne.

2.10 Conclusion

À travers ce chapitre, nous avons mis en évidence le rôle central que joue l'intelligence artificielle dans la transformation des systèmes de santé. En s'appuyant sur des approches avancées d'apprentissage automatique et d'apprentissage profond, l'IA permet non seulement d'automatiser des tâches complexes, mais aussi d'améliorer significativement la qualité des, notamment grâce à une analyse fine et rapide de données médicales massives. Qu'il s'agisse d'imagerie médicale, de prédiction des maladies chroniques ou d'optimisation des ressources hospitalières, les exemples concrets démontrent l'impact tangible de l'IA dans le quotidien des professionnels de santé et des patients.

Cependant, pour que ces technologies puissent être intégrées de manière éthique, fiable et efficace, il est essentiel de comprendre les fondements techniques qui les sous-tendent, tout en tenant compte des limites actuelles, comme la sensibilité aux biais, les problématiques de confidentialité et les contraintes réglementaires. L'IA, bien que puissante, ne peut se substituer à l'humain ; elle doit rester un outil d'aide à la décision, au service d'une médecine plus préventive, personnalisée et équitable.

Ainsi, ce chapitre constitue une base essentielle pour aborder les développements ultérieurs, notamment ceux liés à la prédiction des maladies chroniques comme la maladie rénale chronique (MRC), où l'IA s'annonce comme un levier prometteur pour améliorer le dépistage précoce, le suivi des patients et la réduction de la charge sur les systèmes de santé.

Chapitre 3

L’IA pour la prédiction de la MRC : Revue de la littérature

3.1 Introduction

La maladie rénale chronique (MRC) est aujourd'hui l'une des pathologies non transmissibles les plus préoccupantes, tant par sa prévalence croissante que par ses conséquences lourdes sur la santé des patients. Ce qui la rend particulièrement dangereuse, c'est qu'elle évolue souvent de manière silencieuse, sans signes cliniques évidents jusqu'à un stade avancé. Dans ce contexte, l'identification précoce des patients à risque devient un enjeu crucial pour limiter la progression de la maladie et améliorer la qualité de vie des malades.

Depuis quelques années, l'essor des technologies numériques et la disponibilité croissante de données médicales ont ouvert la voie à de nouvelles approches de prédiction, en particulier grâce aux méthodes issues de l'intelligence artificielle (IA). En exploitant les capacités des algorithmes d'apprentissage automatique et d'apprentissage profond, il devient possible de détecter des schémas subtils dans les données cliniques et biologiques, souvent imperceptibles pour l'œil humain ou les méthodes classiques.

Ce chapitre propose une revue détaillée des travaux de recherche qui ont exploré l'apport de l'IA dans la prédiction de la MRC. Nous commencerons par présenter les approches traditionnelles utilisées dans le domaine de la médecine prédictive, avant de nous intéresser aux apports concrets du machine learning et du deep learning. Seront également abordées les différentes étapes du processus d'analyse, allant de la collecte des données à l'évaluation des performances des modèles. Enfin, une synthèse critique permettra de mettre en lumière les forces et les limites des approches actuelles, ainsi que les pistes d'amélioration pour l'avenir.

3.2 Stratégies récentes basées sur l'IA pour la prédiction de la maladie rénale chronique

Ces dernières années, l'intelligence artificielle (IA) s'est imposée comme un levier puissant dans le domaine de la prédiction des maladies chroniques, notamment la maladie rénale chronique (MRC). Une grande variété de techniques d'apprentissage automatique (machine learning) et d'apprentissage profond (deep learning) sont aujourd'hui mobilisées pour analyser des données médicales complexes et détecter précocement les risques liés à la MRC. Ces approches vont des algorithmes classiques tels que les machines à vecteurs de support (SVM), les arbres de décision ou les forêts aléatoires, jusqu'à des modèles plus sophistiqués comme les réseaux de neurones profonds ou les modèles ensemblistes capables de saisir des relations non linéaires entre les variables cliniques [118].

L'IA permet d'exploiter des sources de données diverses, incluant les résultats biologiques, les dossiers médicaux électroniques, les antécédents cliniques, les habitudes de vie ou encore les données génétiques, pour construire des modèles prédictifs reflétant la complexité de l'étiologie de la MRC. L'intégration de ces données hétérogènes favorise une stratification plus fine des risques et une approche personnalisée du suivi médical [119] .Ces techniques permettent d'optimiser les capacités des modèles en exploitant au mieux les données disponibles [120].

Par ailleurs, plusieurs travaux mettent en évidence l'intérêt des approches hybrides combinant plusieurs algorithmes pour améliorer la robustesse et la généralisation des prédiction [121]. Ces

modèles sont particulièrement pertinents pour traiter des ensembles de données déséquilibrés, fréquents en santé, à l'aide de méthodes comme le suréchantillonnage (oversampling) ou les techniques d'équilibrage basées sur des ensembles.

Enfin, l'intégration de l'IA dans le domaine de la néphrologie ouvre de nouvelles perspectives en médecine prédictive, en permettant un diagnostic plus précoce, plus fiable et adapté aux spécificités individuelles des patients.

3.3 La Médecine prédictive et la prédiction précoce de la MRC

La médecine prédictive, au cœur de la médecine dite « 4P » (prédictive, préventive, personnalisée et participative) [122],[123], marque une évolution majeure dans la manière d'aborder les maladies chroniques comme la maladie rénale chronique (MRC). Contrairement à une approche purement curative, elle vise à anticiper les risques de développement ou de progression d'une pathologie à partir d'une analyse approfondie des données cliniques, biologiques, démographiques et parfois même génétiques. Dans le cas de la MRC, qui évolue de manière silencieuse jusqu'à des stades avancés, la détection précoce revêt une importance particulière. En effet, une identification à un stade précoce permet non seulement de ralentir la progression de la maladie grâce à une prise en charge ciblée, mais aussi d'éviter des traitements lourds comme la dialyse ou la transplantation rénale.

Grâce aux progrès technologiques, notamment en intelligence artificielle et en science des données, il devient possible de construire des modèles capables de détecter des signaux faibles ou des schémas complexes invisibles à l'œil humain. Ces modèles exploitent des données issues de dossiers médicaux électroniques, d'analyses biologiques ou encore de facteurs liés au mode de vie pour évaluer de manière fine le risque de développer une insuffisance rénale [124]. Par exemple, l'utilisation de variables comme le débit de filtration glomérulaire estimé (eGFR), la créatininémie, la pression artérielle ou encore la présence d'albuminurie permet d'établir des scores de risque utiles en pratique clinique[125]. Plusieurs travaux récents ont démontré que l'IA, appliquée à des cohortes de patients, peut prédire l'apparition de la MRC plusieurs mois, voire plusieurs années avant les premiers symptômes, ouvrant la voie à une prévention véritablement personnalisée [126].

3.4 Les méthodes traditionnelles pour la prédiction des maladies chronique

Au fil des décennies, les méthodes statistiques traditionnelles, telles que la régression logistique, ont constitué l'approche de référence pour la prédiction des maladies chroniques. Elles offrent une interprétabilité simple, sont faciles à mettre en œuvre et permettent de dégager des relations linéaires entre variables cliniques. Ces techniques ont été largement utilisées dans la création de scores de risque pour diverses pathologies, notamment les maladies cardiovasculaires, le diabète ou encore l'insuffisance rénale chronique. Toutefois, malgré leur robustesse, ces approches présentent certaines limites, notamment leur capacité réduite à gérer des ensembles de données complexes, volumineux et hautement dimensionnels, caractéristiques fréquentes des données de santé modernes. Des études récentes ont mis en évidence que, dans des cas simples, la régression logistique pouvait offrir des performances comparables aux algorithmes d'apprentissage automatique [127]. Cependant, d'autres travaux

ont montré que l'intelligence artificielle, notamment le machine learning, surpassait les approches traditionnelles lorsqu'il s'agissait de modéliser des interactions complexes ou de traiter des données non structurées [128] ,. Ainsi, bien que les méthodes classiques restent pertinentes, leur combinaison ou leur remplacement progressif par des modèles plus avancés devient une nécessité dans une médecine de plus en plus orientée vers la personnalisation et l'exploitation intelligente des données.

3.5 Intégration de l'apprentissage automatique dans la prédiction de la MRC

Dernièrement, dans le champ de la recherche en santé, l'intégration de l'apprentissage automatique dans le domaine de la néphrologie a ouvert de nouvelles perspectives pour la prédiction précoce et la prédiction de la maladie rénale chronique (MRC). Grâce à leur capacité à analyser des ensembles de données médicaux complexes et multidimensionnels, les algorithmes d'apprentissage automatique permettent d'identifier des schémas subtils souvent indétectables par les approches conventionnelles. Ces techniques offrent la possibilité de construire des modèles prédictifs robustes, capables d'estimer avec précision le risque de développement ou de progression de la MRC. Dans cette section, nous présentons une vue d'ensembles des approches basées sur l'apprentissage automatique appliquées à la prédiction de la MRC, en mettant en lumière les principaux algorithmes utilisés, les caractéristiques exploitées ainsi que les mesures couramment adoptées pour évaluer leurs performances.

3.5.1 Collecte et structuration des données dans le domaine de la santé

L'expansion des sources de données dans le domaine de la santé est un phénomène marquant, avec l'émergence des informations provenant des consultations médicales, des hospitalisations, des appareils connectés (comme les montres intelligentes, les tensiomètres ou les smartphones), ainsi que des données relatives à la consommation de soins. L'intérêt de ces données réside non seulement dans leur volume croissant, mais surtout dans la possibilité de croiser et d'analyser des informations qui étaient auparavant inaccessibles, grâce à l'utilisation de technologies avancées telles que l'intelligence artificielle. Ces nouvelles sources de données apportent des opportunités pour améliorer la prise en charge des patients et personnaliser les traitements.

Cependant, bien que ces données aient un potentiel considérable pour améliorer la santé publique et personnaliser les soins, leur collecte et leur utilisation soulèvent des questions importantes, notamment en matière de confidentialité et de protection de la vie privée des individus. Les données doivent être manipulées avec précaution, car elles révèlent des informations sensibles sur le comportement, l'état de santé et la vie privée des patients. Les acteurs de la santé, tels que les établissements de soins et les compagnies d'assurance, utilisent ces données pour divers objectifs : la gestion des coûts, l'amélioration des soins et la surveillance de l'état de santé des individus à différentes échelles. Toutefois, l'accès à ces données reste complexe en raison des réglementations strictes et des enjeux commerciaux, ce qui peut constituer un obstacle pour les chercheurs indépendants.

3.5.1.1 Les types de données dans le domaine médicale

Les données dans le domaine médical se présentent sous plusieurs formes, chacune ayant des caractéristiques particulières et des applications spécifiques dans le cadre de l'analyse des soins de santé. Voici les principaux types de données utilisés :

- **Données d'imagerie médicale**

Les données provenant des techniques de radiologie telles que les radiographies, les échographies, les scanners CT, les IRM, et les scanners PET, qui produisent des images des tissus internes. Ces données nécessitent des techniques avancées de traitement d'images pour en extraire des informations significatives et sont couramment utilisées pour le diagnostic visuel par des spécialistes [129].

- **Données textuelles non structurées**

Ces données incluent des notes manuscrites ou tapées par des médecins et infirmiers, contenant souvent des informations sur les visites des patients. Elles peuvent être en partie désorganisées, avec des phrases incomplètes, des abréviations ou des erreurs typographiques. Leur extraction nécessite l'utilisation de techniques avancées de traitement du langage naturel pour les structurer et les analyser efficacement [130].

- **Données structurées**

Il s'agit de données organisées dans des formats prédéfinis tels que des bases de données ou des tableurs. Ces données incluent les informations démographiques des patients (nom, âge, sexe), les résultats de tests médicaux (analyses sanguines, tests d'urine), les signes vitaux (pression artérielle, fréquence cardiaque), et les données de traitement (médicaments prescrits). Elles sont facilement analysables à l'aide de techniques statistiques et sont essentielles pour l'application de modèles d'intelligence artificielle [129].

- **Données des dossiers de santé électroniques (EHR)**

Les EHR contiennent des informations complètes et détaillées sur l'historique médical des patients, telles que les antécédents médicaux, les traitements, les résultats des tests, et les visites médicales. Ces données sont stockées de manière numérique et permettent un suivi précis de l'état de santé du patient tout au long de son parcours médical [131].

- **Données de remboursement**

Les données de remboursement, généralement collectées par les assureurs santé, incluent des informations sur les services médicaux reçus par un patient, les traitements administrés, ainsi que les coûts associés. Bien que ces données soient principalement centrées sur les aspects financiers, elles offrent également un aperçu précieux de l'utilisation des services de santé par les patients [130].

- **Données des dispositifs médicaux connectés**

Les appareils médicaux et non médicaux connectés, tels que les montres intelligentes, les tensiomètres, et les pèse-personnes, collectent des données en temps réel sur l'état de santé des

patients, comme la fréquence cardiaque, la pression artérielle, ou le poids. Ces données permettent un suivi continu et peuvent être intégrées dans des systèmes prédictifs pour améliorer la gestion des soins de santé.

3.5.2 Les outils de préparation et de prétraitement des données

Avant d'appliquer des techniques d'analyse ou des algorithmes d'apprentissage automatique, il est essentiel de transformer les données brutes en un format cohérent, fiable et exploitable. Dans le domaine médical, les données collectées proviennent de sources variées et présentent souvent des problèmes tels que des valeurs manquantes, des erreurs de saisie, des formats hétérogènes ou des doublons. Le prétraitement des données vise à garantir leur qualité, à réduire les biais et à améliorer la pertinence des analyses futures. Cette étape englobe un ensembles de méthodes et d'outils, allant du nettoyage et de la normalisation à la réduction de dimensions ou encore à l'encodage des variables catégorielles. Les sections suivantes décrivent les techniques les plus utilisées pour préparer efficacement les données médicales à l'analyse.

3.5.2.1 Nettoyage des données

Le nettoyage des données constitue une étape fondamentale dans le processus de prétraitement, en particulier dans le domaine médical où la qualité des données influe directement sur la fiabilité des résultats. Cette phase consiste à identifier et corriger les erreurs, à gérer les valeurs manquantes, à supprimer les doublons et à détecter les incohérences. Un nettoyage rigoureux permet d'assurer la cohérence et la crédibilité des analyses subséquentes.

3.5.2.1.1 Stratégies de traitement des valeurs manquantes

Dans tout processus d'analyse de données, il est fréquent de rencontrer des données incomplètes, résultant de diverses causes telles que des erreurs humaines, des défaillances techniques ou des limitations lors de la collecte. Ces lacunes peuvent compromettre la qualité des analyses et la performance des modèles prédictifs. Plusieurs méthodes permettent de traiter ce problème efficacement :

a) Suppression des données manquantes

Cette approche consiste à éliminer les observations ou les variables contenant des valeurs absentes. Elle est recommandée lorsque la proportion de données manquantes est faible et que leur suppression n'entraîne pas de perte significative d'information. Cette méthode est simple mais doit être appliquée avec prudence pour éviter une réduction excessive de la taille de l'échantillon.

b) Détection des valeurs aberrantes

Les outliers ou valeurs extrêmes sont des données qui s'écartent fortement des tendances générales. Ils peuvent provenir d'erreurs de saisie, de dysfonctionnements matériels ou de cas rares mais légitimes. Leur identification repose souvent sur des techniques statistiques (comme

la méthode de l'écart interquartile ou la détection par z-score), et leur traitement dépend du contexte : on peut les corriger, les transformer ou les exclure.

c) Correction des erreurs

Les données peuvent contenir des erreurs de saisie, des fautes de format ou des incohérences d'unité de mesure. Ces erreurs doivent être identifiées manuellement ou automatiquement, puis corrigées pour garantir l'intégrité du jeu de données. Cette étape est essentielle, notamment pour les variables critiques utilisées dans les modèles de prédiction.

d) Normalisation des noms de catégories

Il arrive fréquemment que des catégories identiques soient libellées différemment (par exemple : "masculin", "M", "Homme"). Une harmonisation s'impose alors pour éviter la duplication inutile d'étiquettes. Ce processus de standardisation facilite l'analyse et améliore la cohérence des données qualitatives.

e) Suppression des doublons

Les données redondantes, résultant souvent d'une saisie multiple ou d'une fusion incorrecte de fichiers, peuvent fausser les analyses. L'élimination des enregistrements dupliqués permet de garantir l'unicité de chaque observation et d'éviter les biais statistiques.

3.5.2.1.2 Imputation des données

L'imputation des données manquantes est une étape essentielle dans le traitement des données pour garantir la qualité et la fiabilité des modèles d'apprentissage automatique. En particulier, dans le cadre de la prédiction des maladies chroniques, cette étape permet de remplir les valeurs manquantes sans perdre d'information cruciale. Voici les principales méthodes d'imputation :

a) Imputation par la moyenne (mean imputation)

Cette méthode consiste à remplacer les valeurs manquantes d'une variable numérique par sa moyenne calculée à partir des valeurs disponibles. Elle est facile à mettre en œuvre, mais peut réduire la variance des données et introduire un biais si les valeurs manquantes ne sont pas aléatoires.

b) Imputation par la médiane (median imputation)

La médiane est souvent utilisée lorsqu'il y a des valeurs extrêmes ou une distribution asymétrique. Elle remplace les valeurs manquantes par la valeur centrale des observations disponibles, ce qui permet de réduire l'impact des outliers.

c) Imputation par le mode (mode imputation)

Pour les variables catégorielles, l'imputation par le mode est généralement adoptée. Elle consiste à remplacer les valeurs manquantes par la modalité la plus fréquente dans la variable concernée. Cette approche est simple mais peut fausser la distribution des classes si les données manquantes sont nombreuses.

d) Imputation par K-plus proches voisins (K-NN)

Cette méthode remplace les valeurs manquantes par une moyenne pondérée des valeurs enregistrées des voisins les plus proches dans l'ensembles des données [132]. L'imputation par K-NN est souvent utilisée lorsque les variables sont fortement corrélées. Cependant, elle peut être limitée par la complexité des relations entre les variables dans des ensembles de données à grande échelle.

e) Imputation par réseaux de neurones

L'imputation basée sur des réseaux de neurones utilise des architectures d'apprentissage profondes pour prédire les valeurs manquantes à partir des relations complexes entre les données enregistrées [133],[134]. Cette méthode est particulièrement utile lorsque les données présentent des relations non linéaires. Par exemple, l'imputation par réseaux de neurones auto-encodeurs est souvent employée pour gérer les ensembles de données avec de nombreuses valeurs manquantes tout en préservant la structure sous-jacente des données [135].

f) Imputation hybride

Une approche hybride combine plusieurs méthodes d'imputation afin d'améliorer la précision des prédictions. Par exemple, une combinaison d'imputation par K-NN et de régression linéaire peut être utilisée pour traiter les variables continues et catégorielles de manière plus robuste. Cette méthode est efficace lorsque les données manquantes ne sont pas réparties de manière aléatoire, ce qui permet de réduire les biais et d'augmenter la précision du modèle [132].

g) Imputation par régression linéaire

L'imputation par régression linéaire consiste à prédire les valeurs manquantes en utilisant les autres variables comme prédicteurs. Cette méthode est adaptée aux ensembles de données où il existe une relation linéaire entre les variables, et elle peut être particulièrement utile lorsque les autres méthodes d'imputation échouent à capturer des dépendances complexes.

h) Imputation par analyse en composantes principales (PCA)

Dans cette approche, les valeurs manquantes sont estimées en fonction des composantes principales extraites du jeu de données. L'idée est d'identifier les directions de variance maximale dans les données et de les utiliser pour estimer les valeurs manquantes en minimisant la perte d'information due à l'absence de données.

Il est important de souligner que le choix de la méthode d'imputation dépend fortement de la nature des données et de la tâche de prédiction. Une imputation incorrecte peut introduire des erreurs systémiques dans les résultats du modèle. Par conséquent, il est essentiel de sélectionner la méthode la plus appropriée en fonction des caractéristiques spécifiques du jeu de données et des objectifs du modèle [135].

3.5.2.1.3 Standardisation et normalisation des données

La standardisation et la normalisation sont des étapes cruciales dans le prétraitement des données, particulièrement lorsqu'on utilise des algorithmes sensibles à l'échelle des variables. Elles visent à rendre les variables comparables entre elles, à accélérer le processus d'apprentissage, et à améliorer la précision des modèles. La normalisation consiste à transformer les données pour qu'elles respectent certaines propriétés statistiques ou qu'elles soient contenues dans une plage donnée, sans pour autant altérer la structure relative des valeurs.

Selon la littérature, plusieurs techniques sont disponibles pour normaliser ou standardiser les données [102],[136]. Les plus couramment utilisées sont présentées ci-dessous :

i) Normalisation Min-Max

La normalisation Min-Max est une transformation linéaire qui réduit les valeurs d'une variable dans une plage spécifique, typiquement [0, 1] ou [-1, 1]. La formule est la suivante :

$$x' = \text{low} + \frac{(high-low)(x_m-min(X))}{\max(X)-\min(X)} \qquad (1)$$

Où xmx_mxm est la valeur initiale, xm′x'_mxm′ la valeur transformée, et low, high les bornes de la plage cible.

Cette technique est simple à mettre en œuvre et efficace lorsque les données sont bornées et ne contiennent pas de valeurs aberrantes. Cependant, son principal inconvénient réside dans sa dépendance aux valeurs minimales et maximales, qui peuvent ne pas être connues à l'avance pour les données hors échantillon [102].

j) Normalisation Z-score

La standardisation Z-score transforme les données de manière à ce que la moyenne de la variable soit nulle et son écart-type égal à un :

$$x'_m = \frac{X_m-\mu}{\delta}$$

Où μ\muμ représente la moyenne et σ\sigmaσ l'écart-type de la variable. Cette méthode est largement utilisée lorsque les données suivent une distribution normale. Elle permet de réduire l'impact des unités différentes entre les variables. Toutefois, elle n'impose pas une plage bornée pour les valeurs transformées, et sa sensibilité aux valeurs aberrantes peut constituer une limitation dans certains contextes [136].

k) Normalisation par la médiane

Dans cette méthode, chaque valeur d'une variable est divisée par la médiane de cette variable :

$$X'_m = \frac{X_m}{médiane(X)}$$

Ce type de normalisation est robuste face aux valeurs extrêmes et convient particulièrement aux distributions asymétriques. Il est souvent utilisé lorsque l'on souhaite comparer des échantillons hétérogènes ou dans des contextes nécessitant une meilleure résistance aux outliers [136].

l) Normalisation par la méthode tangente hyperbolique "Tanh"

La normalisation par tangente hyperbolique est une transformation non linéaire qui s'exprime comme suit :

$$X'_{m=\frac{1}{2}} \left[\tanh\left(0.01.\frac{X_{m-\mu}}{\sigma}\right)+1\right]$$

Où μ\muμ est la moyenne et σ\sigmaσ l'écart-type des données. Cette méthode permet de comprimer les valeurs extrêmes tout en conservant la structure relative des données. Elle est particulièrement adaptée aux modèles d'apprentissage profond, car elle renforce la stabilité numérique tout en maintenant les valeurs dans l'intervalle [0, 1] [136].

3.5.2.1.4 Encodage des variables catégoriques

Les variables catégoriques représentent des attributs non numériques tels que le sexe, la couleur ou la classe de risque. Ces variables ne peuvent pas être directement utilisées dans la majorité des algorithmes de machine learning, qui exigent des entrées numériques. Ainsi, il est essentiel de transformer ces données textuelles en valeurs numériques via des techniques d'encodage adaptées. Le choix de la méthode dépend du type de variable (nominale ou ordinale), de la nature du problème et de l'algorithme utilisé.

a) Encodage par étiquetage (Label *Encoding*) : Le label encoding attribue un entier unique à chaque catégorie d'une variable. Par exemple, pour une variable "couleur" contenant les valeurs "rouge", "vert" et "bleu", on pourrait obtenir respectivement les valeurs 0, 1 et 2. Cette méthode est simple et efficace, particulièrement utile pour les algorithmes qui peuvent exploiter une relation d'ordre. Cependant, elle peut introduire un ordre artificiel entre les catégories, ce qui peut fausser les résultats si la variable est purement nominale.

b) Encodage ordinal (Ordinal *Encoding*) : Conçu pour les variables catégoriques ordonnées, l'encodage ordinal associe à chaque catégorie un entier reflétant son rang. Par exemple, pour une variable "taille" avec les valeurs "petit", "moyen" et "grand", on peut leur attribution respectivement 0, 1 et 2. Cette technique conserve l'ordre naturel des catégories et peut être différentielle pour les modèles sensibles à la hiérarchie, comme les arbres de décision ou les régressions linéaires.

c) Encodage par fréquence (Frequency *Encoding*) : Cette méthode remplace chaque catégorie par sa fréquence d'apparition dans l'ensembles des données. Si la catégorie « ville » contient « Paris » (50 %), « Lyon » (30 %) et « Nice » (20 %), ces valeurs peuvent être directement utilisées comme codes numériques. L'encodage par fréquence est simple, rapide à mettre en œuvre, et peut s'avérer utile lorsque la fréquence de certaines catégories est corrélée à la variable cible.

d) Encodage binaire (Binary *Encoding*) : L'encodage binaire combine les avantages du label encoding et du one-hot encoding. Il commence par convertir chaque catégorie en un entier (comme dans le label encoding), puis transformer cet entier en représentation binaire, répartie

sur plusieurs colonnes. Par exemple, si une catégorie reçoit le code entier 5, elle sera transformée en "101" sur trois colonnes. Cette technique est particulièrement adaptée aux variables comportant un grand nombre de modalités, car elle réduit la dimensionnalité tout en conservant une structure informative

e) **Encodage cible (Target *Encoding*) :** L'encodage cible consiste à remplacer chaque catégorie par une statistique calculée à partir de la variable cible, généralement la moyenne. Par exemple, si la variable « quartier » est encodée en fonction du taux moyen de diabète observé dans chaque quartier, chaque quartier se verra attribuer cette valeur moyenne. Bien que cette méthode permette d'introduire des informations pertinentes dans les données, elle nécessite une attention particulière pour éviter le surapprentissage, notamment par l'utilisation de techniques de validation croisée.

f) **Encodage one-hot (One-*Hot Encoding*) :** L'encodage one-hot crée une colonne binaire pour chaque catégorie. Si une variable "animal" contient les valeurs "chat", "chien" et "lapin", trois colonnes seront générées : "animal_chat", "animal_chien" et "animal_lapin". Chaque observation aura un '1' dans la colonne correspondante à sa catégorie et '0' dans les autres. Cette méthode est particulièrement adaptée pour les variables nominales sans ordre hiérarchique, bien qu'elle puisse augmenter significativement la dimension du jeu de données si le nombre de catégories est élevé.

3.5.3 Augmentation des données

L'augmentation des données est une technique essentielle en science des données et en apprentissage automatique, visant à enrichir le jeu de données d'origine pour améliorer la performance et la robustesse des modèles prédictifs. Elle est particulièrement utile lorsque les données sont peu nombreuses, déséquilibrées ou sujettes à l'overfitting. Cette étape consiste à générer de nouvelles instances à partir des données existantes ou à modifier celles-ci de manière contrôlée, sans changer leur signification ou leur répartition globale. L'augmentation des données peut s'appliquer aussi bien aux données tabulaires qu'aux données images, texte ou audio. Il existe plusieurs approches selon la nature des données et les objectifs visés.

a) **Suréchantillonnage :** Est une technique qui consiste à augmenter artificiellement le nombre d'exemples de la classe minoritaire dans un jeu de données déséquilibré. Cela permet d'équilibrer les classes afin d'éviter que les modèles d'apprentissage ne deviennent biaisés en faveur de la classe majoritaire. La méthode la plus simple de suréchantillonnage est la duplication aléatoire des instances de la classe minoritaire. Toutefois, cette duplication peut entraîner un surapprentissage si elle est utilisée de manière excessive.

b) **Sous-échantillonnage :** Le sous-échantillonnage vise à réduire le nombre d'exemples de la classe majoritaire afin de rétablir l'équilibre entre les classes. Cela peut se faire en supprimant aléatoirement des exemples ou en sélectionnant les plus représentatifs. Bien que cette méthode réduise le biais du modèle envers la classe majoritaire, elle peut aussi entraîner une perte d'information importante, surtout si la classe majoritaire contient une grande diversité de cas.

c) **SMOTE (technique de suréchantillonnage synthétique des minorités) :** est une méthode avancée de sur échantillonnage qui crée de nouvelles instances synthétiques de la classe minoritaire plutôt que de simplement les dupliquer. Elle fonctionne en interpolant entre un échantillon minoritaire et ses plus proches voisins dans l'espace des caractéristiques. Cela permet de générer des exemples plus diversifiés et réalistes, améliorant ainsi la capacité du modèle à généraliser. Cependant, SMOTE peut parfois générer des points peu représentatifs si les classes sont fortement imbriquées.

d) **ADASYN (échantillonnage synthétique adaptatif) :** est une variante de SMOTE qui se concentre davantage sur les exemples minoritaires difficiles à apprendre, c'est-à-dire ceux proches de la frontière entre les classes. Elle génère plus d'exemples synthétiques autour de ces cas, adaptant le suréchantillonnage aux régions les plus complexes du jeu de données. ADASYN améliore ainsi la précision locale du modèle dans les zones où le déséquilibre est la plus problématique.

e) **Augmentation aléatoire (Random Over/Under Sampling) :** L'augmentation aléatoire combine les approches de suréchantillonnage et de sous-échantillonnage de manière simple. Le RandomOverSampler (de la bibliothèque *imblearn*) génère des copies aléatoires des instances minoritaires, tandis que le RandomUnderSampler sélectionne aléatoirement un sous-ensembles d'instances majoritaires. Ces techniques sont faciles à mettre en œuvre et utiles comme base de comparaison.

f) **Génération de données synthétiques :** Les techniques modernes de génération de données utilisent des modèles génératifs comme les GANs (Generative Adversarial Networks) ou le**s** VAE (Variational Autoencoders) pour produire de nouvelles données qui imitent la distribution réelle. Ces approches sont particulièrement puissantes pour les données complexes (images, textes, signaux biomédicaux), car elles permettent de générer des échantillons très réalistes. Bien que plus coûteux en calcul, elles ouvrent la voie à une augmentation de données très précise et contrôlée.

g) **SMOTENC (SMOTE pour Nominal et Continu) :** est une extension de SMOTE conçue pour gérer les ensembles de données contenant à la fois des variables continues et catégorielles. Contrairement au standard SMOTE, qui fonctionne uniquement sur des variables numériques, SMOTENC identifie les colonnes catégorielles et les traite différemment lors de la génération des exemples synthétiques, provoquant ainsi de produire des combinaisons incohérentes. Dans votre étude, SMOTENC a permis d'augmenter efficacement les instances minoritaires tout en respectant la nature mixte des variables, améliorant ainsi la qualité des prédictions sur des données hétérogènes.

h) **Random Under Sampler :** est une méthode simple de sous-échantillonnage qui consiste à réduire la taille de la classe majoritaire en sélectionnant aléatoirement un sous-ensemble de ses instances. Cette technique permet de rééquilibrer les classes sans modifier la classe minoritaire. Elle est utile lorsque le volume de données de la classe majoritaire est très élevé, ce qui peut ralentir l'entraînement ou biaiser les prédictions. Toutefois, en supprimant une partie informative des données, elle peut également entraîner une perte de diversité ou d'informations importantes. C'est pourquoi cette méthode est souvent utilisée en combinaison avec des techniques de suréchantillonnage ou d'optimisation du modèle.

3.5.4 Extraction et sélection des caractéristiques

L'extraction et la sélection des caractéristiques constituent une étape essentielle dans tout processus d'analyse de données. Elles visent à identifier, transformer ou choisir les variables les plus pertinentes pour améliorer la performance des modèles d'apprentissage. Une base de données peut contenir un grand nombre de variables, dont certaines sont redondantes, faiblement informatives ou bruitées. Travailler avec un ensemble de caractéristiques réduit mais pertinent permet non seulement de diminuer la complexité du modèle, mais aussi d'améliorer la précision, de réduire le risque de surapprentissage et d'accélérer les temps de calcul. L'extraction consiste à créer de nouvelles variables à partir des données initiales, tandis que la sélection vise à choisir, parmi les variables disponibles, celles qui apportent le plus d'informations pour la tâche d'apprentissage. Plusieurs méthodes peuvent être mobilisées à cette fin, allant des approches statistiques classiques aux techniques avancées de réduction de dimensionnalité ou d'optimisation automatique.

3.5.4.1 Algorithme de sélection et de réduction minimale absolue (LASSO)

L'algorithme LASSO (Least Absolute Shrinkage and Selection Operator) est largement utilisé dans le cadre de la régression linéaire pour effectuer à la fois la régularisation des coefficients et la sélection des variables. Son principal objectif est d'améliorer la capacité de généralisation des modèles tout en simplifiant leur interprétation. En introduisant une pénalité basée sur la somme des valeurs absolues des coefficients, LASSO tend à réduire certaines d'entre eux à zéro, ce qui conduit à l'élimination automatique des variables les moins pertinentes [137].

Cette méthode est particulièrement efficace dans les contextes où les données présentent une forte multicolinéarité, car elle permet d'identifier un sous-ensembles restreint de variables ayant une influence significative sur la variable cible [138]. LASSO repose sur une régularisation de type L1, où la contrainte de pénalisation favorise la parcimonie du modèle. Ainsi, au fur et à mesure que le paramètre de régularisation augmente, de plus en plus de coefficients deviennent nuls, ce qui se traduit par un modèle plus épuré.

Le fonctionnement du LASSO consiste à mettre à jour les valeurs absolues des coefficients des variables. Les caractéristiques associées à des coefficients nuls ou négatifs sont supprimées de l'ensembles des variables retenues [137] . Toutefois, il est possible que certaines variables non pertinentes soient temporairement conservées dans le processus de sélection. En fin

d'évaluation, seules les caractéristiques les plus fortement corrélées à la variable cible sont enregistrées.

3.5.4.2 Élimination des caractéristiques récursives (RFE)

La méthode RFE (Recursive Feature Elimination) est une technique de sélection de caractéristiques de type *Wrapper* [139], qui applique une stratégie de sélection séquentielle par élimination arrière. Elle repose sur un algorithme d'apprentissage spécifique (comme SVM, régression logistique ou arbre de décision) pour évaluer l'importance des variables. RFE procède de manière itérative : à chaque étape, elle supprime la variable la moins significative jusqu'à ce qu'un nombre prédéfini de caractéristiques soit atteint. Cette méthode a démontré une grande efficacité dans le domaine de la sélection de gènes, notamment en combinaison avec les machines à vecteurs de support (SVM). Depuis, elle est devenue une méthode de référence dans plusieurs domaines de recherche [140].

RFE utilise une approche gloutonne basée sur le classement des caractéristiques. Elle commence par inclure l'ensembles des variables du jeu de données, puis évalue leur importance via le modèle sous-jacent choisi. À chaque itération, la variable la moins pertinente est retirée, et le processus est répété jusqu'à atteindre la taille de sous-ensembles souhaitée [141]. Ce paramètre la taille du sous-ensembles de caractéristiques peut être défini par l'utilisateur, permettant de contrôler la complexité du modèle et de limiter le surapprentissage [142].

L'objectif de RFE est de conserver uniquement les variables les plus pertinentes pour la tâche de prédiction considérée, ce qui améliore non seulement la précision du modèle, mais aussi sa capacité de généralisation sur de nouvelles données. Cette approche permet de construire des modèles plus performants, plus efficaces et mieux adaptés au domaine d'application en ne retenant que les caractéristiques les plus informatives et les plus représentatives.

3.5.4.3 RELIEF

L'algorithme Relief, initialement proposé par Kira et Rendell [143] ,est une méthode de sélection de variables fondée sur l'apprentissage par instance. Il s'agit d'un filtre qui évalue individuellement la pertinence de chaque caractéristique en lui attribuant un score (ou poids), reflétant son utilité pour prédire la variable cible. Pour cela, l'algorithme sélectionne aléatoirement un certain nombre d'instances d'apprentissage et, pour chacune, identifie ses deux plus proches voisines : l'une appartenant à la même classe (appelée *nearest hit*), et l'autre à la classe opposée (*nearest miss*). Le poids d'une caractéristique est ensuite ajusté selon les différences de valeur observées avec ces voisins : une différence avec la *miss* indique une caractéristique informative, dont le poids est augmenté ; une différence avec le hit suggère le contraire, et le poids est diminué. Cette mise à jour est normalisée pour que les poids finaux se situent dans l'intervalle [-1, 1]. La fonction diff, utilisée pour mesurer ces écarts, est adaptée aussi bien aux variables discrètes qu'aux variables continues, bien que certains travaux aient montré qu'elle peut sous-estimer la qualité des variables continues dans les ensembles de données mixtes [144]. Enfin, bien que Relief utilise la distance euclidienne pour identifier les plus proches voisins, d'autres métriques peuvent être utilisées. Il est important de noter que la version originale de Relief est restreinte aux tâches de classification binaire et ne gère pas les

données manquantes. Pour pallier ces limitations, des variantes comme Relief ont été développées [144].

3.5.4.4 Analyse en Composantes Principales (PCA)

L'Analyse en Composantes Principales (PCA) est une technique linéaire de réduction de dimensionnalité qui permet de transformer un ensembles initialement large de variables en un nombre plus restreint tout en conservant l'essentiel de l'information d'origine [145]. Elle consiste à rechercher une combinaison linéaire des variables maximisant la variance, puis à supprimer cette composante pour en extraire une seconde, et ainsi de suite, de manière itérative. Ce procédé, appelé méthode des axes principaux, génère des composantes orthogonales, donc non corrélées entre elles. Techniquement, PCA repose sur le calcul des valeurs propres et des vecteurs propres de la matrice de covariance. Les vecteurs propres sont ensuite triés selon l'ordre décroissant de leurs valeurs propres, puis les données initiales sont projetées sur ces nouvelles directions [146].

Par ailleurs, PCA est également considéré comme une méthode de sélection de caractéristiques, car elle permet de réduire la dimensionnalité de l'ensembles des variables. Elle applique une transformation linéaire orthogonale des données vers un nouveau système de coordonnées. Dans ce système, le premier axe correspond à la direction présentant la plus grande variance des données, le second à la deuxième plus grande, etc. Si l'on considère un ensembles de données représenté par une matrice X de dimensions m×n, où chaque colonne correspond à une variable centrée (moyenne empirique nulle), et chaque ligne représente une répétition expérimentale, PCA permet de reformuler ces données selon de nouvelles composantes plus significatives [147].

3.5.4.5 Boruta

La méthode Boruta est une technique de sélection de caractéristiques basée sur l'algorithme des forêts aléatoires. Contrairement aux approches classiques qui visent à identifier un sous-ensembles minimal de variables optimisant un modèle donné, Boruta cherche à sélectionner l'ensembles des variables significativement liées à la variable dépendante [148]**.** Pour cela, elle introduit des "shadow features" générées par permutation aléatoire des variables d'origine, puis entraîne un modèle (généralement un Random Forest) sur l'ensembles des variables réelles et d'ombre combinées. Ensuite, elle compare l'importance de chaque variable réelle à celle de ses équivalents aléatoires à l'aide de métriques telles que l'impureté de Gini ou le gain d'information. Une variable est considérée comme pertinente si son score d'importance dépasse significativement celui des variables d'ombre. Ce processus est répété jusqu'à ce que toutes les variables soient classées comme importantes ou non. Boruta présente plusieurs avantages : il prend en charge différents types de modèles (arbres ou non), gère les redondances et interactions entre variables, et est robuste au bruit. Toutefois, il peut être coûteux en calcul pour les ensembles de données à forte dimension, et ses performances dépendent du choix du seuil d'importance et de la qualité des mesures utilisées [149],[150].

3.5.4.6 Autres techniques de sélection des caractéristiques

En complément des techniques déjà présentées, d'autres méthodes de sélection de caractéristiques sont largement utilisées en apprentissage automatique pour réduire la dimensionnalité des données, supprimer les redondances et améliorer la performance des modèles. Parmi ces techniques, l'analyse de variance (*ANOVA*) est une méthode statistique qui permet d'identifier les caractéristiques ayant un effet significatif sur la variable cible, particulièrement utile pour les problèmes de classification impliquant des variables numériques continues. Le *test du Chi-carré* est quant à lui adapté aux variables catégorielles, et sert à évaluer l'indépendance entre chaque caractéristique et la classe cible. Une autre approche importante repose sur *l'information mutuelle*, qui mesure la dépendance (linéaire ou non linéaire) entre une variable explicative et la variable à prédire. Elle permet de sélectionner les caractéristiques qui partagent le plus d'informations avec la sortie. Par ailleurs, les méthodes basées sur l'importance des caractéristiques dans les arbres de décision, comme *le gain d'information, l'indice Gini* ou encore *l'importance moyenne des permutations*, sont également très employées, notamment avec des modèles tels que les forêts aléatoires (Random Forest). Ces techniques permettent de quantifier directement la contribution de chaque variable à la performance du modèle. Enfin, certaines méthodes modernes comme *SelectKBest* ou *SelectPercentile*, intégrées dans la bibliothèque Scikit-learn, permettent de sélectionner automatiquement un sous-ensemble des meilleures caractéristiques selon un score donné. Le choix de la méthode dépend en grande partie du type de données manipulées et de la nature du problème (classification ou régression).

3.5.5 Les modèles de ML utilisés dans la littérature

Dans le domaine de la prédiction des maladies chroniques, en particulier la maladie rénale chronique (MRC), de nombreux travaux de recherche ont exploré et appliqué diverses approches d'apprentissage automatique (ML) afin de développer des outils de diagnostic précoces, précis et automatisés. La littérature récente montre un recours croissant à ces modèles, en raison de leur capacité à gérer des ensembles de données complexes, à découvrir des schémas cachés, et à améliorer la prise de décision médicale. Ces modèles varient selon leur nature (linéaire, non linéaire, probabiliste, ensembliste, etc.) et leurs performances dépendent généralement de plusieurs facteurs, notamment la qualité des données, la méthode de sélection des caractéristiques, et les techniques de validation employées. Dans ce qui suit, nous présentons un aperçu des modèles les plus utilisés, en les regroupant selon leur type, à commencer par les classificateurs traditionnels.

3.5.5.1 Classificateurs traditionnels

Les classificateurs traditionnels constituent la base de nombreux systèmes prédictifs dans le domaine médical. Ils se caractérisent par leur simplicité d'implémentation, leur interprétabilité, et leur efficacité dans des contextes où les relations entre les variables sont relativement bien définies.

3.5.5.1.1 Régression logistique

Est un algorithme d'apprentissage statistique supervisé couramment utilisé comme modèle pour prédire la probabilité d'une variable binaire cible. Elle permet de classer les observations en deux catégories, généralement représentées par 1 (succès) et 0 (échec). Contrairement à la régression linéaire, qui prédit une valeur continue, la régression logistique s'intéresse à la probabilité qu'un événement appartenant à une classe spécifique [151].

Cette méthode est largement utilisée dans des domaines tels que les sciences sociales et la médecine, notamment pour des applications comme la détection du spam, le diagnostic du diabète ou la détection du cancer. Elle se distingue par sa capacité à modéliser la relation entre les variables indépendantes et une variable dépendante binaire, en se concentrant uniquement sur la probabilité d'occurrence de l'événement étudié [151].

3.5.5.1.2 Random Forest

La méthode Random Forest (RF) appartient à la catégorie des techniques d'ensembles, c'est-à-dire qu'elle combine plusieurs algorithmes pour construire un modèle plus performant[152] ,[153]. Plus précisément, elle repose sur l'utilisation de multiples arbres de décision (Decision Trees), chacun construit à partir de sous-ensembles aléatoires et non corrélés issus des données d'entraînement. Durant la phase d'apprentissage, on sélectionne aléatoirement K échantillons à partir de l'ensembles d'entraînement, puis on génère un arbre de décision pour chacun de ces échantillons. Ce processus est répété N fois, correspondant au nombre total d'arbres à créer. Lors de la phase de prédiction, chaque arbre fournit une classification, et la forêt décide de la classe finale en appliquant une règle de vote majoritaire. Bien que l'utilisation de multiples arbres augmente la complexité informatique par rapport à un arbre unique, Random Forest démontre généralement une meilleure précision, notamment sur des ensembles de données inconnues [154] ,[155]

3.5.5.1.3 K-Nearest Neighbors

L'algorithme K-Nearest Neighbors (KNN) est un classificateur basé sur les instances. Il attribue une classe à une donnée inconnue en comparant cette dernière aux exemples connus à l'aide d'une mesure de distance, généralement la distance euclidienne. On commence par définir K, le nombre de voisins à prendre en compte. Ensuite, on calcule la distance entre l'instance à classer et chaque échantillon de l'ensembles d'entraînement. Les K plus proches voisins sont identifiés, et la classe la plus représentée parmi eux est attribuée à l'instance inconnue [156].

$$\text{Distance}(X, Z) = \sqrt{\Sigma(x_i - z_i)^2}$$

Où x_i sont les variables des données d'entraînement, et z_i celles de l'entrée, (Formule de la distance euclidienne)**.**

3.5.5.1.4 Machine à vecteurs de support

Le modèle SVM (Support Vector Machine) repose sur la construction d'un hyperplan dans un espace de grande dimension afin de mieux séparer les données en différentes classes. Cet hyperplan est positionné de manière à maximiser la marge, c'est-à-dire la distance entre lui et

les points de données les plus proches de chaque classe, appelés vecteurs de support. L'objectif principal de SVM est de trouver l'hyperplan optimal qui assure une bonne séparation des données tout en garantissant une bonne généralisation sur des données non vues. Pour améliorer sa capacité de séparation, SVM utilise le truc du noyau *(kernel trick),* qui permet de projeter implicitement les données d'entrée dans un espace de dimension supérieure, facilitant ainsi la recherche d'un hyperplan séparateur. Grâce à cette approche, SVM est particulièrement adapté aux ensembles de données linéairement séparables comme non linéairement séparables, offrant robustesse et flexibilité dans les tâches de classification [157].

3.5.5.1.5 Naive Bayes

Le classificateur de type Naive Bayes repose sur l'hypothèse que l'effet d'un attribut sur une classe donnée est indépendant des autres attributs. Cette approche calcule la probabilité a posteriori de chaque classe à l'aide de la formule de Bayes. L'étiquette attribuée à une instance est celle qui maximise cette probabilité a posteriori [158]:

$$P(C|X) = (P[X|C] \times P[C]) / P[X]$$

Où X représente les caractéristiques de l'instance, et C la classe correspondante.

3.5.5.2 Méthodes d'ensembles

3.5.5.2.1 AdaBoost

L'algorithme AdaBoost (pour *Adaptive Boosting*) est une méthode de Boosting conçue principalement pour la classification binaire. Son principe consiste à combiner plusieurs classificateurs faibles afin de créer un classificateur global plus puissant et plus fiable. Le processus débute par l'attribution de poids initial aux instances de l'ensembles d'apprentissage, avec pour objectif d'améliorer la précision des prédictions sur un grand nombre d'exemples, par exemple 1000. Dans certaines études, AdaBoost cherche à construire un classificateur performant à l'aide d'une méthode d'apprentissage appelée *low* learning, qui vise à identifier, à chaque itération, le classificateur faible le plus apte à distinguer efficacement les exemples positifs des négatifs. Durant ce processus, un seuil optimal est déterminé pour chaque attribut, de façon à réduire au maximum les erreurs de classification [152],[159],[160].

3.5.5.2.2 Bagging

Est une méthode d'apprentissage ensembliste développée pour accroître la performance des algorithmes de classification. Elle consiste à générer plusieurs sous-ensembles aléatoires de l'ensembles de données initiales, par échantillonnage avec remise, puis à entraîner un modèle distinct sur chacun de ces sous-ensembles. Par la suite, les prédictions issues de ces modèles sont combinées, souvent par vote majoritaire pour les tâches de classification ou par moyenne dans le cas de la régression, afin d'obtenir une prédiction consolidée [160].

L'un des principaux avantages du Bagging réside dans sa capacité à réduire la variance du modèle, limitant ainsi les risques de surapprentissage (overfitting), sans pour autant augmenter de manière significative le biais. En mutualisant les forces de plusieurs modèles faiblement corrélés, cette approche permet de générer des prédictions globalement plus fiables et précises.

3.5.5.2.3 Gradient Boosting

Est un algorithme d'apprentissage supervisé largement utilisé pour des tâches de classification, de régression et d'autres applications complexes[161]. Contrairement aux apprenants faibles individuels, tels que les arbres de régression simples, cette méthode améliore progressivement la précision des prédictions en augmentant séquentiellement de nouveaux apprenants faibles au modèle existant. Ce processus d'ensembles repose sur la réduction progressive d'une fonction de perte, qui évalue l'écart entre les valeurs attendues et celles réellement prédites [161].

En affinant le modèle à chaque itération, le Gradient Boosting réduit efficacement le biais et la variance, ce qui conduit à un notable des performances prédictives. L'un de ses atouts réside dans sa simplicité d'implémentation, notamment avec la régression par moindres carrés, ce qui en fait une solution attrayante dans de nombreux domaines[161] .

Généralement efficace même avec un nombre réduit d'échantillons (environ 100), le Gradient Boosting convient aussi bien à la classification qu'à la régression. Son fonctionnement repose sur trois éléments clés : une fonction de perte soigneusement définie, un modèle de base utilisé pour la prédiction, et un mécanisme d'agrégation des apprenants faible visant à minimiser la fonction de perte [162]. Cette stratégie permet non seulement de limiter le surapprentissage (*overfitting*), mais aussi d'améliorer la précision des prédictions, notamment dans des contextes où les classes sont déséquilibrées [163].

Alors que les méthodes traditionnelles d'apprentissage automatique se concentrent sur l'optimisation directe de l'espace des fonctions, le Gradient Boosting adopte une stratégie d'amélioration itérative pour générer une fonction optimale, assurant ainsi des performances supérieures sur des ensembles de données variées [164].

3.5.5.2.4 Arbres extrêmement aléatoires

La méthode Extra Tree repose sur l'agrégation des prédictions issues de multiples arbres de décision générés aléatoirement. En régression, elle procède par moyennage des résultats de ces arbres afin d'augmenter la précision des prédictions et de limiter le surapprentissage (*overfitting*). Pour ce faire, un méta-prédicteur est utilisé pour entraîner un grand nombre d'arbres de décision sur différents sous-échantillons du jeu de données.

L'un des avantages notables de cette méthode est sa rapidité d'exécution et son faible coût informatique, ce qui en fait une solution performante dans des contextes à ressources limitées. Cependant, cette technique présente certaines limites, notamment une sélection aléatoire des points de séparation et l'absence d'optimisation systématique pour identifier les meilleurs seuils de découpage [165].

3.5.5.3 Réduction de dimension et régression

3.5.5.3.1Ridge Regression

La régression Ridge est une extension de la méthode des moindres carrés qui permet d'obtenir des estimateurs biaisés mais avec une variance réduite. Elle s'avère particulièrement utile dans les situations où la multicolinéarité est présente entre les variables explicatives, car elle atténue

ce problème en pénalisant les coefficients, ce qui permet d'éviter le surapprentissage. En pratique, cela revient à sacrifier une petite part de précision (introduire un biais) pour gagner en stabilité des estimations [166].

Contrairement à la régression Lasso, qui peut annuler certains coefficients, la régression Ridge tend à réduire la valeur des coefficients des variables corrélées sans les supprimer, leur ainsi de partager l'information entre elles. Dans une perspective bayésienne, cette méthode est bien adaptée lorsque de nombreux prédicteurs sont impliqués et que tous ont une influence non nulle, comme dans le cas d'une distribution normale des coefficients. Par ailleurs, l'analyse de la régression Ridge nécessite de prendre en compte des éléments comme le biais, la variance, et l'erreur quadratique moyenne de l'estimateur pour mieux évaluer sa performance globale [166].

3.5.5.3.2 Linear Discriminant Analysis (LDA)

Est une méthode supervisée couramment utilisée en statistique, en apprentissage automatique et en reconnaissance de formes. Elle vise à projeter les données dans un espace de plus faible dimension tout en maximisant la séparabilité entre plusieurs classes. Pour cela, LDA cherche une combinaison linéaire de variables qui permet de distinguer au mieux les groupes d'observations en maximisant la variance entre les classes tout en minimisant la variance au sein de chaque classe [167].

3.5.6 Les modèles de DL utilisés dans la littérature

3.5.6.1 Artificial Neural Network

Le réseau de neurones artificiels (ANN) s'inspire du fonctionnement du cerveau humain pour traiter l'information. Ce modèle comprend une couche d'entrée (où les données sont introduites), une ou plusieurs couches cachées, et une couche de sortie qui fournit la prédiction finale. L'apprentissage se fait via l'algorithme de rétropropagation (BPN), largement utilisé pour entraîner ce type de réseau. Ce réseau est constitué d'un grand nombre de neurones interconnectés capables de collaborer pour résoudre des problèmes complexes. Grâce à une architecture multicouche, les ANN peuvent être appliqués à des tâches de classification et de prédiction. Les poids associés aux connexions entre les neurones sont ajustés lors de la phase d'apprentissage, ce qui permet au système de s'adapter aux nouvelles données et de généraliser à des situations inconnues [168].

3.5.6.2 Réseau de neurones convolutifs (CNN)

Est un modèle d'apprentissage profond particulièrement efficace pour le traitement des données structurées en grille, comme les images ou les signaux. Il se compose de couches spécialisées, notamment les couches de convolution, qui extraient automatiquement les caractéristiques pertinentes, les couches de pooling, qui réduisent la dimensionnalité tout en conservant les informations essentielles, et les couches entièrement connectées, utilisées pour la classification finale. Les CNN sont largement utilisés pour des tâches telles que la reconnaissance d'images, la détection d'objets et l'analyse de signaux médicaux [164].

3.5.7 Évaluation des performances des modèles IA

L'évaluation des modèles d'intelligence artificielle est une étape cruciale afin de juger de leur efficacité dans la prédiction de la maladie rénale chronique. Cette évaluation repose sur plusieurs problèmes métriques classiques de la matrice de confusion, qui permettent de mesurer différents aspects de la performance des modèles.

La **matrice de confusion**, représentée dans la Figure ci-dessous, distingue quatre catégories principales de prédictions :

- **VP (Vrai Positif)** : nombre de cas positifs correctement prédits.
- **VN (Vrai Négatif)** : nombre de cas négatifs correctement prédits.
- **FP (Faux Positif)** : nombre de cas négatifs incorrectement prédits comme positifs.
- **FN (Faux Négatif)** : nombre de cas positifs incorrectement prédits comme négatifs.

  À partir de cette matrice, plusieurs indicateurs sont calculés :

- **Exactitude (Accuracy)** : mesure globale du taux de prédictions correctes, positives et négatives.

$$\text{Exactitude} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Précision (Precision)** : proportion des prédictions positives qui sont effectivement correctes.

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

- **Sensibilité (Recall)** : mesurez la capacité du modèle à identifier correctement les cas positifs.

$$Sensibilité = \frac{VP}{VP + FN}$$

- **Spécificité (Specificity)** : mesurer la capacité à identifier correctement les cas négatifs.

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

- **Valeur Prédictive Négative (VPN)** : proportion des prédictions négatives qui sont réellement négatives.
- **Score F1** : mesure harmonique entre la précision et la sensibilité, particulièrement utile en cas de déséquilibre entre les classes.

$$\text{Score F1} = 2 * \frac{précision * Sensibilité}{précision + Sensibilité}$$

- **Courbe ROC et AUC (Area Under Curve)** : utilisées pour évaluer la capacité de discrimination des modèles, en particulier dans les contextes de classification binaire.

L'ensembles de ces métriques permet une comparaison approfondie des différents modèles testés (ML et DL), en mettant en évidence leurs points forts et leurs faiblesses. Ces indicateurs sont particulièrement importants dans les contextes médicaux, où l'identification correcte des patients à risque est primordiale pour une prise en charge précoce.

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 3.1: Matrice de confusion illustrant les principaux indicateurs de performance

3.6 Analyse des travaux récents sur la prédiction de la MRC

L'utilisation d'algorithmes d'apprentissage automatique pour la prédiction des maladies chroniques a connu un essor considérable ces dernières années, principalement grâce à leur capacité à améliorer la précision et l'efficacité du diagnostic précoce. Ceci est particulièrement crucial pour des maladies comme la maladie rénale chronique (MRC), souvent asymptomatique à ses débuts. L'application d'outils basés sur l'IA facilite une intervention médicale rapide, améliorant ainsi les résultats pour les patients et optimisant les ressources de santé.

Ani et al. [156] ont développé un système d'aide à la décision visant à aider les cliniciens à diagnostiquer l'IRC à l'aide de divers algorithmes d'apprentissage automatique, notamment l'analyse discriminante linéaire (LDA), la rétropropagation basée sur un réseau neuronal (BPN), la méthode bayésienne naïve, le classificateur de sous-espaces aléatoires, l'arbre de décision et les K-plus proches voisins (KNN). Leur analyse comparative a montré que le modèle Random Forest offrait les meilleures performances, avec une précision de 94 %, tandis que le modèle LDA obtenait la précision la plus faible, soit 76 %. Cette étude a mis en évidence l'intérêt de la comparaison de modèles pour identifier les algorithmes prédictifs les plus efficaces pour la MRC.

Dans une autre étude, Xiao et al. [169] ont appliqué neuf algorithmes d'apprentissage automatique (régression logistique, Elastic Net, régression Lasso, régression Ridge, machine à vecteurs de support (SVM), forêt aléatoire, XGBoost, réseau neuronal et KNN) sur un ensemble de données de 551 patients. Parmi ceux-ci, les modèles basés sur la régression, notamment Elastic Net et Lasso, ont démontré une puissance prédictive supérieure avec des précisions respectives de 82 % et 81 %. Cela suggère que les techniques de régularisation peuvent

améliorer considérablement la généralisation des modèles, en particulier lorsqu'il s'agit d'ensembles de données médicales.

Hassan et al. [170] ont proposé un modèle de prédiction de la MRC à un stade précoce en se concentrant sur sept caractéristiques très pertinentes (hémoglobine, gravité spécifique, volume globulaire concentré, albumine, œdème pédieux, créatinine sérique et bilan phosphaté) sélectionnées à l'aide de la technique d'importance des caractéristiques XGBoost. Ils ont évalué plusieurs modèles (réseau neuronal, forêt aléatoire, SVM, arbre aléatoire et arbre de bagging) sur l'ensembles de données optimisées. Le modèle SVM a atteint une précision de 100 % avec les caractéristiques sélectionnées, et le réseau neuronal a obtenu un résultat similaire lorsqu'il a été appliqué à l'ensembles de données du référentiel d'apprentissage automatique de l'UCI, renforçant l'importance de la sélection ciblée des caractéristiques dans les modèles de prédiction clinique.

Pour faire face à la progression de la MRC, Islam et al. [171] ont mis en œuvre douze classificateurs d'apprentissage automatique, dont AdaBoost, Decision Tree, XGBoost, CatBoost, KNN, Random Forest, Gradient Boosting, Stochastic Gradient Boosting, LightGBM, Extra Tree, SVM et ANN. Ces modèles ont été entraînés sur dix caractéristiques cliniquement significatives, telles que l'hémoglobine, l'albumine, la créatinine sérique et les taux de glucose. L'analyse en composantes principales (PCA) a été utilisée pour la sélection des caractéristiques. XGBoost s'est distingué par une précision de 98 %, soulignant sa force dans le traitement des données médicales structurées.

Baidya et al. [172] ont proposé une approche d'apprentissage d'ensembles combinant huit modèles d'apprentissage automatique : Extra Tree, AdaBoost, KNN, Gradient Boosting, XGBoost, Decision Tree, Gaussian Naive Bayes et Random Forest. Ils ont utilisé une stratégie de vote souple pour l'intégration du modèle, qui s'est avérée efficace pour augmenter la fiabilité prédictive. L'étude a fait état de performances élevées, KNN et Extra Tree obtenant les meilleurs résultats avec une précision de 99 %. De plus, le temps d'exécution du modèle hybride a été optimisé à 317 ms, démontrant son applicabilité pratique dans les systèmes de diagnostic en temps réel.

Chittora et al. [151] ont réalisé une comparaison exhaustive de sept classificateurs, dont ANN, C5.0, CHAID, régression logistique, SVM linéaire avec pénalités L1 et L2, et arbre aléatoire, en utilisant un ensemble de données du référentiel UCI. Ils ont appliqué trois techniques de sélection de caractéristiques (basée sur la corrélation, Wrapper et LASSO) pour évaluer leur impact sur les performances du modèle. La précision la plus élevée, 98,86 %, a été obtenue par le classificateur LSVM avec régularisation L2 et toutes les caractéristiques incluses, mettant en évidence l'effet combiné d'un choix de classificateur robuste et d'une utilisation exhaustive des caractéristiques.

Yashfi et al. [173] ont présenté un modèle de prédiction exploitant les algorithmes ANN et Random Forest, utilisant des tests du Chi carré pour identifier les caractéristiques significatives. L'algorithme Random Forest a atteint une précision de 97,12 %. L'étude a souligné les défis posés par la progression silencieuse de la MRC et la nécessité d'utiliser des techniques de réduction des caractéristiques pour améliorer la précision prédictive sans compromettre les connaissances diagnostiques.

Saurabah [155] a présenté un modèle intégrant des classificateurs de base grâce à une stratégie de vote majoritaire afin d'améliorer la prédiction précoce de la MRC. Le processus de classification a été divisé en trois étapes selon les types d'attributs : catégorique, non catégorique et combiné. Le modèle a été évalué à l'aide de SVM, RF et ANN, qui ont donné des précisions respectives de 91 %, 93 % et 89 %. L'amélioration progressive de 3 % de la précision par rapport aux modèles précédents démontre le potentiel des stratégies d'ensembles dans les applications médicales.

Mustafizur Rahman et al. [163] ont abordé les défis sanitaires spécifiques à l'Asie du Sud en concevant un modèle diagnostique robuste utilisant huit classificateurs d'ensembles, dont LightGBM. Pour améliorer la qualité des données, ils ont appliqué l'imputation MICE pour les valeurs manquantes et Borderline-SMOTE pour l'équilibrage des classes. L'élimination récursive des caractéristiques (RFE) et Boruta ont été utilisées pour la sélection des caractéristiques. Leur méthode a atteint une précision remarquable de 99,75 %, apportant de nouvelles stratégies pour le diagnostic de la MRC et soulignant l'importance d'un prétraitement des données personnalisé dans les contextes régionaux.

Ghosh et al. [162] ont évalué plusieurs techniques de classification, notamment SVM, AdaBoost, LDA et Gradient Boosting, à l'aide de l'ensembles de données UCI. Leur analyse a révélé que le Gradient Boosting atteignait la plus grande précision (99,80 %), surpassant les autres modèles et confirmant la pertinence des techniques de boosting pour les tâches de prédiction médicale.

Navaneeth et al.[164] ont présenté une architecture hybride d'apprentissage profond combinant CNN et SVM pour relever les défis de classification dans la prédiction de la MRC. Leur modèle intégrait le clustering dynamique et l'élagage des caractéristiques, appliqués à la concentration d'urée dans les échantillons de salive, un nouveau biomarqueur non invasif de l'IRC. Le modèle CNN-SVM a atteint une précision de 96,51 %, surpassant les modèles CNN standard et offrant une approche diagnostique innovante et moins invasive.

Ebiaredoh-Mienye et al. [174] ont présenté un modèle AdaBoost sensible aux coûts, avec une sélection des caractéristiques basée sur le gain d'information. Cette approche a minimisé le nombre de tests cliniques requis, réduisant ainsi les coûts de dépistage tout en maintenant des performances diagnostiques élevées. Leur modèle a surpassé plusieurs autres, notamment la régression logistique, la RF, la SVM, XGBoost et l'AdaBoost traditionnel, avec une précision de 99,8 %, soulignant l'efficacité de l'apprentissage sensible aux coûts dans le secteur de la santé.

Pal et al. [175] ont développé un modèle de prédiction de la maladie rénale chronique utilisant la régression logistique, l'arbre de décision et le SVM sur un ensembles de données de 400 instances. Parmi eux, l'arbre de décision a atteint une précision de 95,92 %, qui est passée à 97,23 % après l'application du bagging. Cette étude souligne l'utilité des techniques d'ensembles comme le bagging pour améliorer le pouvoir prédictif, même avec des ensembles de données de taille limitée.

Dans cette étude, Hossain et al [176] proposent une méthode avancée pour la prédiction précoce de la maladie rénale chronique (MRC) à l'aide de techniques d'apprentissage automatique et de

modèles d'ensembles. En exploitant deux ensembles de données publiques (CKD15 et CKD21), ils appliquent deux techniques d'encodage : Discrete Encoding et Ranged Encoding pour harmoniser les données et créer deux jeux fusionnés. La phase de prétraitement comprend la normalisation, l'équilibrage des classes par sur-échantillonnage synthétique et la sélection des attributs par cinq méthodes, dont RFECV et la corrélation de Pearson. Les auteurs introduisent un nouveau modèle d'ensembles, le Tri-phase Ensembles, combinant les approches Voting, Bagging et Stacking, ainsi que deux autres modèles en couches : Multi-layer Stacking et Multi-layer Blending. Les résultats expérimentaux montrent que le Tri-phase Ensembles et le Multi-layer Stacking atteignent 99,5 % de précision avec le jeu de données Discrete Merged, tandis qu'AdaBoost atteint 97,5 % avec le Ranged Merged. Pour la validation croisée, la régression logistique obtient 99,5 % de précision avec le jeu discret, et Random Forest ainsi que SVM atteint 100 % avec le jeu de données encodé par plages. Enfin, l'utilisation de la méthode LIME permet d'expliquer les prédictions des modèles, renforçant l'interprétabilité de cette approche. Cette étude souligne l'efficacité des modèles d'ensembles pour la prédiction automatique de la MRC, tout en évoquant des perspectives d'amélioration par l'intégration de données multimodales et temporelles.

Rajib et al. [177] proposer un modèle intelligent de prédiction de la maladie rénale chronique (CKD) basé sur l'application de plusieurs techniques d'apprentissage automatique. Le processus méthodologique inclut un prétraitement des données (gestion des valeurs manquantes, encodage, normalisation), suivi par une phase de sélection des variables à l'aide de diverses méthodes telles que la corrélation, le chi-deux, le seuil de variance, la régression Lasso, Ridge et l'analyse de variance (ANOVA). Les auteurs ont évalué les performances de sept classificateurs : Random Forest (RF), AdaBoost, XGBoost, Gradient Boosting (GB), K-Nearest Neighbours (KNN), Support Vector Machine (SVM) et Decision Tree (DT). Parmi ces modèles, Random Forest et AdaBoost ont obtenu les meilleures performances, atteignant une précision et une AUC de 100 % dans certaines configurations. Les résultats ont été validés à l'aide de plusieurs techniques de validation croisée et d'évaluation croisée des ensembles de données. En complément, une application web en temps réel a été développée pour permettre l'utilisation pratique du modèle, notamment dans les environnements à faibles ressources.

Dans leur étude, Elias et al. [178] proposent une méthodologie de prédiction de la maladie rénale chronique (MRC) reposant sur des techniques d'apprentissage supervisé. Après avoir appliqué une méthode d'équilibrage des classes pour corriger la distribution déséquilibrée des données, les auteurs ont effectué un classement des variables avant d'évaluer plusieurs modèles de machine learning, incluant des approches probabilistes, basées sur les arbres de décision et les méthodes d'ensembles. Les classificateurs testés comprennent le Support Vector Machine (SVM), la régression logistique (LR), la descente de gradient stochastique (SGD), le réseau de neurones artificiels (ANN), le k-Nearest Neighbors (k-NN) et le Rotation Forest (RotF). Parmi ces modèles, le Rotation Forest a montré les meilleures performances, atteignant une AUC de 100 %, et une précision, un rappel, une F-mesure et une exactitude de 99,2 %.

R. Sawhney et al. [179] ont développé des modèles d'intelligence artificielle pour la prédiction et l'évaluation de la maladie rénale chronique, en combinant un Perceptron Multi-Couche (MLP) avec rétropropagation, deux méthodes d'extraction de caractéristiques et trois techniques de sélection de variables. Leur approche a permis d'améliorer la performance des prédictions.

Le réseau de neurones artificiels (ANN) s'est révélé particulièrement performant, obtenant une précision de test parfaite de 100 %, surpassant ainsi la régression logistique et la machine à vecteurs de support, qui ont respectivement obtenu 96 % et 82 %.

Arif et al. [180] ont proposé un modèle prédictif reposant sur l'apprentissage automatique, intégrant des techniques avancées de prétraitement, la sélection de caractéristiques par l'algorithme Boruta et une optimisation des hyperparamètres. Leur méthodologie comprend l'imputation itérative des valeurs manquantes et une technique originale de normalisation des données combinant mise à l'échelle robuste, standardisation z-score et transformation min-max. Grâce à ces améliorations, leur modèle utilisant l'algorithme des k plus proches voisins (k-NN) optimisé par GridSearchCV a atteint une précision de 100%.

Poonia et al. [181] ont conçu une approche basée sur la sélection de caractéristiques pour détecter la maladie rénale chronique, en appliquant plusieurs algorithmes de classification tels que k-NN, ANN, SVM et Naive Bayes. Ils ont utilisé l'élimination récursive de variables (RFE) et le test du chi carré pour identifier les caractéristiques les plus pertinentes. Parmi les modèles testés, la régression logistique optimisée avec les variables sélectionnées par chi carré a obtenu la meilleure performance, avec une précision atteignant 98,75%.

Ilyas et al. [182] ont mis au point un système de diagnostic reposant sur des arbres de décision, en utilisant les algorithmes J48 et Random Forest. Leur étude a souligné l'importance du débit de filtration glomérulaire (DFG), calculé à l'aide de la formule CKD-EPI. Les modèles ont été testés sur différents stades de la maladie, démontrant d'excellentes performances, en particulier lors des phases précoces, avec une précision pouvant atteindre 98 %. Les performances sont conservées élevées même aux stades plus avancés.

3.7 Analyse critique des approches existantes et opportunités d'amélioration

Malgré les avancées considérables de l'intelligence artificielle (IA) dans le domaine médical, son application à la prédiction des maladies chroniques comme la maladie rénale chronique (MRC) reste confrontée à plusieurs défis majeurs. L'un des obstacles les plus importants concerne la qualité et la diversité des données médicales disponibles notamment de la maladie rénale chronique. Ces données sont souvent fragmentées, hétérogènes, incomplètes et limitées en taille, ce qui freine le développement de modèles prédictifs généralisables et fiables. Par exemple, certaines bases de données utilisées pour la MRC ne couvrent qu'un nombre restreint d'échantillons, sans prendre en compte des facteurs essentiels liés au mode de vie, aux déclarations médicales déclarées ou à des indicateurs cliniques cruciaux tels que le débit de filtration glomérulaire (DFG), qui joue pourtant un rôle déterminant dans la prédiction, le diagnostic et la stadification de la MRC. Afin de pallier cette limitation, il devient nécessaire d'utiliser à des techniques d'intelligence artificielle permettant d'enrichir les ensembles de données existantes, notamment par le calcul de variables cliniques manquantes ou dérivées, comme le DFG, et par l'augmentation de l'ensembles de données. Par ailleurs, la confidentialité des données de santé soulève des préoccupations importantes, car l'entraînement de modèles performants nécessite l'accès à des données sensibles, ce qui impose des exigences strictes en matière de sécurité et de conformité éthique. Les biais algorithmiques constituent également un défi majeur : ils peuvent apparaître lorsque certaines catégories de patients sont sous-

représentées dans les données, ce qui peut conduire à des prédictions inexactes ou inéquitables. La complexité des données médicales impose en outre une sélection rigoureuse des caractéristiques et une réduction adéquate de la dimensionnalité pour optimiser la performance des modèles. L'interprétabilité des systèmes d'IA représente une autre barrière importante à leur adoption, car les professionnels de santé doivent pouvoir comprendre et justifier les décisions proposées. Notre étude s'inscrit dans cette dynamique de recherche en proposant une approche expérimentale fondée sur des techniques avancées de sélection de caractéristiques et d'apprentissage automatique, avec pour objectif de contribuer à l'amélioration de la prédiction précoce de la MRC.

3.8 Conclusion

Ce chapitre a permis de mettre en lumière les principales approches développées dans la littérature pour la prédiction de la maladie rénale chronique à l'aide de techniques d'intelligence artificielle, ainsi que les enjeux méthodologiques et cliniques associés à leur mise en œuvre. Bien que les résultats rapportés soient encourageants, de nombreuses limites subsistent, notamment en ce qui concerne la qualité et la représentativité des données utilisées, la présence de biais algorithmiques, la faible interprétabilité des modèles, ainsi que la difficulté d'intégration de ces solutions dans les environnements médicaux réels.

Dans le cas spécifique de la maladie rénale chronique, la rareté des données complètes et normalisées, l'absence de certaines variables cliniques clés comme le DFG dans certains jeux de données, ainsi que le manque d'approches généralisables constituent autant d'obstacles à une prédiction fiable et applicable. Face à ces défis, il apparaît essentiel de concevoir des modèles intelligents capables de tirer le meilleur parti des données disponibles, tout en assurant une rigueur dans le choix des variables et la transparence des prédictions.

Chapitre 4

Contributions et Evaluation

(Partie 1)

4.1 Introduction

La prédiction précoce des maladies chroniques, en particulier la maladie rénale chronique (MRC), constitue un enjeu majeur pour la santé publique en raison de l'impact silencieux et progressif de cette pathologie. Face aux limites des approches traditionnelles, les techniques d'intelligence artificielle, notamment l'apprentissage automatique, offrent des solutions prometteuses pour améliorer le diagnostic précoce et la prise en charge des patients.

Dans ce contexte, plusieurs travaux ont été réalisés, mettant en œuvre différentes méthodes de sélection de caractéristiques et divers modèles d'apprentissage supervisé. Ce chapitre rassemble ces contributions, en présentant de manière synthétique les approches adoptées ainsi que les résultats expérimentaux obtenus. L'objectif est de mettre en évidence la pertinence des modèles explorés et leur performance dans la prédiction de la MRC.

4.2 Contribution 1 : Efficient Chronic Kidney Disease Prediction : A Comparative Analysis Using Feature Selection and Machine Learning models

Dans le cadre de cette contribution, nous avons exploré l'utilisation de techniques d'intelligence artificielle, en particulier l'apprentissage automatique, pour la prédiction précoce de la maladie rénale chronique (MRC), un problème de classification binaire (MRC ou non-MRC). L'objectif principal de notre étude est de proposer une approche efficace permettant d'identifier les modèles les plus performants pour cette tâche de prédiction. Pour ce faire, nous avons utilisé un ensemble de données issues du référentiel UCI, que nous avons préalablement nettoyé à l'aide de techniques d'imputation afin de traiter les valeurs manquantes. Après un processus rigoureux de prétraitement des données, nous avons appliqué trois méthodes de sélection de caractéristiques : LASSO, RELIEF et RFE, afin de limiter le surapprentissage, d'améliorer la précision et de réduire le temps d'apprentissage. Les données ont ensuite été divisées en deux sous-ensembles : 70 % pour l'entraînement et 30 % pour le test. Nous avons entraîné et évalué plusieurs modèles d'apprentissage supervisé, à savoir Random Forest, AdaBoost, Bagging, ainsi qu'un modèle hybride, en analysant leur performance sur l'ensembles des variables ainsi que sur les sous-ensembles sélectionnés.

4.2.1 Description de l'ensembles de données :

L'ensembles de données exploitées dans ce travail provient du UCI Machine Learning Repository[183] et représente, à ce jour, la seule base de données publique largement reconnue pour l'étude de la maladie rénale chronique (Chronic Kidney Disease – CKD). Il se compose de 400 instances de patients, collectés sur une période de deux mois, chacun étant décrit par 24 attributs cliniques et biologiques (dont 11 numériques et 13 catégories) ainsi qu'un attribut cible binaire indiquant la présence ou l'absence de la maladie. Les variables incluent, entre autres, le taux de créatinine, la pression artérielle, la densité urinaire, le taux d'hémoglobine, ou encore la présence d'anémie ou de protéines dans l'urine. L'attribut cible permet de classer les individus en deux catégories : CKD (malade) ou Not-CKD (non malade). Ce jeu de données constitue une référence précieuse pour les travaux de recherche en intelligence artificielle appliquée à la néphrologie, en raison de sa structure complète et de son accessibilité libre. Les données ont été

intégrées dans l'environnement Python à l'aide de la bibliothèque Pandas pour les étapes de prétraitement, d'analyse et de modélisation. La description détaillée des attributs de cet ensembles de données est présentée dans le tableau suivant :

Tableau 4.1: Description des attributs de l'ensembles de données et leurs typologies

N°	Variable	Description	Type	Valeurs possibles
1	age	Âge	Numérique	Âge en années
2	al	Albumine	Nominal	0, 1, 2, 3, 4, 5
3	sg	Densité urinaire	Nominal	1.005, 1.010, 1.015, 1.020, 1.025
4	bp	Pression artérielle	Numérique	En mmHg
5	su	Sucre	Nominal	0, 1, 2, 3, 4, 5
6	rbc	Globules rouges	Nominal	Normal, anormal
7	pc	Cellules de pus	Nominal	Normal, anormal
8	pcc	Amas de cellules de pus	Nominal	Présent, non présent
9	ba	Bactéries	Nominal	Présent, non présent
10	bgr	Glycémie aléatoire	Numérique	En mg/dl
11	bu	Urée sanguine	Numérique	En mg/dl
12	sc	Créatinine sérique	Numérique	En mg/dl
13	sod	Sodium	Numérique	En mEq/L
14	pot	Potassium	Numérique	En mEq/L
15	hemo	Hémoglobine	Numérique	En g/dl
16	pcv	Volume globulaire moyen	Numérique	-
17	wc	Numération globulaire blanche	Numérique	En cellules/mm³
18	rc	Numération globulaire rouge	Numérique	En millions/mm³
19	htn	Hypertension	Nominal	Oui, non
20	dm	Diabète sucré	Nominal	Oui, non
21	cad	Maladie coronarienne	Nominal	Oui, non

22	appet	Appétit	Nominal	Bon, mauvais
23	pe	Œdème des membres inférieurs	Nominal	Oui, non
24	ane	Anémie	Nominal	Oui, non
25	class	Classe (Classification)	Nominal	ckd, notckd

4.2.2 Visualisation de la distribution de l'ensembles de données :

Dans le cadre de l'analyse exploratoire du jeu de données, il est essentiel de commencer par examiner la répartition de la variable cible. Celle-ci représente les deux états de santé étudiés : les patients atteints de la maladie rénale chronique (CKD) et ceux non atteints (Not-CKD). La figure ci-dessous illustre la répartition de cette variable, mettant en évidence un déséquilibre modéré entre les deux classes. En effet, sur les 400 instances disponibles, 250 correspondantes aux cas de CKD et 150 aux cas Not-CKD. Cette répartition est un facteur important à prendre en compte dans l'évaluation des performances des modèles de classification, notamment pour éviter les biais liés aux classes majoritaires. Une bonne compréhension de cette distribution permet d'anticiper les besoins en techniques de rééchantillonnage ou en métriques adaptées (comme le score F1 ou la matrice de confusion) lors de la phase d'évaluation des modèles prédictifs.

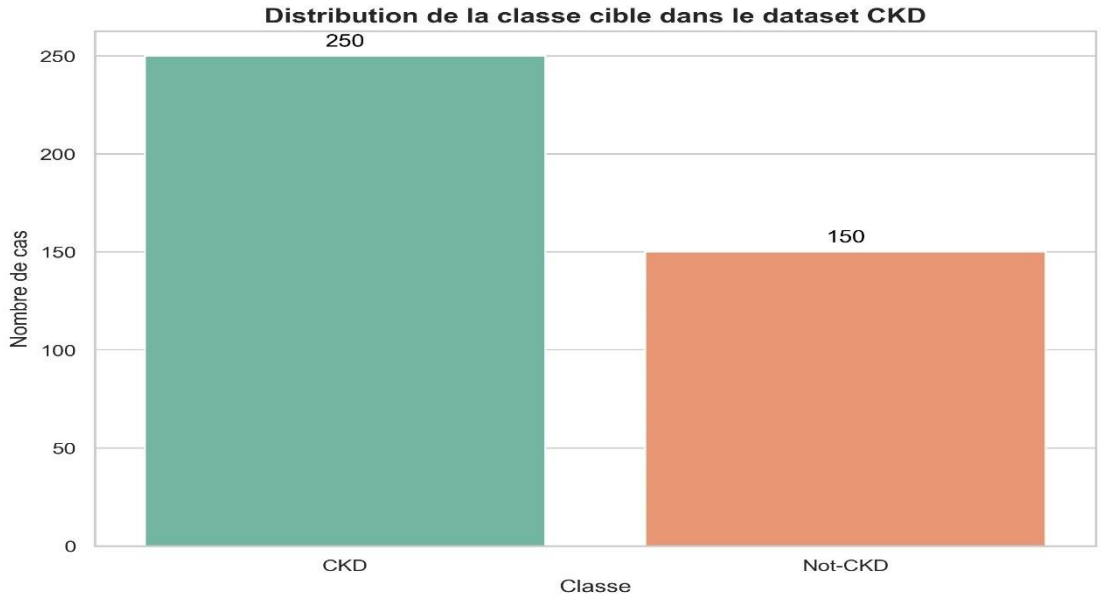


Figure 4.1: Distribution de la classe cible (Classification)

Afin de mieux comprendre la répartition des variables contenues dans le jeu de données utilisées pour cette étude, une série d'histogrammes a été générée pour illustrer la distribution des principales caractéristiques numériques, telles que l'âge, la pression artérielle, les paramètres biochimiques (urée, créatinine sérique, sodium, potassium, etc.) ainsi que les valeurs hématologiques (hémoglobine, volume globulaire moyen, etc.). Cette visualisation permet d'identifier d'éventuelles asymétries, la présence de valeurs extrêmes, ainsi que les tendances

générales de chaque variable. Une bonne compréhension de ces distributions est essentielle pour guider les étapes suivantes du prétraitement, notamment la gestion des valeurs manquantes et la normalisation des données. La figure ci-dessous présente les histogrammes correspondants.

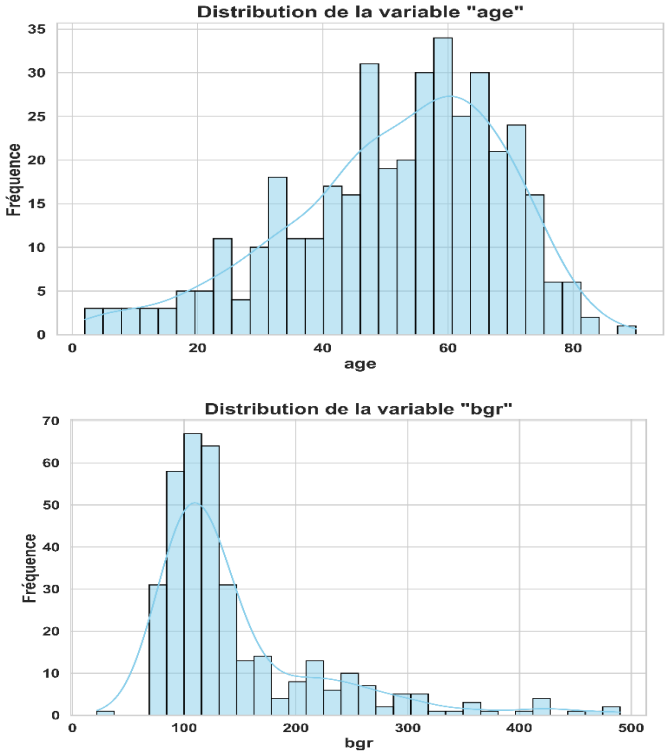


Figure 4.2: Distribution des variables l'Age et la glycémie

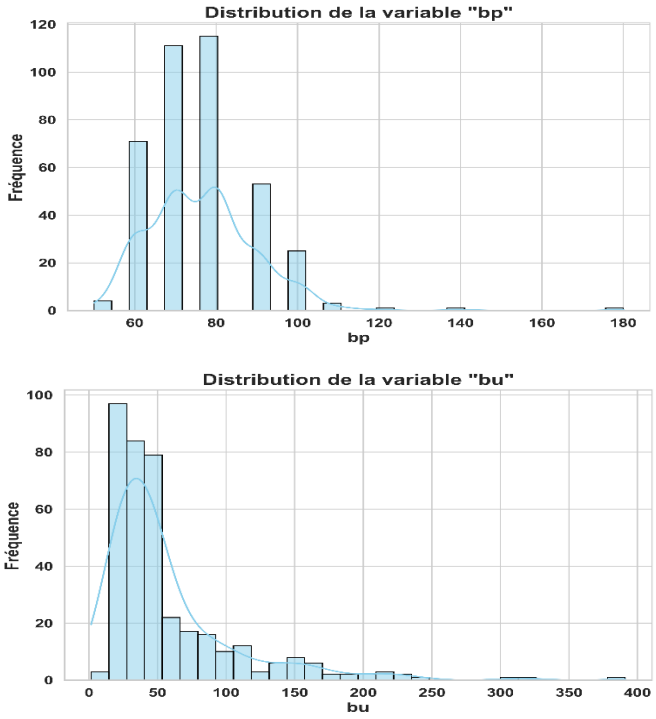


Figure 4.3: Distribution des variables Pression artérielle et Urée sanguine

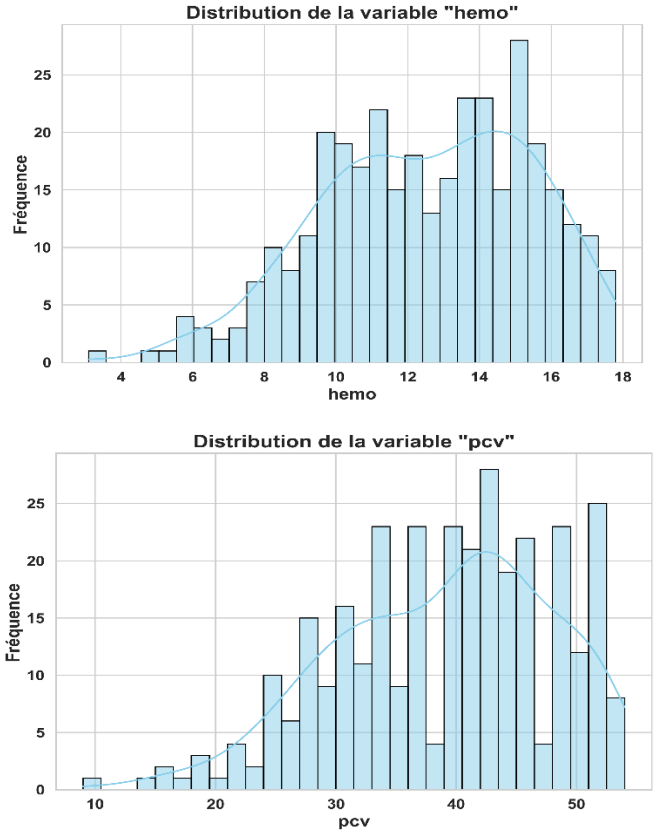


Figure 4.4: Distribution des variables l'hémoglobine et Volume globulaire

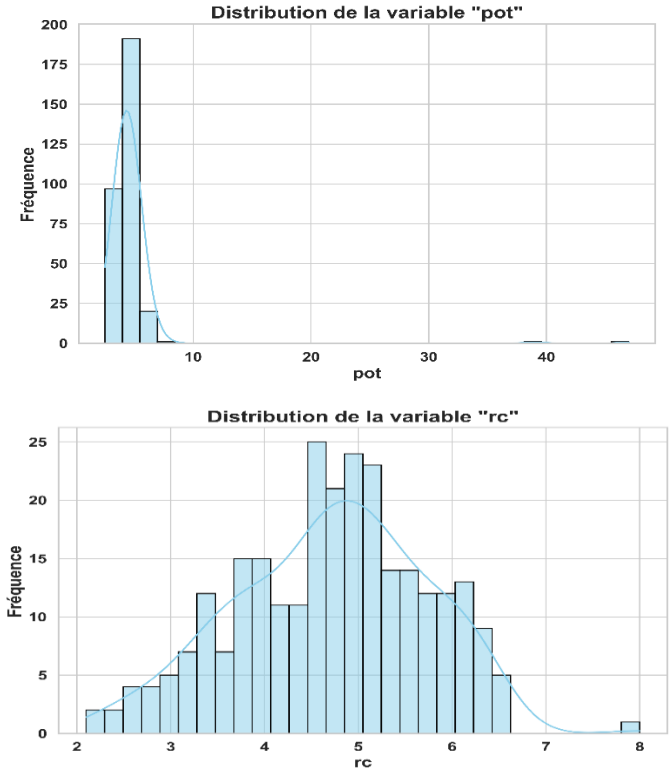


Figure 4.5: Distribution des variables potassium et globules rouges

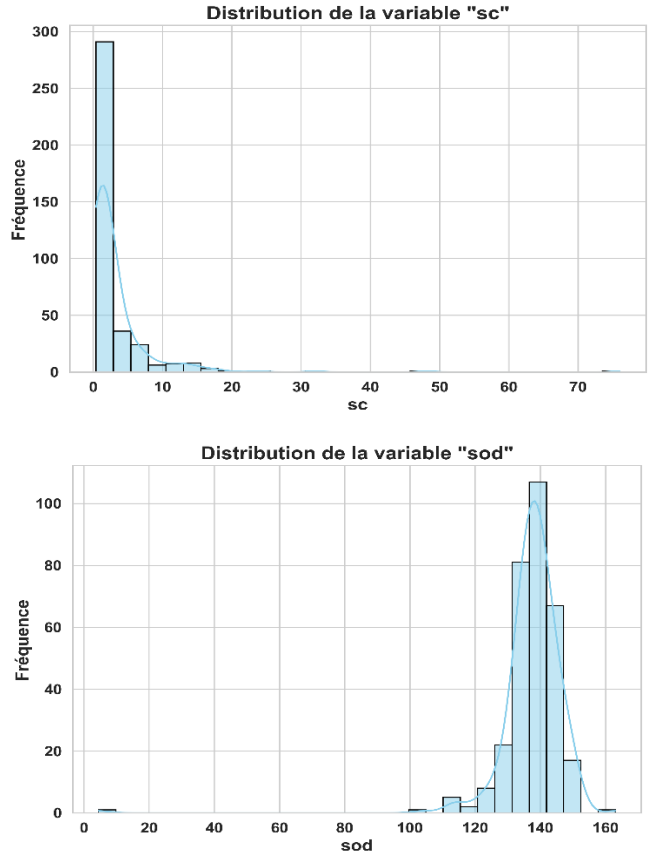


Figure 4.6: Distribution de Créatinine sérique et sodium

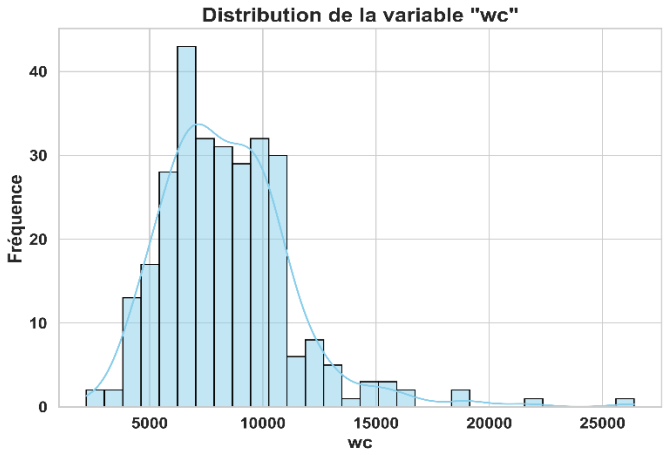


Figure 4.7: Distribution des variables globule blanche

4.2.3 Corrélation entre les variables

Pour explorer les relations linéaires entre les variables numériques de notre ensemble de données, nous avons eu recours à une carte de corrélation (heatmap). Cette visualisation permet de mesurer l'intensité et la direction des associations entre chaque paire de variables. Chaque cellule de la matrice représente le coefficient de corrélation de Pearson, dont les valeurs varient entre -1 et 1 : une valeur proche de 1 indique une forte corrélation positive, tandis qu'une valeur

proche de -1 traduit une forte corrélation négative. En revanche, une valeur proche de 0 suggère une absence de relation linéaire.

L'analyse de cette matrice de corrélation est essentielle pour identifier les redondances entre variables indépendantes, ce qui peut impacter négativement les performances des modèles prédictifs, en particulier ceux sensibles à la multicolinéarité. Dans le cas contraire, une forte corrélation entre une variable indépendante et la variable cible est généralement souhaitable, car elle indique un fort pouvoir explicatif. Ainsi, cette étape permet de guider les choix en matière de sélection ou de réduction des variables.

Dans notre étude, la carte de corrélation révèle, par exemple, une forte corrélation positive entre l'hémoglobine (*hemo*), le volume globulaire moyen (*pcv*) et le nombre de globules rouges (*rc*), ce qui est cohérent sur le plan médical. En revanche, des corrélations négatives notables apparaissent entre la créatinine sérique (*sc*) et certains paramètres comme le sodium (*sod*) ou l'hémoglobine, ce qui reflète une dégradation de la fonction rénale. Ce type de visualisation permet également de repérer d'éventuelles anomalies ou variables peu pertinentes, facilitant ainsi l'élaboration d'un modèle plus robuste et plus interprétable. La figure ci-dessous illustre la matrice de corrélation des principales variables numériques de l'ensembles de données, mettant en évidence les relations linéaires positives ou négatives entre elles à l'aide d'un code couleur intuitif.

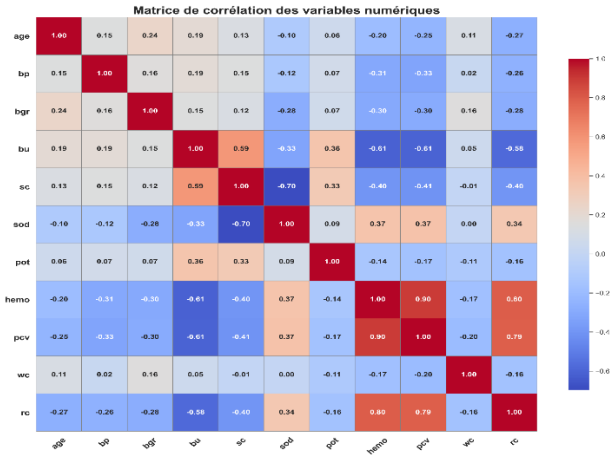


Figure 4.8: Carte de corrélation des variables numériques de l'ensembles de données

4.2.4 Prétraitement des données :

L'étape de prétraitement constitue une phase fondamentale dans toute étude basée sur l'apprentissage automatique, car la qualité des données détermine en grande partie la performance des modèles. Dans notre étude, nous avons porté une attention particulière au traitement des valeurs manquantes et des valeurs aberrantes, étant donné leur impact significatif sur les résultats.

Le jeu de données initial comprenait un certain nombre de valeurs manquantes, notamment dans des attributs tels que les globules rouges (151 valeurs manquantes), la densité urinaire (47), ou encore le sodium (85). Ignorer ces lacunes ou supprimer les lignes incomplètes aurait entraîné une perte considérable d'informations précieuses, compromettant ainsi la robustesse de

notre modèle. Par conséquent, nous avons adopté une stratégie rigoureuse de traitement de ces données absentes.

Pour les variables catégorielles, telles que *rbc* (globules rouges), *pc* (cellules de pus), *htn* (hypertension), *dm* (diabète), *cad* (maladie coronarienne), *pe* (œdème), *ane* (anémie), et *appétit* (appétit), nous avons appliqué une imputation par la valeur modale. Cette méthode consiste à remplacer les valeurs manquantes par la modalité la plus fréquemment mentionnée dans chaque variable. Par exemple, si la majorité des patients présentent une pression artérielle normale (*htn* = *no*), cette modalité est utilisée pour combler les valeurs manquantes de cette variable. Cette approche est particulièrement pertinente pour les variables nominales, car elle permet de maintenir la structure des classes et de minimiser l'introduction de biais, tout en entraînant la suppression d'instances précieuses.

En ce qui concerne les variables numériques, comme l'âge, la pression artérielle (*bp*), la créatinine sérique (sc), l'hémoglobine (hemo), ou encore la concentration de sodium (*sod*) et de potassium (*pot*), nous avons opté pour une imputation par la moyenne arithmétique. Chaque valeur manquante a ainsi été remplacée par la moyenne de la variable concernée, calculée à partir des valeurs disponibles. Cette méthode, bien qu'élémentaire, est particulièrement efficace lorsque les données sont distribuées de manière relativement symétrique, sans forte asymétrie ni concentration de valeurs extrêmes. Elle présente l'avantage de préserver la tendance centrale des données sans affecter excessivement la variance globale du jeu de données.

Le choix de ces méthodes d'imputation repose sur leur simplicité, leur efficacité et leur adaptabilité aux types de données rapprochées. Elles permettent d'obtenir un ensemble de données complet, cohérent, et immédiatement exploitable pour les étapes ultérieures du processus de modélisation.

Une fois les valeurs manquantes imputées, une étude approfondie des valeurs aberrantes a été retenue. Ces valeurs extrêmes peuvent être dues à des erreurs de saisie, mais dans un contexte médical, elles peuvent également représenter des cas cliniques significatifs. Afin d'éviter d'exclure des informations utiles, nous avons opté pour une stratégie d'imputation conditionnelle : les valeurs situées en dehors des bornes définies par la méthode de l'intervalle interquartile (IQR) (Q1 - 1.5 IQR et Q3 + 1.5 IQR) ont été remplacées par la moyenne de la variable correspondante. Cette méthode a été répétée de manière itérative jusqu'à ce que toutes les variables soient stabilisées, c'est-à-dire sans nouvelles détections d'anomalies.

Enfin, un encodage par étiquetage (Label Encoding) a été appliqué sur toutes les variables catégorielles afin de les transformer en valeurs numériques, condition nécessaire à l'apprentissage par les algorithmes utilisés.

Ces différentes étapes de préparation ont permis d'obtenir un jeu de données complet, homogène et nettoyé, prêt à être exploité par les modèles d'apprentissage automatique et de Deep Learning. Ce soin apporté au nettoyage des données représente l'un des points forts majeurs de notre étude, renforçant la fiabilité des résultats obtenus lors de la phase de modélisation.

4.2.5 Méthode de sélection des caractéristiques :

Afin d'optimiser la performance prédictive de notre modèle pour la prédiction précoce de la maladie rénale chronique (MRC), nous avons appliqué la méthode Recursive Feature Elimination (RFE), qui constitue l'une des techniques de sélection des variables les plus efficaces dans le domaine de l'apprentissage automatique supervisé. Cette approche vise à identifier les variables les plus informatives pour la tâche de classification, tout en éliminant progressivement les caractéristiques les moins pertinentes. RFE repose sur un processus itératif qui consiste à entraîner un modèle (généralement un classificateur linéaire ou basé sur des arbres), à attribuer un score d'importance à chaque variable, puis à supprimer celles ayant les scores les plus faibles. Ce processus est répété jusqu'à ce qu'un nombre optimal de variables soit conservé, améliorant ainsi la robustesse, la lisibilité et l'efficacité calculée du modèle [184].

L'application de la méthode RFE sur notre jeu de données, composée de 400 instances annotées (250 atteintes de MRC et 150 non atteintes), a permis de hiérarchiser les caractéristiques selon leur contribution à la prédiction. Les résultats sont illustrés à la Figure suivante, qui présente l'importance relative des variables sélectionnées, exprimée en scores.

Parmi ces caractéristiques, la densité urinaire spécifique (sg) s'est distinguée comme la plus déterminante, avec un score de 3,25. Ce paramètre reflète la capacité des reins à concentrer ou diluer l'urine, et constitue donc un indicateur direct de la fonction rénale. Sa position en tête de classement souligne son utilité dans le diagnostic précoce de la MRC.

Les variables hémoglobine (hemo) et volume globulaire (pcv) occupent respectivement la deuxième et la troisième position avec des scores de 1,98 et 1,82. Ces indicateurs sont fortement liés à l'état hématologique du patient. La MRC étant souvent associée à une anémie secondaire à la baisse de la production d'érythropoïétine, ces deux paramètres jouent un rôle crucial dans l'évaluation de la gravité de la maladie.

D'autres variables ont également montré une contribution significative, telle que les cellules de pus (pc), la natrémie (sod), ou encore les globules rouges urinaires (rbc). Leur présence ou variation témoigne d'anomalies biologiques fréquentes chez les patients atteints de pathologies rénales.

À l'inverse, certaines variables ont été associées à des scores négatifs, ce qui peut refléter leur rôle comme facteurs de risque ou comme indicateurs inverses dans la classification. Parmi celles-ci figurent notamment l'albuminurie (al) avec -1.72, la créatinine sérique (sc) avec -1.53 , et le diabète sucré (dm) avec -1.82. Ces résultats sont cliniquement cohérents : l'élévation de la créatinine et la présence d'albumine dans l'urine sont des signes caractéristiques d'une altération de la fonction rénale, tandis que le diabète constitue une des causes principales de la MRC à l'échelle mondiale.

Enfin, des caractéristiques comme l'hypertension artérielle (htn), l'appétit (appet) ou encore les cellules leucocytaires (wc) figurent également dans la liste des variables pertinentes, renforçant l'idée que la MRC est une pathologie multifactorielle dont la prédiction nécessite une prise en compte globale des paramètres cliniques, biologiques et métaboliques.

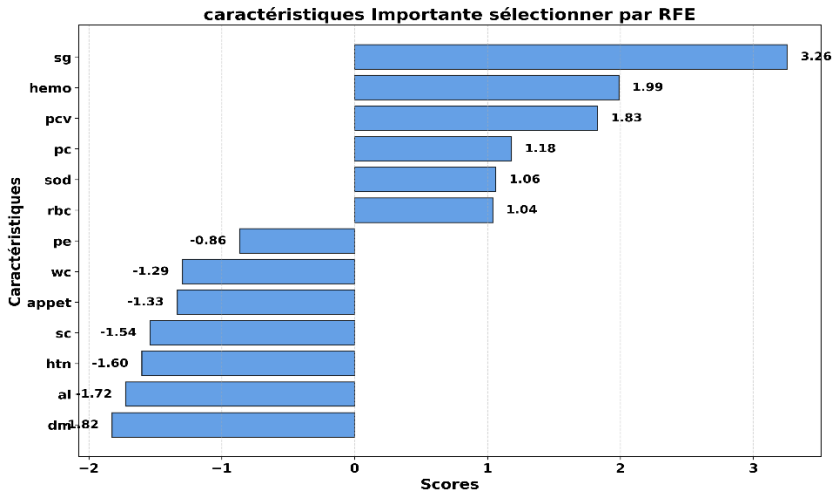


Figure 4.9: Les caractéristiques importantes sélectionnés par RFE et leurs scores

4.2.6 Les algorithmes d'apprentissage automatique utilisés :

Dans notre étude, nous avons opté pour l'utilisation de techniques d'apprentissage automatique supervisé, en raison de la disponibilité d'un ensemble de données étiquetées indiquant l'état clinique des patients (atteints ou non de la maladie rénale chronique). Ce type d'approche est bien adapté aux problèmes de classification binaire, et se révèle particulièrement pertinent dans un contexte médical où la fiabilité des prédictions est cruciale pour soutenir les décisions cliniques.

La sélection d'un algorithme efficace dans un problème de prédiction médicale dépend de nombreux facteurs, notamment la nature des données, leur volume, leur qualité, ainsi que la complexité des relations entre les variables. Dans cette optique, il est essentiel d'évaluer plusieurs méthodes afin d'identifier celles qui s'adaptent mieux aux caractéristiques du problème étudié. En effet, il n'existe pas de solution universelle garantissant de bons résultats dans tous les contextes. Chaque algorithme possède ses propres avantages et limites selon les cas d'utilisation.

Dans cette contribution, nous avons mis en œuvre trois modèles d'ensembles utilisés pour la classification : Random Forest (RF), AdaBoost et Bagging Classifier. Ces techniques sont reconnues pour leur capacité à améliorer la précision en combinant plusieurs modèles de base afin de réduire le risque d'erreurs liées à l'instabilité ou au surapprentissage. Chaque modèle a été réalisé de manière indépendante, puis évalué selon plusieurs métriques afin de comparer leurs performances.

L'évaluation de tous ces modèles a été réalisée à l'aide de techniques rigoureuses (comme la validation croisée et les métriques classiques : exactitude, précision, rappel, F1-score, etc.), afin d'assurer une comparaison équitable et fiable. Cette étape est indispensable pour valider la robustesse des modèles et leur potentiel à être appliqués à des données réelles en contexte clinique.

4.2.7 Modèle hybride proposé :

Afin d'améliorer davantage la performance de la classification et de tirer profit de la complémentarité entre les algorithmes d'ensembles testés, nous avons proposé un modèle hybride basé sur une stratégie de vote. Ce dernier repose sur la combinaison des prédictions issues de trois classificateurs performants individuellement, à savoir : Random Forest (RF), AdaBoost, et Bagging Classifier.

Ce type de combinaison est connu sous le nom de Voting Classifier, une technique simple mais efficace, qui consiste à agréger les prédictions des modèles de base selon une règle de majorité (vote dur) ou en prenant la moyenne pondérée des probabilités de classe (vote souple). Dans notre cas, l'approche soft vote a été retenue, car elle permet de tenir compte du degré de confiance de chaque classificateur pour chaque prédiction. Cette méthode présente plusieurs avantages. D'une part, elle évite les redondances computationnelles associées aux structures à plusieurs couches (comme les modèles empilés complexes). D'autre part, elle améliore la robustesse du modèle final en tirant partie des forces spécifiques de chaque classificateur. Par exemple, tandis que Bagging est efficace pour réduire la variance, AdaBoost permet de corriger les erreurs sur des exemples difficiles, et Random Forest combine stabilité et capacité à gérer les données bruitées.

L'intégration de ces modèles dans une architecture de vote permet donc de stabiliser les prédictions finales, en limitant les erreurs propres à chaque algorithme pris individuellement. Ainsi, le modèle hybride proposé vise une meilleure généralisation, critère essentiel dans le domaine médical où les décisions issues de ces systèmes doivent être fiables et interprétables. L’évaluation du modèle hybride a montré une signification en termes de précision globale, ainsi qu'une meilleure sensibilité dans la prédiction des cas positifs de maladie rénale chronique, comparée aux modèles pris séparément.

4.2.8 Résultats expérimentaux de la contribution 1 :

Dans cette première contribution, nous avons évalué l'efficacité de plusieurs modèles d'apprentissage automatique appliqués à un sous-ensembles de caractéristiques sélectionnées à l'aide de la méthode RFE (Recursive Feature Elimination). Treize variables jugées pertinentes ont été retenues pour la prédiction de la maladie rénale chronique, notamment : la densité urinaire spécifique (sg), l'hémoglobine (hemo), le volume cellulaire packé (pcv), les cellules épithéliales (pc), le sodium (sod), les globules rouges (rbc), l'œdème périphérique (pe), le nombre de globules blancs (wc), l'appétit (appet), la créatinine sérique (sc), l'hypertension (htn), l'albumine (al) et le diabète sucré (dm).

L'expérimentation a été réalisée sous Jupyter Notebook avec la bibliothèque scikit-learn, sur un ensemble de données structurées et prétraitées selon les bonnes pratiques. Les performances des modèles ont été mesurées à l'aide de métriques standard en classification binaire, à savoir : la précision (precision), le rappel (recall), le score F1 et l'exactitude (accuracy). Une validation croisée a été employée afin de garantir la robustesse des résultats.

Tableau 4.2: Résultats de performance de chaque modèle

Métrique d'évaluation	**Modèles de ML**			
	Random forest	AdaBoost	Bagging	Modèle hybride
Accuracy	0.985	0,978	0,986	0,993
Precision	0,990	0,985	0,992	1.000
Recall	0,980	0,970	0,981	0.995
F1_Score	0,985	0,977	0,986	0.997

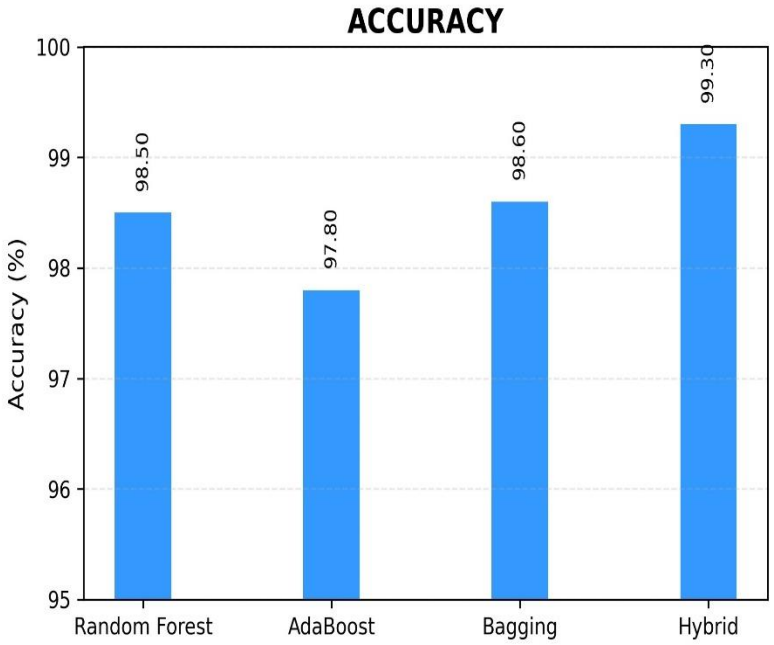


Figure 4.10: Résultats de l'accuracy de chaque modèle

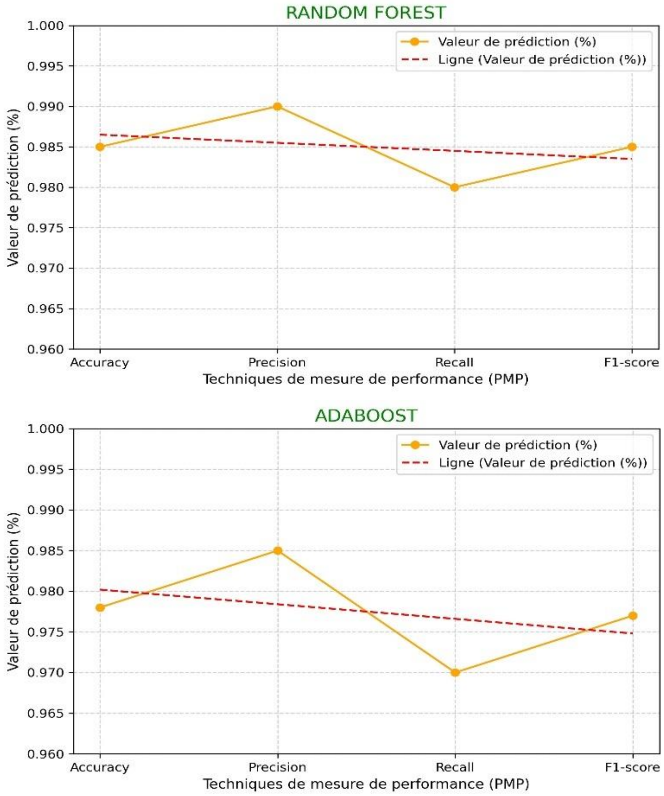


Figure 4.11: Analyse des résultats de la prédiction de la maladie rénale chronique à l'aide de RF et AdaBoost

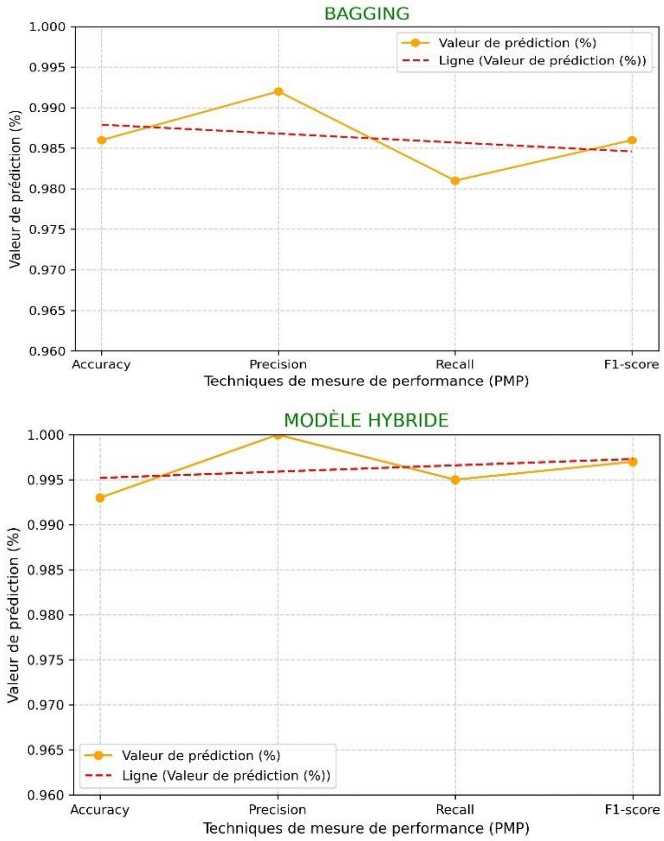


Figure 4.12: Analyse des résultats de la prédiction de la maladie rénale chronique à l'aide de Bagging et le modèle hybride

4.2.9 Discussion des résultats de la contribution 1

En analysant les résultats expérimentaux présentés dans la figure suivante**,** nous constatons que le modèle hybride proposé a surpassé les autres modèles de classification testés, avec une exactitude remarquable de 99,3%, comparée à Bagging (98,6%), Random Forest (98,5%) et AdaBoost (97,8%). Cette supériorité est en grande partie grâce à l'intégration d'une stratégie de sélection des caractéristiques efficace, à savoir la méthode RFE (Recursive Feature Elimination).

L'utilisation de RFE a permis de réduire le nombre de variables en conservant uniquement les attributs les plus pertinents pour la prédiction, ce qui a contribué à éliminer le bruit, éviter le sur-apprentissage (overfitting) et améliorer la capacité de généralisation des modèles. Grâce à cette étape de sélection, les modèles ont été entraînés sur un espace de caractéristiques optimisées, ce qui a significativement renforcé leurs performances, notamment en termes de rappel et de précision.

Le modèle hybride a ainsi obtenu une précision de 100%, un rappel de 99,5%, et un score F1 de 99,7%, traduisant un excellent équilibre entre la prédiction correcte des cas positifs et la minimisation des faux positifs. Les autres modèles, bien que performants, restent légèrement en retrait, avec notamment AdaBoost qui affiche les performances les plus faibles de l'étude.

Enfin, cette première contribution montre que l'association de techniques de sélection de variables comme RFE avec des approches d'ensembles et hybrides permet d'atteindre des niveaux de performance élevés, particulièrement adaptés aux problématiques sensibles de prédiction précoce dans le domaine médical. Le gain observé souligne également l'importance d'une étape de prétraitement rigoureux dans tout processus de modélisation basé sur des données médicales.

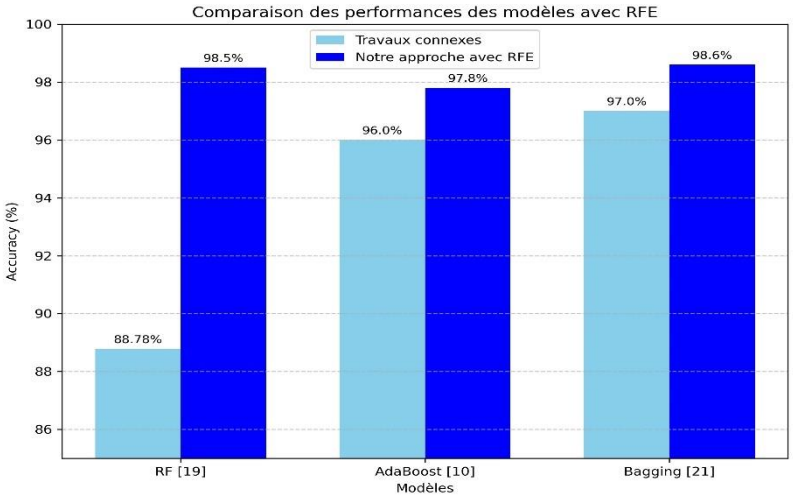


Figure 4.13: Comparaison des résultats de la contribution 1 avec des travaux connexes

4.3 Contribution 2 : Prédiction précoce de la néphropathie chronique à l'aide de la sélection de caractéristiques et de l'apprentissage

Dans cette deuxième contribution, nous avons développé une approche prédictive avancée visant à détecter précocement la néphropathie chronique (CKD), en tirant profit d'algorithmes d'apprentissage automatique supervisé associés à une méthode rigoureuse de sélection de caractéristiques. Deux ensembles de données issues de la base UCI ont été exploités :

- Le premier dataset, identique à celui utilisé dans la première contribution (voir section 4.2.1), comprend 400 échantillons et a été utilisé pour la phase d'entraînement ;
- Le deuxième ensembles de données, totalement distinct, contient 202 échantillons et a été utilisé exclusivement pour tester et valider les performances de notre approche [185].

Chaque ensemble de données intègre un panel de variables cliniques et biologiques pertinentes, incluant les principaux facteurs de risque associés à la MRC. Les caractéristiques détaillées de ces deux ensembles de données seront présentées dans un tableau afin de mieux illustrer leur contenu et leur complémentarité.

Tableau 4.3: Description de l'ensembles de données de test

Caractéristiques	**Type**
Âge	Catégorielle
Pression artérielle diastolique (bp)	Numérique
Limite de pression artérielle (limite bp)	Numérique
Densité (sg)	Catégorielle
Albumine (al)	Catégorielle
Sucre (su)	Catégorielle
Globules rouges (rbc)	Catégorielle
Cellule de pus (pc)	Numérique
Amas de cellules pusiques (pcc)	Numérique
Bactéries (ba)	Numérique
Glycémie aléatoire (bgr)	Catégorielle
Urée sanguine (bu)	Catégorielle
Créatinine sérique (sc)	Catégorielle

Sodium (sod)	Catégorielle
Potassium (pot)	Catégorielle
Hémoglobine (hémo)	Catégorielle
Volume globulaire total (pcv)	Catégorielle
Numération des globules blancs (wbcc)	Catégorielle
Numération des globules rouges (rbc)	Numérique
Hypertension artérielle (htn)	Numérique
Diabète sucré (dm)	Numérique
Maladie coronarienne (cad)	Numérique
Appétit (aappet)	Numérique
Œdème des pieds (pe)	Numérique
Anémie (ane)	Numérique
Stade (stage)	Catégorielle
Débit de filtration glomérulaire (grf)	Catégorielle
Affecté (afdected)	Numérique
Classe (étiquette)	Binaire (0 ou 1)

4.3.1 Méthode de sélection des caractéristiques utilisé dans la contribution 2 :

Dans le but d'améliorer la précision des modèles prédictifs tout en définissant leur complexité, nous avons eu recours à une technique de sélection automatique des caractéristiques, à savoir l'algorithme LASSO (Least Absolute Shrinkage and Selection Operator). Cette méthode repose sur la régularisation L1, qui consiste à contraindre la somme des valeurs absolues des coefficients du modèle, ce qui pousse certains d'entre eux à devenir exactement nuls. Ainsi, les variables jugées non pertinentes sont automatiquement écartées du modèle, tandis que les plus influentes sont conservées.

L'intérêt de cette méthode réside dans sa capacité à identifier les attributs les plus informatifs pour la tâche de classification, notamment dans des contextes où certaines variables présentent un pouvoir discriminant limité. L'utilisation de LASSO permet donc à la fois de réduire la dimensionnalité des données et de prévenir le surapprentissage, tout en conservant une bonne

interprétabilité des résultats. Il est à noter que des variantes de LASSO, comme le LASSO aléatoire (randomized LASSO), permettent d'améliorer la stabilité des sélections en répétant le processus plusieurs fois et en retenant les variables les plus fréquemment sélectionnées.

L'analyse des résultats a mis en évidence les caractéristiques les plus influentes dans la prédiction de la néphropathie chronique. Parmi celles-ci, la variable "sg" (densité urinaire) présente le coefficient le plus élevé avec une valeur positive de 2.009, indiquant une forte association directe avec la présence de la maladie. D'autres caractéristiques telles que "hemo" (hémoglobine) et "pcv" (volume cellulaire emballé) affichent également des coefficients positifs, traduisant leur impact significatif dans la prédiction. En revanche, des variables comme "al" (albumine), "sc" (créatinine sérique), "htn" (hypertension), "dm" (diabète) et "appet" (appétit) présentent des coefficients négatifs, ce qui suggère une relation inverse avec le résultat cible. L'ensembles des variables sélectionnées et leurs coefficients respectifs sont illustrés dans la Figure suivante, mettant en évidence le rôle crucial de certaines caractéristiques biologiques dans le processus de prédiction précoce de la CKD.

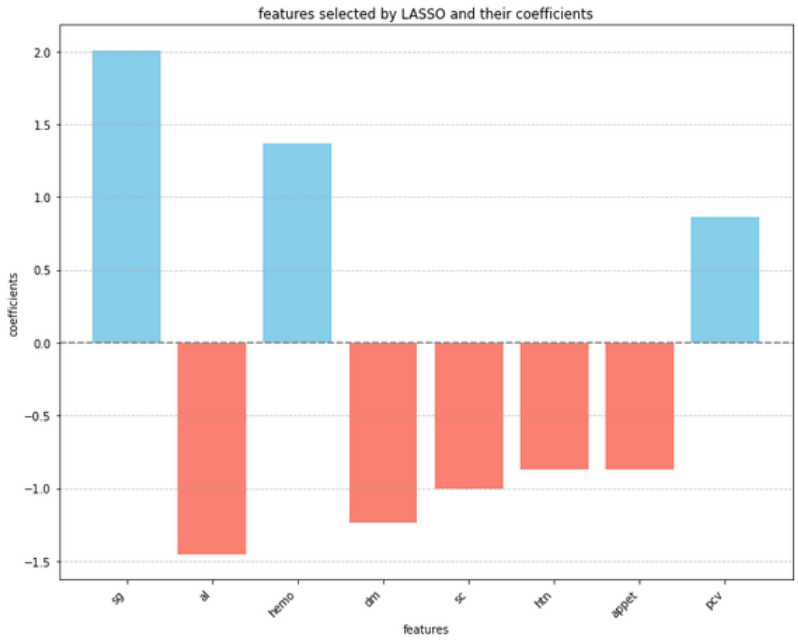


Figure 4.14: Les caractéristiques importantes sélectionnées par la méthode LASSO et leurs coefficients

4.3.2 Modélisation

À l'issue de la phase de sélection des caractéristiques, deux algorithmes d'apprentissage supervisé, à savoir K-Nearest Neighbours (KNN) et Gradient Boosting, ont été appliqués à l'ensembles des données traitées. Ces modèles ont été choisis en raison de leur efficacité démontrée dans des contextes médicaux complexes, et sont décrits en détail dans la section dédiée à l'état de l'art (voir chapitre 3).

Le jeu de données d'entraînement a été divisé en deux sous-ensembles : 80 % pour l'apprentissage des modèles et 20 % pour la validation croisée. Les performances ont été réalisées à la fois avec et sans sélection de caractéristiques (via LASSO), afin de mesurer

l'impact de cette étape sur la qualité des prédictions. Ensuite, les modèles obtenus ont été testés sur un deuxième ensemble de données indépendant contenant 202 échantillons, dans le but de valider la robustesse et la généralisation des résultats. Chaque modèle a été évalué selon plusieurs métriques de performance : précision (accuracy), rappel (recall), F1-score, précision et aire sous la courbe ROC (AUC). Cette étape a permis d'identifier le classificateur le plus performant pour la prédiction précoce de la néphropathie chronique.

4.3.3 Résultats expérimentaux de la contribution 2 :

L'évaluation expérimentale a été menée afin d'examiner l'efficacité de deux classificateurs supervisés, à savoir K-Nearest Neighbours (KNN) et Gradient Boosting (GB), appliqués à la prédiction précoce de la néphropathie chronique. Pour garantir une comparaison objective, les modèles ont été testés selon deux configurations : d'une partie, en utilisant l'ensembles complet des caractéristiques issues du jeu de données, et d'autre partie, en appliquant une sélection préalable des variables par la méthode LASSO, destinée à ne conserver que les attributs les plus informatifs.

Les résultats obtenus démontrent sans équivoque l'intérêt de la sélection de caractéristiques. En effet, lorsque LASSO est appliqué en amont de la phase de modélisation, le classifieur KNN atteint une exactitude de 97,5 %, un rappel parfait de 100 %, une précision de 94,1 %, et un score F1 de 96,9 %. Ce niveau de performance indique que le modèle est capable d'identifier tous les patients atteints, sans pour autant nuire à sa fiabilité par une génération excessive de faux positifs.

Tableau 4.4: Évaluation des modèles avec et sans sélection des caractéristiques (LASSO)

Modèle	Accuracy avec LASSO	Précision avec LASSO	Rappel avec LASSO	F1 Score avec LASSO	Accuracy avec toutes les fonctionnalités
KNN	97,5%	94.1%	100%	96.9%.	89%
Gradient Boosting	100%	100%	100%	100%	90,3%

Les performances obtenues par les deux modèles avec la sélection de caractéristiques via LASSO sont illustrées dans la figure suivante.

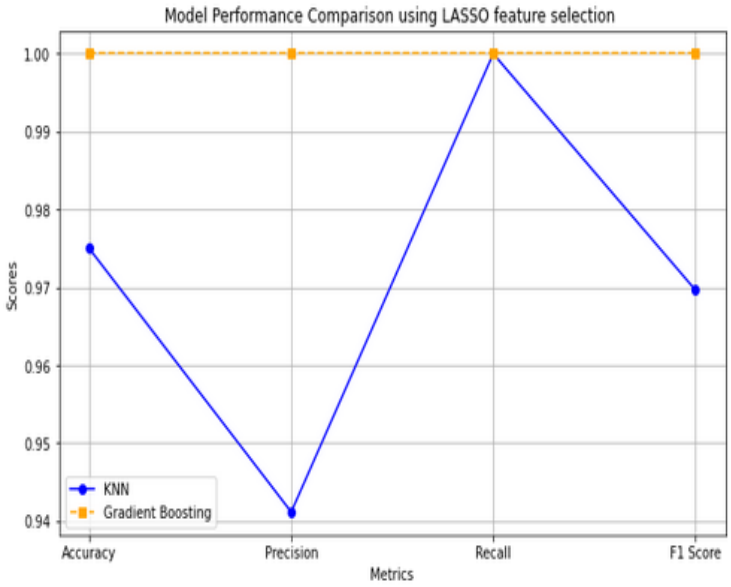


Figure 4.15: Performances des modèles utilisant la sélection de fonctionnalités LASSO

La courbe d'apprentissage du modèle KNN est présentée ci-dessous afin d'illustrer son comportement lors de l'entraînement.

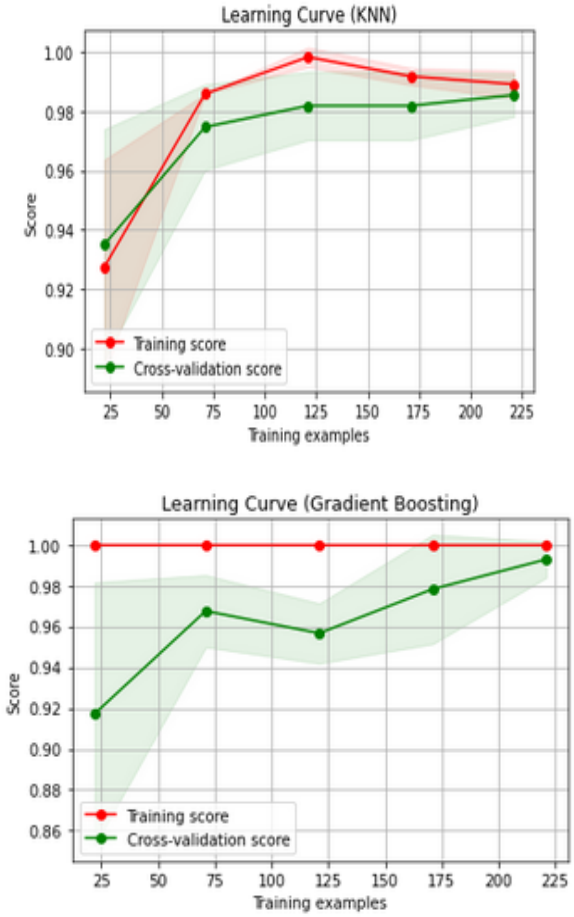


Figure 4.16: Courbe d'apprentissage pour KNN et Gradient Boosting

Le modèle Gradient Boosting, quant à lui, se distingue par des performances parfaites sur l'ensembles des indicateurs, atteignant 100 % pour l'exactitude, la précision, le rappel et le score

F1. Ce résultat témoigne d'une capacité de généralisation remarquable, confirmée par une aire sous la courbe ROC (AUC) de 1.00, traduisant une séparation totale entre les classes prédictives. Le modèle KNN présente également une AUC très élevée de 0.99, ce qui atteste de sa robustesse.

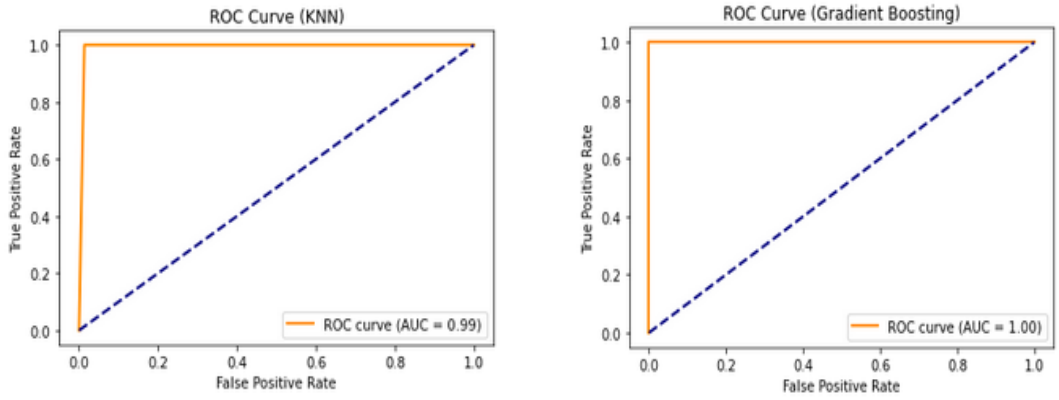


Figure 4.17: Courbe ROC des modèles KNN et Gradient Boosting

À titre comparatif, l'utilisation de l'ensembles des caractéristiques sans sélection préalable entraîne une baisse sensible des performances : l'exactitude du KNN tombe à 89 %, tandis que celle du Gradient Boosting diminue à 90,3 %. Cette observation souligne l'importance cruciale de la réduction dimensionnelle dans les systèmes prédictifs appliqués à des données médicales souvent bruitées ou redondantes. Ces résultats révèlent que la combinaison de la sélection de caractéristiques via LASSO et du modèle Gradient Boosting constitue une approche particulièrement efficace pour la prédiction précoce de la MRC. Par ailleurs, les performances obtenues laissent entrevoir un potentiel réel d'intégration dans des outils d'aide à la décision clinique, permettant une identification rapide et fiable des individus à risque, et contribuant ainsi à une prise en charge médicale plus réactive et personnalisée.

4.3.4 Discussion

Les résultats obtenus dans cette contribution ont été comparés à ceux de plusieurs travaux récents portant sur la prédiction de la néphropathie chronique à l'aide de techniques d'apprentissage automatique. L'objectif de cette analyse comparative est de situer l'apport de notre approche dans le contexte des recherches existantes, tant en termes de performances que de méthodologie.

Le classifieur K-Nearest Neighbours (KNN) a atteint une précision de 97,5 %, tandis que le modèle Gradient Boosting a obtenu un score parfait de 100 % sur l'ensembles de test. Ces performances dépassent nettement celles rapportées dans la littérature. Par exemple, Devika et al.[186] ont obtenu une précision de 87,87 % avec KNN, tandis que Xioa et al.[169] n'ont pas dépassé 74 % avec le même algorithme. Du côté de Gradient Boosting, les résultats précédemment publiés par Baidya et al.[171] et Hema et al.[187] se situaient respectivement à 98 % et 97 %.

Un autre aspect distinctif de notre étude réside dans l'utilisation conjointe de deux ensembles de données séparées pour l'entraînement et la validation, contrairement à la plupart des travaux

précédents qui utilisent un seul ensemble de données partagées entre les deux phases. Ce choix méthodologique renforce la robustesse de l'évaluation et la capacité de généralisation des modèles.

De plus, l'intégration de la méthode de sélection de caractéristiques LASSO a permis d'améliorer de manière significative la performance des modèles. Cette technique a facilité l'élimination des variables redondantes ou peu informatives, tout en conservant les attributs les plus pertinents pour la classification. Son effet bénéfique est clairement observé dans l'augmentation des scores des modèles, notamment sur la précision et l'AUC, et ce même avec un nombre réduit de caractéristiques.

En somme, les résultats démontrent que la combinaison de LASSO avec des algorithmes performants tels que KNN et Gradient Boosting permet d'obtenir une meilleure précision que celle rapportée dans les études existantes. Cette approche peut ainsi constituer une base solide pour le développement d'outils intelligents d'aide à la décision clinique, destinés à la prédiction précoce de la MRC.

Les performances comparées sont résumées dans le Tableau suivant qui mettent en évidence la supériorité de notre méthodologie par rapport aux approches classiques.

Tableau 4.5: Comparaison de la performance de nos modèles avec des modèles issus de travaux connexes

Travail	Modèle	Accuracy	Dataset
Devika et al. [186]	KNN	87.87%	1(Pour entrainement et test)
Hema et al. [187]	Gradient Boosting	97%	1 (Pour entrainement et test)
Xioa et al [169]	KNN	74%	1(Pour entrainement et test)
Baidya et al [171]	Gradient Boosting	98%	1 (Pour entrainement et test)
Notre approche	KNN Gradient Boosting	97,5% 100%	2 (pour entrainement test et validation)

4.4 Conclusion

Ce chapitre a présenté deux contributions majeures consacrées à l'amélioration de la prédiction précoce de la maladie rénale chronique (MRC), en exploitant des approches avancées d'apprentissage automatique et de sélection de caractéristiques.

Dans la première contribution, nous avons réalisé une étude comparative approfondie entre plusieurs classificateurs supervisés (Random Forest, AdaBoost, Bagging, et un modèle hybride), combinés à différentes techniques de sélection de variables, notamment RFE, LASSO et RELIEF. Les expérimentations ont révélé que le modèle hybride associé à RFE offrait les meilleures performances, atteignant une exactitude de 99,3 %, une précision de 100 %, un rappel de 99,5 % et un score F1 de 99,7 %. Ces résultats démontrent que l'intégration d'une

sélection rigoureuse des caractéristiques peut améliorer significativement les performances des modèles, tout en précisant le risque de surapprentissage.

La deuxième contribution s'est appuyée sur la méthode LASSO pour affiner la sélection des variables les plus discriminantes, avant d'entraîner deux classificateurs performants, à savoir K-Nearest Neighbours et Gradient Boosting. L'évaluation sur un deuxième ensembles de données indépendant a permis de valider la robustesse des modèles. Avec LASSO, KNN a atteint une exactitude de 97,5 % et Gradient Boosting une précision parfaite de 100 %, surpassant ainsi les résultats d'études antérieures. Ce travail a également mis en lumière l'intérêt de coupler une réduction dimensionnelle pertinente à des algorithmes puissants pour améliorer la qualité des prédictions.

Le chapitre suivant propose une approche avancée visant à améliorer davantage la précision et la robustesse des modèles prédictifs. Cette nouvelle contribution explore des stratégies méthodologiques plus affinées et des techniques d'analyse optimisées, dans le mais de développer un système prédictif plus fiable, mieux adapté aux données cliniques réelles, et capable de répondre aux exigences du diagnostic médical précoce.

Chapitre 5

Contribution et évaluation

(Partie 2)

5.1 Introduction

Dans le prolongement des travaux précédents, ce chapitre présente deux nouvelles contributions visant à renforcer la précision et la robustesse des modèles de prédiction appliqués à la maladie rénale chronique (MRC), en agissant cette fois-ci sur l'enrichissement du jeu de données. La première contribution consiste à ajouter au dataset existant une variable clinique essentielle : le taux de filtration glomérulaire (DFG ou eGFR). Calculé à partir de variables biologiques disponibles, cet indicateur est reconnu comme un critère déterminant pour l'évaluation de la fonction rénale et du stade de la MRC. L'intégration du DFG dans le jeu de données a permis d'améliorer la représentation clinique des patients et d'affiner les prédictions. Dans une seconde contribution, nous avons poursuivi l'amélioration du dataset en procédant à l'augmentation du nombre d'échantillons, avec pour objectif d'améliorer la représentativité des données, de limiter les biais liés aux petits ensembles, et d'augmenter la capacité de généralisation des modèles. Ces deux apports successifs permettent de réévaluer l'efficacité des méthodes de classification étudiées, en observant l'impact direct de l'enrichissement du jeu de données sur les performances des modèles prédictifs.

5.2 Conception de l'approche proposé (Contribution 3) : An innovative approach for predictive modeling and staging of chronic kidney disease

Dans le cadre de cette contribution, nous avons élaboré une approche prédictive avancée et méthodiquement structurée, visant à renforcer la prédiction précoce de la maladie rénale chronique (MRC). Notre démarche ne se limite pas à l'application d'algorithmes d'apprentissage automatique, mais s'inscrit dans une vision globale, où la pertinence médicale, la qualité des données et la robustesse des étapes de traitement sont considérées comme des leviers complémentaires au service d'une prédiction fiable et exploitable en pratique clinique.

Conscients des limites méthodologiques référencées dans de nombreuses études antérieures, nous avons structuré notre approche autour d'un pipeline intégrant de manière approfondie des phases de nettoyage, de transformation, d'enrichissement et d'optimisation des données. Une attention particulière a été portée à l'intégration de critères médicaux à forte valeur ajoutée, ainsi qu'à la réduction de la complexité du modèle, afin d'éviter les risques de surapprentissage tout en améliorant la lisibilité des résultats.

L'ensemble de données final a été scindé en deux sous-ensembles distincts, avec 80 % des données dédiées à l'apprentissage des modèles et 20 % réservées à l'évaluation, permettant ainsi d'apprécier de manière objective leurs performances sur des données non vues. Ce processus se décline en six étapes principales, allant de la préparation des données à l'évaluation finale des performances des modèles, comme illustré dans la Figure 5.1. Chaque étape a été conçue de manière à garantir la fiabilité, la cohérence statistique et la validité clinique des résultats obtenus. En combinant rigueur algorithmique et sensibilité au contexte médical, cette

approche se veut reproductible, généralisable et directement transposable aux problématiques réelles de diagnostic précoce de la MRC.

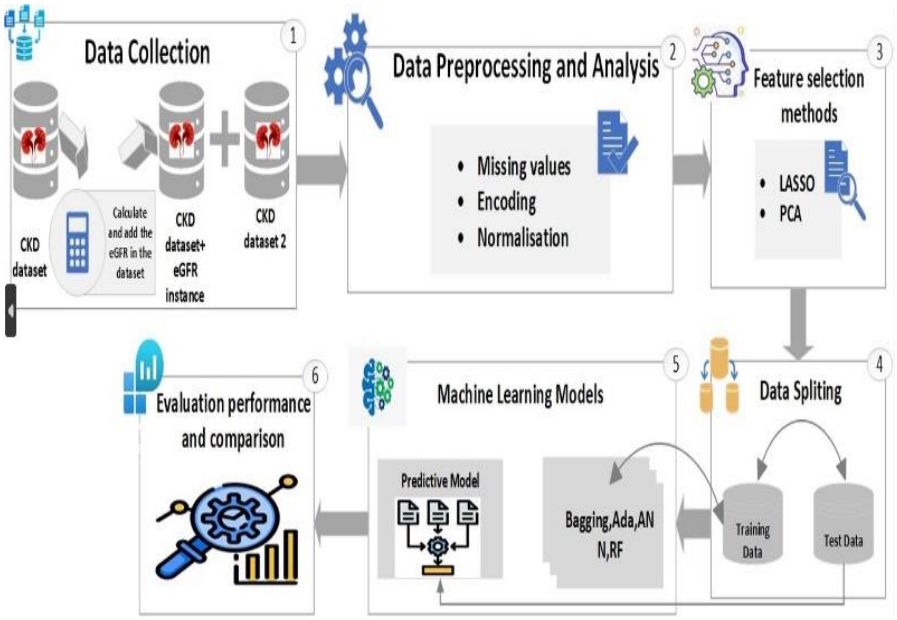


Figure 5.1: Architecture de notre approche

5.3 Ensembles de données :

Dans cette nouvelle contribution, notre démarche repose sur l'idée que la qualité des données joue un rôle tout aussi déterminant que les algorithmes dans la réussite d'un système de prédiction fiable, notamment dans le domaine médical où l'exactitude des diagnostics est cruciale. Bien que le jeu de données utilisé soit issu d'un référentiel reconnu (UCI Machine Learning Repository) et largement exploité dans la littérature, nous avons identifié l'absence d'une variable clinique majeure : le taux de filtration glomérulaire important (DFG ou eGFR). Cet indicateur est essentiel pour évaluer la fonction rénale, diagnostiquer la MRC, en déterminer le stade, et orienter les choix thérapeutiques. En l'ajoutant à notre jeu de données, calculé via la formule MDRD, nous avons cherché à combler une lacune fréquente dans les études existantes et à enrichir le jeu de données avec une variable à forte valeur clinique ajoutée. Ce choix s'aligne avec notre objectif de produire des modèles prédictifs non seulement performants sur le plan statistique, mais aussi pertinents d'un point de vue médical.

Afin de renforcer davantage la robustesse de notre démarche, nous avons également intégré un deuxième jeu de données, indépendant. Ce jeu de données, composé de 200 instances et de 28 variables, intègre-lui aussi le DFG et inclut des informations sur le stade de la maladie, ce qui en fait une base particulièrement adaptée à l'évaluation externe de nos modèles. L'utilisation de ce deuxième jeu de données exclusivement pour la phase de test permet de valider la généralisabilité de notre approche sur des données distinctes, tout en conservant une certaine cohérence avec le jeu de données d'entraînement. Cette stratégie d'évaluation croisée sur deux bases différentes renforce la crédibilité de nos résultats et leur transférabilité vers d'autres contextes cliniques. Enfin, l'enrichissement du jeu de données initial par l'ajout de l'estimation du DFG, associé à une validation sur une base externe, constitue une contribution

méthodologique majeure, renforçant la pertinence clinique de nos travaux pour la prédiction précoce de la maladie rénale chronique.

age	bp	sg	al	su	bgr	bu	sc	sod	pot	...	pcc	ba	htn	dm	cad	appet	pe	ane	class	eGFR
48.0	80.000000	1.020	1.0	0.0	121.0	36.0	1.2	137.613782	4.634084	...	0	0	1	1	0	0	0	0	0	68.682456
51.0	80.000000	1.010	2.0	0.0	106.0	26.0	1.4	137.613782	4.634084	...	0	0	0	0	0	0	0	0	0	56.786414
60.0	90.000000	1.015	3.0	0.0	74.0	25.0	1.1	142.000000	3.200000	...	0	0	1	1	0	0	1	0	0	72.574004
68.0	70.000000	1.010	0.0	0.0	100.0	54.0	24.0	104.000000	4.000000	...	0	0	0	0	0	0	0	0	0	2.017204
24.0	76.545455	1.015	2.0	4.0	410.0	31.0	1.1	137.613782	4.634084	...	0	0	0	1	0	0	1	0	0	87.410359
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
55.0	80.000000	1.020	0.0	0.0	140.0	49.0	0.5	150.000000	4.900000	...	0	0	0	0	0	0	0	0	1	183.488113
42.0	70.000000	1.025	0.0	0.0	75.0	31.0	1.2	141.000000	3.500000	...	0	0	0	0	0	0	0	0	1	70.569685
12.0	80.000000	1.020	0.0	0.0	100.0	26.0	0.6	137.000000	4.400000	...	0	0	0	0	0	0	0	0	1	202.512861
17.0	60.000000	1.025	0.0	0.0	114.0	50.0	1.0	135.000000	4.900000	...	0	0	0	0	0	0	0	0	1	104.648212
58.0	80.000000	1.025	0.0	0.0	131.0	18.0	1.1	141.000000	3.500000	...	0	0	0	0	0	0	0	0	1	73.075182

Figure 5.2: Aperçu du dataset d'entraînement après ajout du DFG

Tableau 5.1: Comparaison des fonctionnalités entre les ensembles de données d'entrainement et de test

Caractéristiques	CKD dataset d'entrainement	Valeurs manquantes	CKD dataset de test	Valeurs manquantes
Âge	Numérique	2.2	Catégorique	0,0
Pression artérielle (pb)	Numérique	3.0	—	—
Pression artérielle diastolique (bp)	—	—	Catégorique	0,0
Limite de pression artérielle (limit bp)	—	—	Catégorique	0,0
Densité (sg)	Catégorique	11,8	Catégorique	0,0
Albumine (al)	Catégorique	11,5	Catégorique	0,0
Sucre (su)	Catégorique	12.2	Catégorique	0,0
Globules rouges (rbc)	Catégorique	38,0	Catégorique	0,0
Cellule de pus (pc)	Catégorique	16.2	Catégorique	0,0
Amas de cellules pusiques (pcc)	Catégorique	1.0	Catégorique	0,0
Bactéries (ba)	Catégorique	1.0	Catégorique	0,0
Glycémie aléatoire (bgr)	Numérique	11.0	Catégorique	0,0

Urée sanguine (bu)	Numérique	4.8	Catégorique	0,0
Créatinine sérique (sc)	Numérique	4.2	Catégorique	0,0
Sodium (sod)	Numérique	21,8	Catégorique	0,0
Potassium (pot)	Numérique	22.0	Catégorique	0,0
Hémoglobine (hemo)	Numérique	13.0	Catégorique	0,0
Volume globulaire total (pcv)	Numérique	17,5	Catégorique	0,0
Numération des globules blancs (wbcc)	Numérique	26.2	Catégorique	0,0
Numération des globules rouges (rbc)	Numérique	32,5	Catégorique	0,0
Hypertension artérielle (htn)	Catégorique	0,5	Catégorique	0,0
Diabète sucré (dm)	Catégorique	0,5	Catégorique	0,0
Maladie coronarienne (cad)	Catégorique	0,5	Catégorique	0,0
Appétit (appet)	Catégorique	0,2	Catégorique	0,0
Œdème des pieds (pe)	Catégorique	0,2	Catégorique	0,0
Anémie (ane)	Catégorique	0,2	Catégorique	0,0
Stade (stage)	__	__	Catégorique	0,0
Débit de filtration glomérulaire (DFG)	Numérique	0,0	Catégorique	0,0
Affecté (affected)	__	__	Catégorique	0,0
Classe (étiquette)	Catégorique	0,0	Catégorique	0,0

5.4 Prétraitement des données

Le prétraitement représente une étape déterminante dans tout processus d'apprentissage supervisé, en particulier dans le domaine médical où la qualité des données conditionne directement la robustesse et la pertinence des prédictions. Dans notre étude, nous avons travaillé à partir de deux ensembles de données distincts, nécessitant chacun une approche de traitement spécifique.

L'ensemble de données d'entrainement, issu du référentiel UCI, présente un taux global de valeurs manquantes atteignant 9,7 %**,** comme le montre le Tableau précédent. Ces valeurs manquantes sont réparties sur plusieurs attributs cliniques et biologiques, rendant indispensable une phase de traitement adaptée. En revanche, le jeu de données de test**,** utilisé exclusivement pour la phase de test externe, ne présente aucune valeur manquante, et ne nécessite donc aucune étape d'imputation**,** ce qui renforce sa fiabilité pour l'évaluation finale.

Pour le dataset d'entrainement, nous avons tout d'abord catégorisé les attributs en deux types : *catégoriels* (ex. : rbc, htn, appet) et *numériques* (âge, sc, hemo, etc.). Pour les variables catégorielles, les valeurs manquantes ont été imputées par la valeur modale, garantissant la conservation des classes majoritaires. Les variables numériques ont été complétées par imputation moyenne, méthode utilisée qui minimise les perturbations tout en conservant la structure statistique des données.

Dans un second temps, afin de limiter l'effet des valeurs aberrantes pouvant fausser l'apprentissage, nous avons appliqué la méthode de l'intervalle interquartile (IQR). Celle-ci consiste à identifier toute valeur située au-delà de l'intervalle :

$$QR=Q\,3-Q\,1\ ;\ Borne\ supérieure=Q\,3+1{,}5\times IQR\ ;\ Borne\ inferieure=Q\,1-1{,}5\times IQR$$

Les valeurs aberrantes identifiées ont été remplacées par la moyenne de la variable correspondante, stabilisant ainsi les distributions sans altérer la tendance centrale. Cette étape permet d'éviter une sensibilité excessive aux fluctuations extrêmes tout en conservant l'information utile.

Par la suite, les variables catégorielles ont été encodées en variables numériques à l'aide de la classe LabelEncoder, une étape indispensable pour assurer la compatibilité avec les algorithmes de machine Learning. Enfin, les variables numériques ont été normalisées par mise à l'échelle Min-Max, afin de ramener toutes les valeurs dans une plage [0, 1]. Cette homogénéisation des échelles permet d'éviter que certaines caractéristiques dominent les autres lors de l'entraînement des modèles.

Notre approche de prétraitement, à la fois méthodique, progressive et rigoureusement adaptée au contexte médical, constitue un fondement essentiel à la fiabilité des prédictions obtenues. Elle garantit que les modèles développés reposent sur des données nettoyées, équilibrées, cohérentes et comparables, maximisant ainsi leur capacité à généraliser, à détecter précocement la MRC et à s'intégrer dans une logique décisionnelle clinique crédible et reproductible.

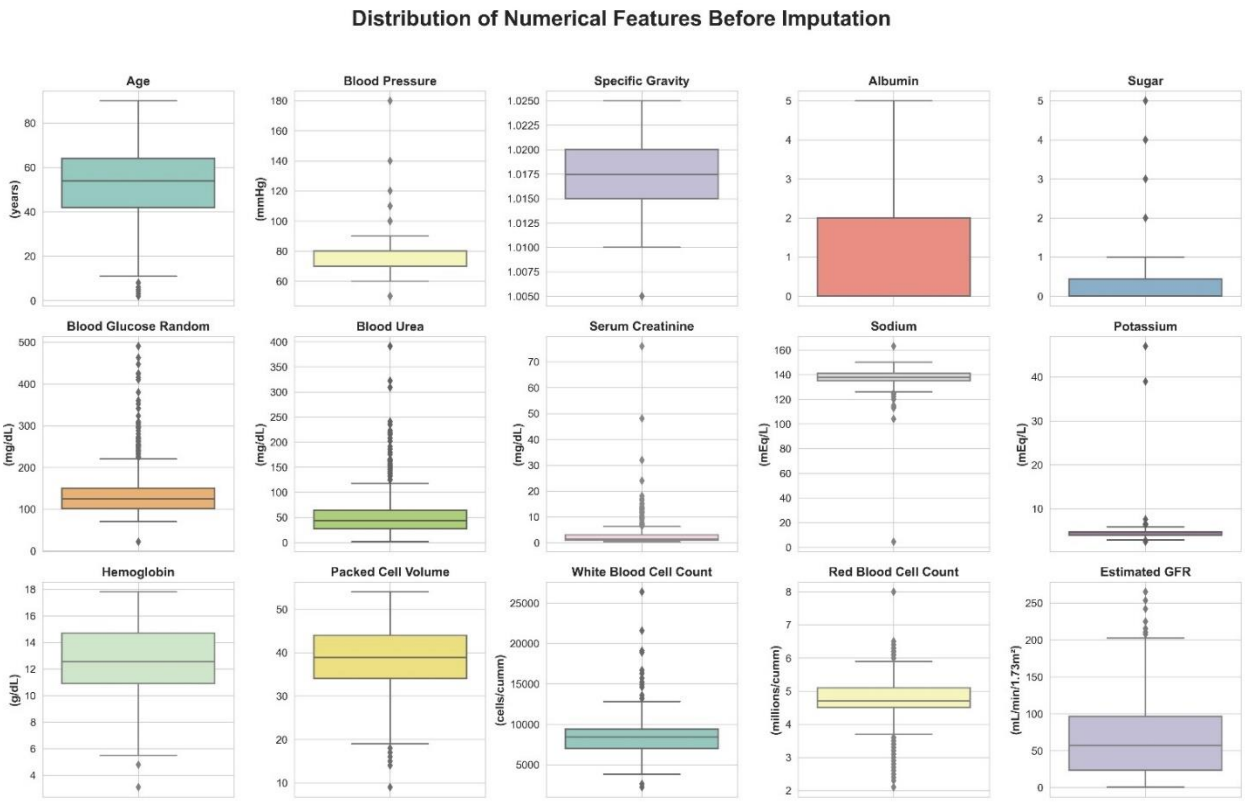


Figure 5.3: Distribution des caractéristiques numériques avant l'imputation

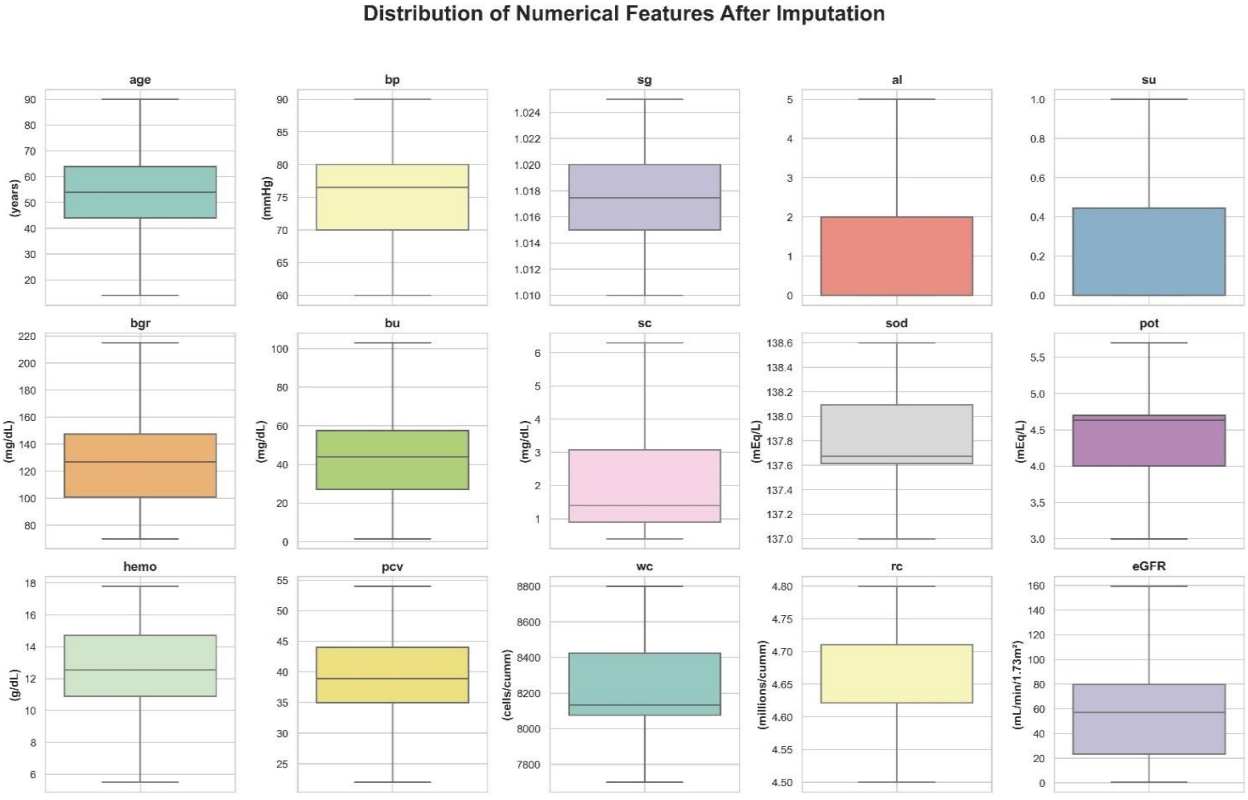


Figure 5.4: Distribution des caractéristiques numériques après l'imputation

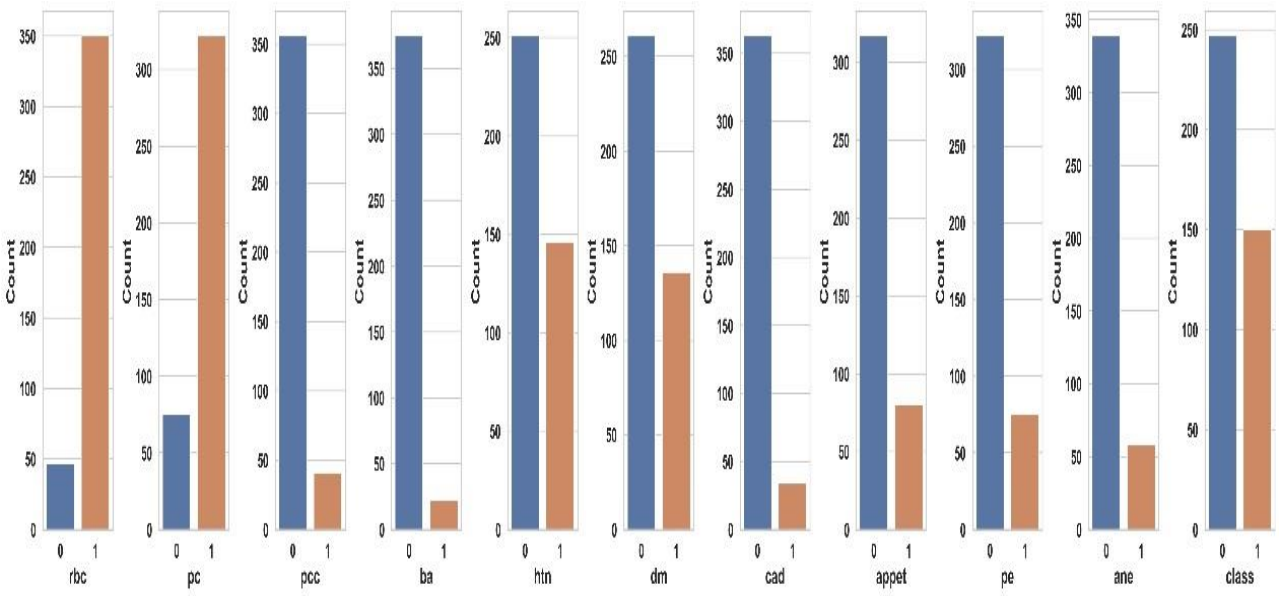


Figure 5.5: Distribution des variables catégorielles

5.5 Sélection des caractéristiques

La sélection des caractéristiques constitue une étape fondamentale dans le développement de tout modèle prédictif fiable. Dans le contexte médical, où les données sont souvent multidimensionnelles, bruitées et redondantes, cette étape permet non seulement d'améliorer la précision des modèles, mais également d'en réduire la complexité, d'accélérer leur exécution, et surtout de renforcer leur interprétabilité clinique. En ne conservant que les attributs les plus significatifs, le modèle devient plus robuste et moins sujet à l'overfitting, tout en facilitant la généralisation sur de nouvelles données.

Dans cette optique, nous avons intégré deux techniques complémentaires de sélection de caractéristiques, qui permettent de tirer parti de la richesse du jeu de données tout en focalisant l'analyse sur les variables les plus déterminantes pour la prédiction de la MRC. Cette approche contribue à supprimer les informations redondantes ou peu discriminantes, améliorant ainsi les performances globales du système.

Les résultats obtenus à l'issue de cette phase confirment la pertinence de notre stratégie. En effet, la sélection par la méthode LASSO a mis en évidence plusieurs variables fortement associées à la progression de la MRC, notamment des indicateurs biologiques tels que la densité urinaire (sg), l'albumine (al), la créatinine sérique (sc), l'hémoglobine (hemo), le volume cellulaire moyen (pcv), ainsi que des facteurs cliniques comme l'hypertension, le diabète ou encore la perte d'appétit. L'intégration du DFG (eGFR) a également été révélée comme une variable hautement contributive à la prédiction, renforçant la validité de notre choix méthodologique d'enrichir le jeu de données par ce paramètre.

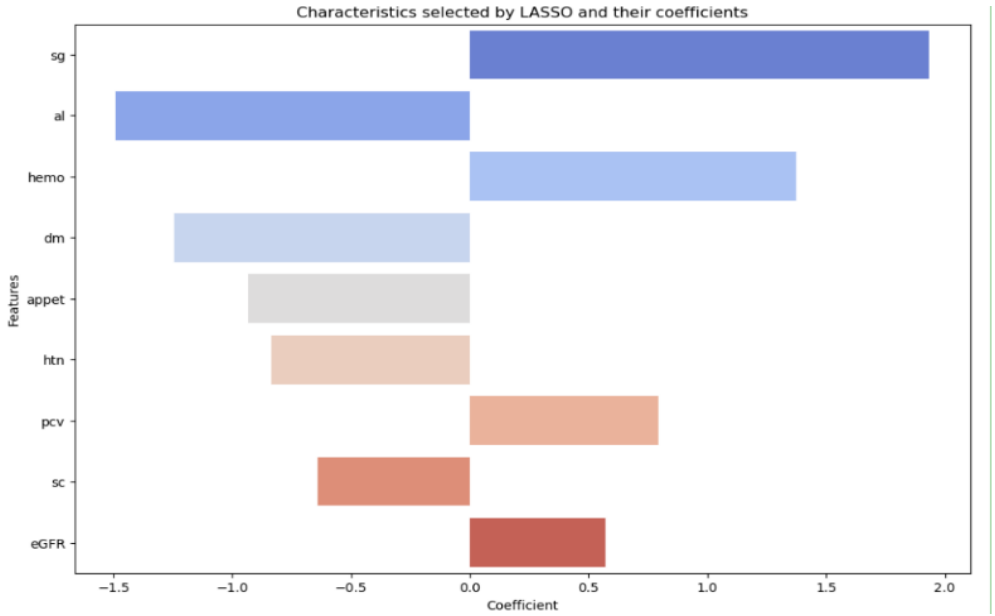


Figure 5.6: Les caractéristiques importantes sélectionnés par LASSO

Parallèlement, l'analyse en composantes principales a permis d'explorer en profondeur les structures latentes du jeu de données, révélant les regroupements entre variables et les corrélations multidimensionnelles qui ne sont pas toujours visibles par une simple analyse bivariée. Cette approche a notamment mis en lumière les axes les plus sensibles à la variabilité de certaines caractéristiques cliniques telles que la créatinine sérique, le sodium, le potassium ou encore le taux de globules rouges. Elle a également permis de mieux comprendre la manière dont certaines variables interagissent entre elles pour contribuer à la prédiction globale de la MRC.

Cette stratégie de sélection repose à la fois sur des principes statistiques rigoureux et sur des considérations médicales pertinentes. Elle permet de construire des modèles plus stables, plus explicables et plus efficaces, tout en garantissant une meilleure adéquation avec les réalités cliniques. Cette phase de sélection s'impose ainsi comme un élément central de notre démarche prédictive, assurant une transition optimale vers la modélisation**.**

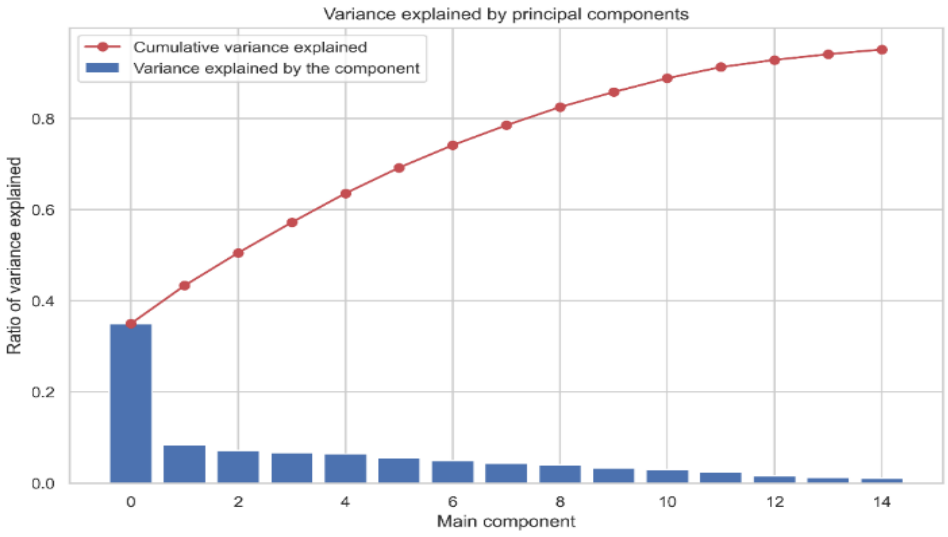


Figure 5.7: Variance expliquée par les composantes principales

5.6 Modélisation

Dans cette phase de notre contribution, nous avons appliqué quatre modèles d'apprentissage automatique supervisé : Bagging Classifier, Random Forest, AdaBoost et Réseaux de Neurones Artificiels (ANN). Ces algorithmes ont été sélectionnés pour leur capacité démontrée à traiter efficacement des problèmes de classification complexes, en particulier dans des contextes médicaux caractérisés par des données bruitées, multidimensionnelles et parfois déséquilibrées.

Le choix de ces modèles repose sur leur complémentarité. Les méthodes d'ensembles telles que Bagging, Random Forest et AdaBoost offrent une meilleure robustesse face au surapprentissage, en combinant plusieurs apprenants faibles ou non corrélés pour renforcer la stabilité des prédictions. Le réseau de neurones artificiels, quant à lui, apporte une approche non linéaire puissante capable de modéliser des relations complexes entre les variables, ce qui est particulièrement pertinent dans les maladies multifactorielles comme la MRC.

L'ensembles des modèles a été évalué sur les ensembles de données enrichis et nettoyés, à la fois avec et sans sélection de caractéristiques, afin de comparer l'impact de cette étape sur les performances de classification. Une validation croisée a été mise en œuvre pour garantir la fiabilité des résultats, en utilisant une division classique en 80 % pour l'entraînement et 20 % pour les tests, tout en assurant l'équilibre entre les classes.

Cette de modélisation vise à identifier les algorithmes les plus performants et les plus adaptés à la stratégie de prédiction précoce de la MRC, tout en évaluant leur capacité à généraliser sur des données externes grâce à l'intégration du second jeu de données indépendant.

5.7 Résultats expérimentaux et discussion

5.7.1 Identification et répartition des stades de la MRC

Dans cette section, nous avons entrepris une analyse fine des stades de la maladie rénale chronique (MRC) à partir des valeurs du DFG calculées pour chaque patient. Conformément aux recommandations cliniques internationales, nous avons classé les patients en six stades : CKD1 (DFG $\geq$ 90), CKD2 (DFG $\geq$ 60), CKD3a (DFG $\geq$ 45), CKD3b (DFG $\geq$ 30), CKD4 (DFG $\geq$ 15) et CKD5 (DFG < 15). Cette classification nous a permis de mieux cerner la progression de la maladie au sein de la population étudiée, et notamment d'identifier les cas précoces.

L'analyse révèle une prédominance des stades précoces CKD1 et CKD2, respectivement avec 69 et 110 patients, ce qui démontre l'efficacité de notre approche prédictive dans la prédiction en amont de la maladie, avant l'apparition de signes cliniques marqués. Fait notable, plusieurs patients initialement non déterminés comme atteints de MRC ont été reclassés dans ces stades précoces après calcul du DFG, mettant en lumière l'importance de cet indicateur dans le diagnostic précoce. À l'autre extrémité du spectre, nous observons également une proportion significative de patients dans les stades avancés : 45 en CKD3b, 68 en CKD4, et 66 en CKD5, soulignant la diversité des profils dans notre cohorte et la nécessité d'une prise en charge adaptée selon le stade de la maladie.

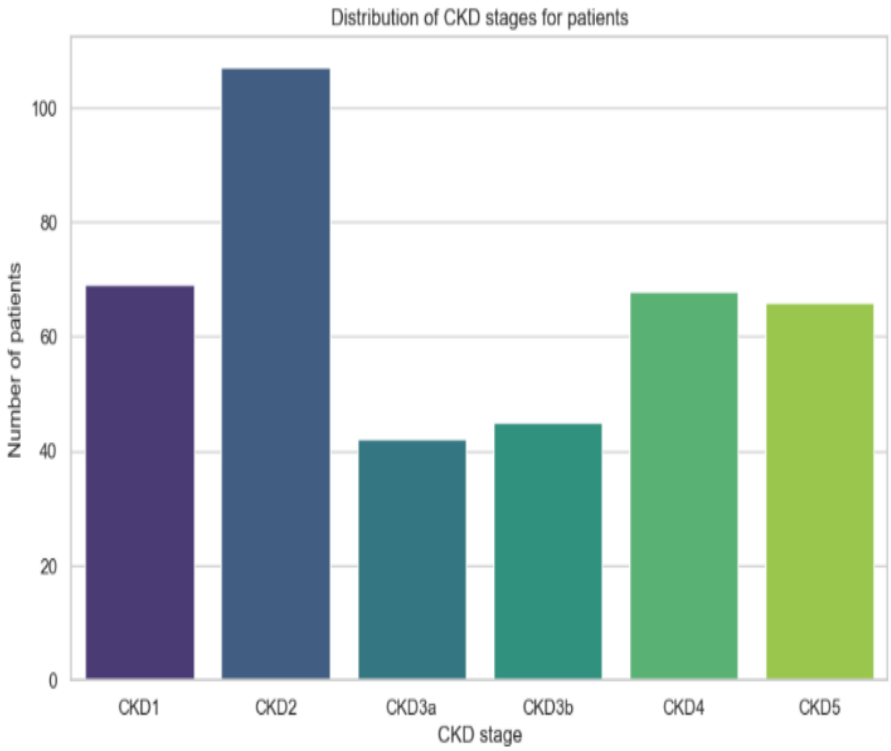


Figure 5.8: Distribution des stades de la MRC

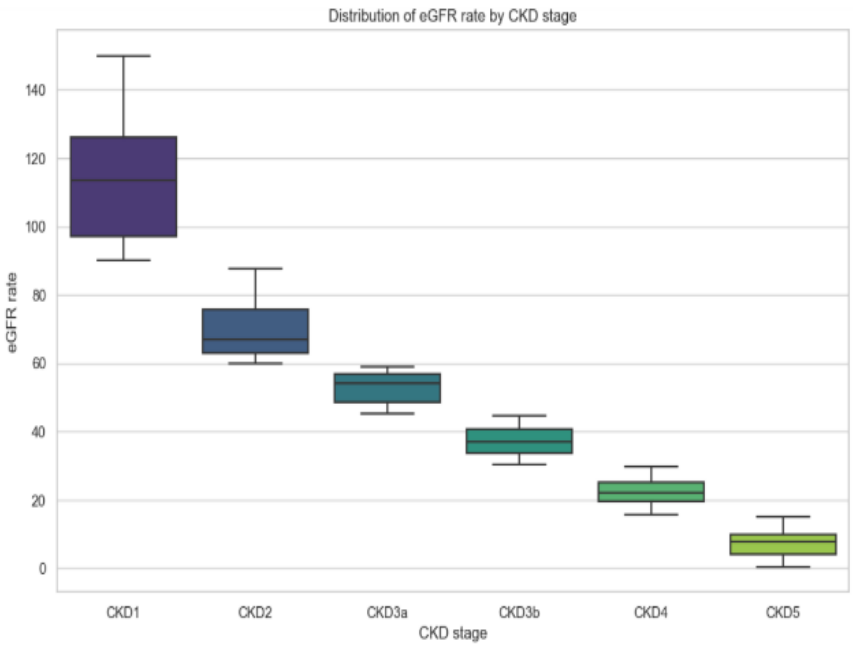


Figure 5.9: Distribution du taux de DFGe par stade de l'MRC

5.7.2 Comparaison des résultats obtenus pour chaque modèle sur le premier jeu de données

Afin de garantir une évaluation rigoureuse de nos modèles, nous avons mis en œuvre une validation croisée à 5 plis, méthode largement reconnue pour sa fiabilité dans les analyses prédictives. Les performances des modèles ont été évaluées à l'aide de quatre métriques essentielles : l'exactitude (accuracy), la précision (precision), le rappel (recall) et le score F1. L'évaluation a été réalisée à la fois sur les ensembles de données transformées par l'Analyse en

Composantes Principales (ACP) et sur ceux réduits via la méthode LASSO. Ces techniques de sélection ont permis de réduire la complexité du modèle, de limiter le surapprentissage et de renforcer l'interprétabilité des résultats, tout en conservant les variables les plus informatives.

5.7.2.1 Comparaison des modèles selon l'exactitude avec PCA

L'exactitude, indicateur fondamental en apprentissage automatique, reflète la capacité globale d'un modèle à effectuer des prédictions correctes. Pour la prédiction de la MRC, cet indicateur est crucial. Les résultats obtenus montrent que Random Forest se distingue avec un score de 98,3 %, traduisant sa forte capacité à gérer la complexité des données. AdaBoost atteint 99,1 %, témoignant de sa performance remarquable en classification binaire. Le modèle Bagging suit avec 97,5 %, confirmant sa robustesse. Quant au modèle ANN, il affiche une exactitude de 96,67 %, soulignant une performance solide malgré une complexité algorithmique plus élevée. Le modèle hybride ANN_AdaBoost enregistre un score exceptionnel de 99,5 %, validant l'intérêt d'associer les forces de deux approches complémentaires.

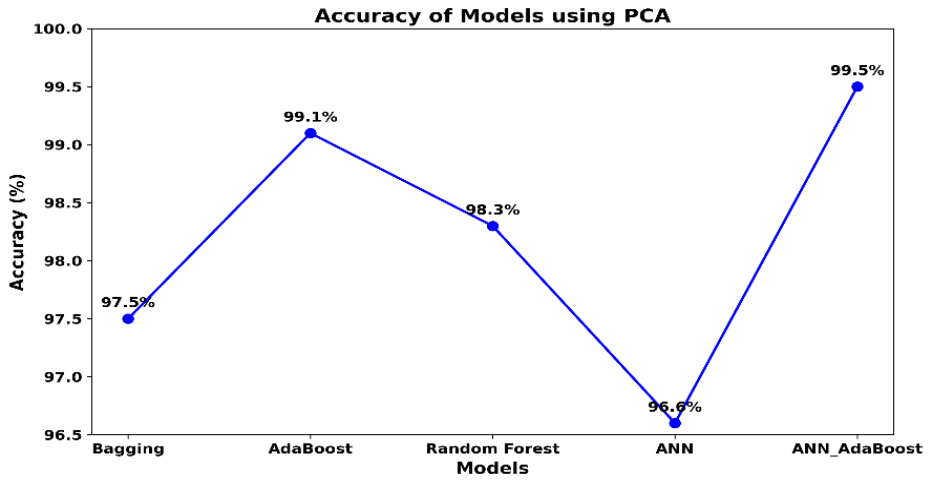


Figure 5.10: Résultats de l'Accuracy obtenus à l'aide de PCA

5.7.2.2 Comparaison des modèles selon la précision avec PCA

La précision permet d'évaluer la proportion de prédictions positives correctes. En contexte médical, elle est primordiale pour limiter les faux positifs et éviter des diagnostics erronés. Random Forest démontre ici une excellente fiabilité avec une précision de 98,3 %, suivi d'AdaBoost à 97,9 %, et de Bagging à 97,7 %. Le modèle ANN obtient 96 %, ce qui reste élevé, tandis que le modèle hybride atteint 97,8 %. Ces performances témoignent d'une grande rigueur dans l'identification des cas positifs réels.

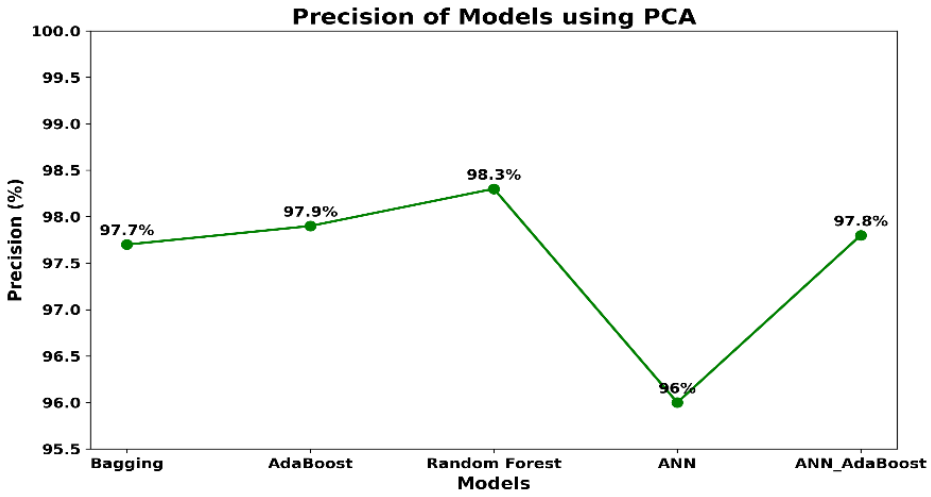


Figure 5.11: Résultats de la précision obtenus à l'aide de PCA

5.7.2.3 Comparaison des modèles selon le rappel avec PCA

Le rappel constitue une métrique déterminante dans les systèmes de diagnostic médical, car il indique la capacité du modèle à détecter l'ensembles des cas positifs, c'est-à-dire tous les patients réellement atteints de la maladie. Dans le cadre de la prédiction de la MRC, cette mesure revêt une importance capitale : un score de rappel faible pourrait se traduire par des patients non diagnostiqués à temps, avec des conséquences cliniques potentiellement graves.

Les résultats expérimentaux montrent qu'AdaBoost atteint un score parfait de 100 %, détectant ainsi tous les cas de MRC sans exception. Ce résultat met en lumière la puissance de cet algorithme dans un contexte aussi critique. Random Forest, avec un rappel de 98,3 %, confirme également une excellente capacité à éviter les faux négatifs, soutenu par son architecture en arbres de décision multiples, qui assure une exploration exhaustive des données.

Notre modèle hybride ANN_AdaBoost affiche un rappel de 97,8 %, illustrant un très bon compromis entre sensibilité et robustesse, grâce à la complémentarité entre les caractéristiques d'un réseau neuronal et celles du boosting. Le modèle ANN seul atteint un score de 96 %, indiquant une performance solide, mais suggérant qu'une optimisation supplémentaire pourrait en améliorer la sensibilité. Enfin, Bagging, avec un rappel de 95,7 %, montre également de très bonnes performances, renforçant la pertinence des méthodes d'ensembles dans les tâches de classification médicale.

Dans l'ensembles, ces résultats confirment l'efficacité des modèles testés dans la prédiction de la MRC, et soulignent l'importance du rappel comme indicateur clé dans des contextes cliniques où la priorité est de ne manquer aucun cas. Ils appuient également le potentiel des approches hybrides et des techniques d'ensembles pour renforcer la fiabilité des systèmes d'aide à la décision médicale.

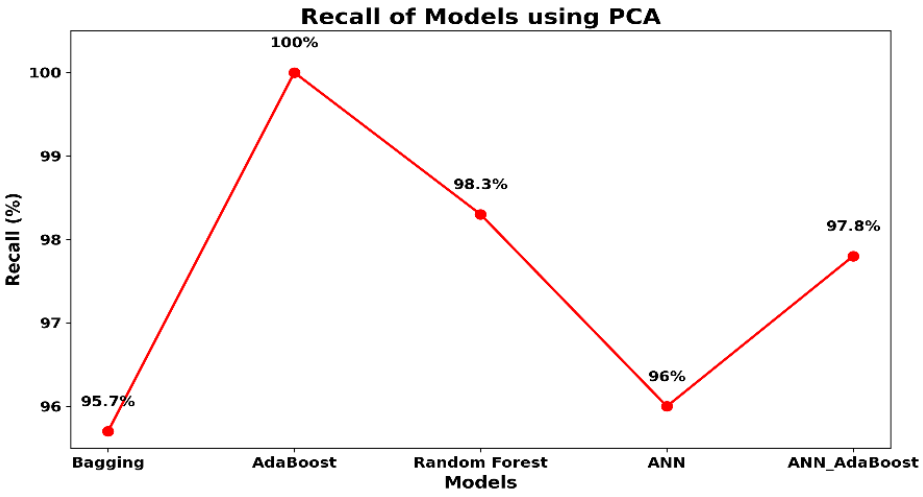


Figure 5.12: Résultats de la précision obtenus à l'aide de PCA

5.7.2.4 Comparaison des modèles selon le score F1 avec PCA

Le score F1, qui représente la moyenne harmonique entre la précision et le rappel, est particulièrement pertinent dans les contextes critiques où les erreurs de classification peuvent avoir de lourdes conséquences, comme c'est le cas en santé. Il offre une vue d'ensembles équilibrée de la performance du modèle, surtout en présence de classes déséquilibrées.

Parmi les modèles évalués, Random Forest se distingue par un score F1 remarquable de 98,3 %, témoignant de sa capacité à maintenir un bon équilibre entre la prédiction correcte des cas et la limitation des erreurs. AdaBoost, avec un score de 98,9 %, confirme sa robustesse et sa fiabilité, grâce à une synergie optimale entre précision et rappel. Le modèle hybride ANN_AdaBoost atteint un score de 97,8 %, illustrant l'efficacité d'une approche combinée qui exploite les avantages des réseaux de neurones et des techniques de boosting. Le modèle ANN, quant à lui, obtient un score de 96 %, ce qui reste compétitif, mais suggère une marge d'amélioration possible pour un meilleur équilibre entre prédiction et exactitude. Le modèle Bagging complète l'évaluation avec un score de 96,7 %, confirmant sa stabilité et son potentiel dans le cadre d'un système de diagnostic fiable. Ces résultats soulignent l'intérêt du score F1 dans l'évaluation de modèles de classification médicale, et confirment la pertinence de notre approche, notamment via des méthodes hybrides et des techniques d'ensembles, pour garantir un diagnostic précoce, fiable et équilibré de la MRC.

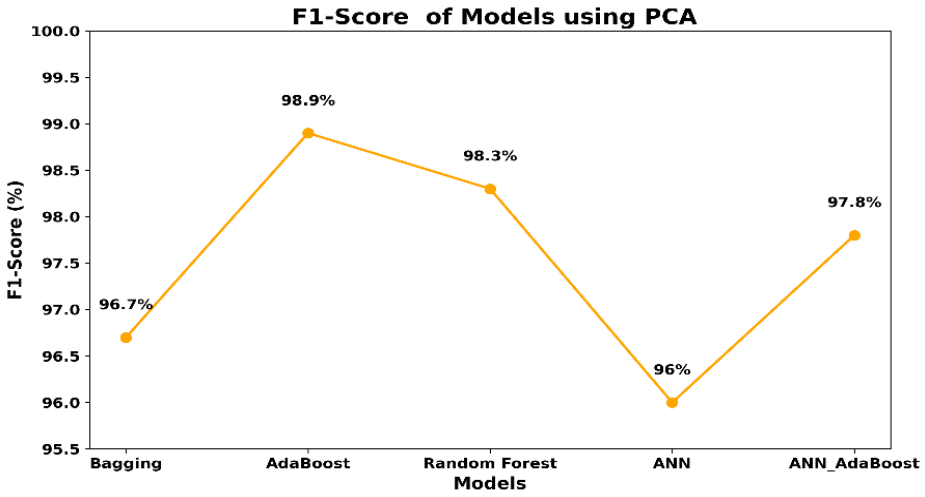


Figure 5.13: Résultats de F1-Score avec PCA

5.7.2.5 Analyse de l'AUC du modèle hybride pour la prédiction de la MRC (PCA)

L'AUC (Area Under the Curve) constitue une métrique cruciale pour évaluer la performance d'un classificateur, en particulier dans les tâches de classification binaire comme la prédiction de la MRC. Un score AUC proche de 1 reflète une capacité exceptionnelle du modèle à distinguer correctement entre les classes positives et négatives. Dans notre étude, le modèle hybride ANN_AdaBoost a atteint un score AUC remarquable de 0,999, témoignant d'une précision quasi parfaite dans la différenciation entre les patients atteints et non atteints de la maladie.

Cette performance est particulièrement significative dans un contexte médical, où une mauvaise classification peut entraîner des diagnostics erronés avec des conséquences graves sur la prise en charge thérapeutique. Un score AUC de 0,5 correspondrait à un modèle sans pouvoir discriminant, équivalent à un choix aléatoire, tandis qu'un score de 1 refléterait une classification parfaite. Le score de 0,999 obtenu par notre modèle le positionne donc à un niveau d'excellence très rare. Ce résultat atteste non seulement de la solidité de l'architecture hybride utilisée, combinant la capacité d'abstraction des réseaux neuronaux à la puissance adaptative du boosting, mais aussi de la qualité du prétraitement et de la sélection des caractéristiques intégrés en amont. Il démontre également que notre modèle est en mesure de généraliser efficacement les apprentissages et de prendre des décisions fiables, même dans des situations cliniques délicates.

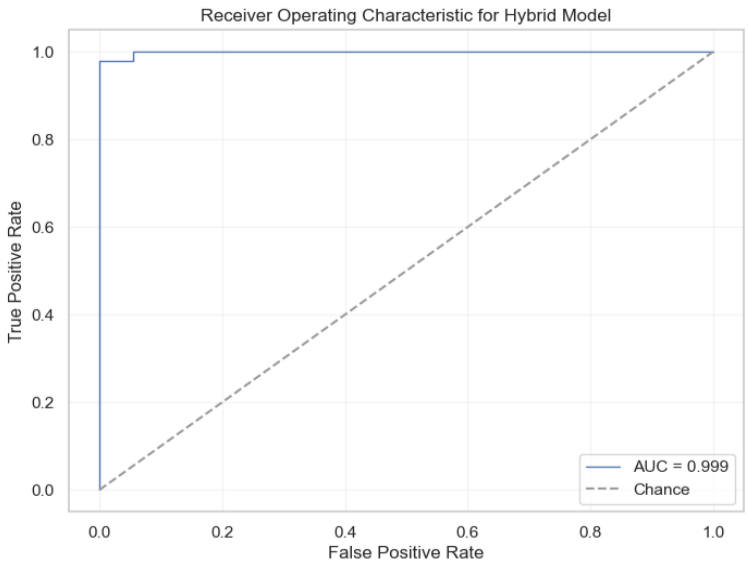


Figure 5.14: Analyse ROC du modèle hybride pour la prédiction de la MRC

5.7.2.6 Comparaison des modèles selon l'exactitude avec LASSO

L'utilisation de la méthode LASSO pour la sélection des variables nous a permis de concentrer l'apprentissage sur un sous-ensembles optimal de caractéristiques, parmi lesquelles : la densité urinaire (sg), l'albumine (al), la créatinine sérique (sc), l'hémoglobine (hemo), le volume cellulaire moyen (pcv), l'hypertension (htn), le diabète (dm), l'appétit (appet), ainsi que le débit de filtration glomérulaire (DFG). Sur cette base, nous avons évalué plusieurs modèles d'apprentissage supervisé.

Le modèle Random Forest s'est particulièrement distingué en obtenant un taux d'exactitude de 98,3 %, démontrant ainsi son efficacité à exploiter ces caractéristiques clés pour fournir des prédictions fiables. Le modèle AdaBoost a affiché une performance honorable avec un score de 96,6 %, soulignant sa robustesse dans la classification des cas de MRC. Le classificateur Bagging a quant à lui obtenu une exactitude de 95,8 %, confirmant sa pertinence en tant qu'approche d'ensembles.

Le modèle ANN a atteint une exactitude de 96,67 %, ce qui traduit une performance compétitive, bien que légèrement inférieure à celle des méthodes d'ensembles. Enfin, notre modèle hybride ANN_AdaBoost a enregistré une exactitude de 96,6 %, mettant en avant l'intérêt de la combinaison de deux paradigmes d'apprentissage distincts. Toutefois, ce résultat montre aussi qu'un ajustement plus poussé des hyperparamètres pourrait permettre de tirer davantage parti de cette synergie.

Ces observations illustrent non seulement la pertinence des variables sélectionnées par LASSO, mais aussi la capacité de certains modèles à maximiser la précision à partir d'un sous-ensemble restreint et cliniquement significatif de caractéristiques, ouvrant ainsi la voie à des systèmes de prédiction plus rapides, moins coûteux et facilement interprétables dans le cadre de la pratique médicale quotidienne.

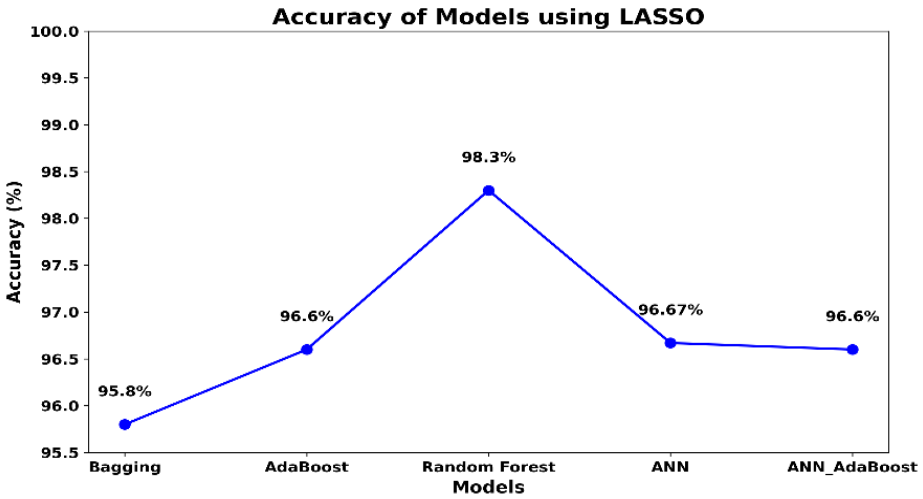


Figure 5.15: Résultats des scores de l'Accuracy à l'aide de LASSO

5.7.2.7 Comparaison des modèles selon la précision avec LASSO

Le modèle Random Forest obtient un score parfait de 100 %, illustrant sa capacité remarquable à éviter les faux positifs. Le modèle ANN suit avec une précision de 97 %, démontrant une capacité robuste à identifier correctement les cas positifs. Bagging atteint une précision de 95,8 %, et le modèle hybride ANN_AdaBoost atteint 92 %, révélant un bon potentiel malgré une légère marge de progression. Quant à AdaBoost, bien qu'il affiche également un score de 92 %, des ajustements supplémentaires pourraient améliorer sa spécificité. Ces résultats soulignent l'importance de la précision comme indicateur de confiance dans les prédictions positives, et confirment la pertinence de notre sélection de caractéristiques via LASSO pour optimiser les performances prédictives dans le cadre de la prédiction de la MRC.

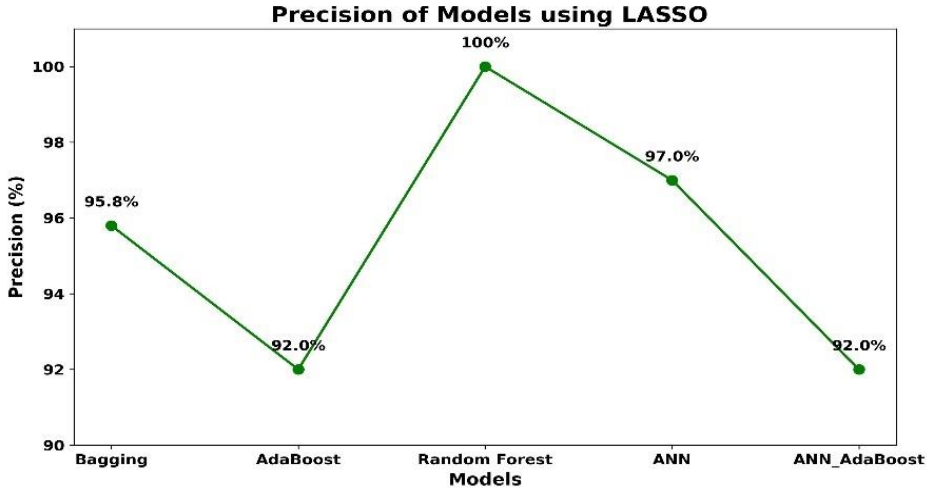


Figure 5.16: Résultats des scores de précision à l'aide de LASSO

5.7.2.8 Comparaison des modèles selon le rappel avec LASSO

En utilisant les neuf caractéristiques sélectionnées par LASSO, deux modèles : AdaBoost et le modèle hybride – se distinguent avec un score parfait de 100 %. Ce résultat atteste de leur aptitude à identifier l'intégralité des cas de MRC, sans omission. Le modèle ANN obtient un rappel de 97 %, témoignant d'une forte performance, bien que légèrement inférieure. Random

Forest suit avec un score de 95,65 %, mettant en évidence son efficacité globale, tandis que le modèle Bagging atteint 93,47 %. Bien que légèrement moins performant, ce dernier conserve un potentiel d'amélioration et reste pertinent dans le cadre d'un diagnostic étendu. Dans une optique clinique, où chaque cas non détecté peut retarder la prise en charge médicale, ces résultats soulignent l'importance capitale du rappel. Ils permettent également de guider les professionnels de santé dans la sélection de modèles les plus adaptés à un dépistage exhaustif de la MRC.

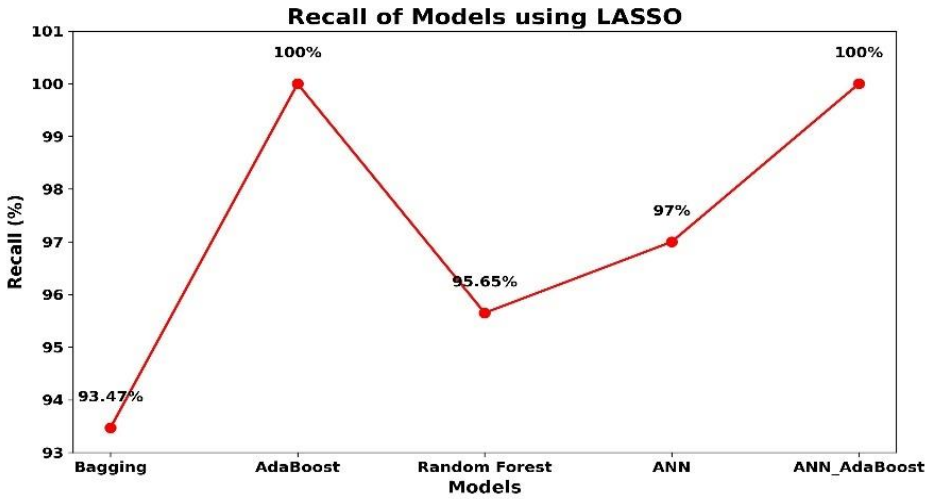


Figure 5.17: Résultats des scores de Recall à l'aide de LASSO

5.7.2.9 Comparaison des modèles selon le F1-Score avec LASSO

Dans notre analyse, le modèle Random Forest se démarque avec un score F1 de 97,7 %, témoignant de sa capacité à assurer un équilibre optimal entre prédiction correcte des cas de MRC et limitation des erreurs de classification. Le modèle ANN suit de près avec un score de 97 %, confirmant sa stabilité et sa cohérence globale. AdaBoost et ANN_AdaBoost affichent tous deux un score de 95,8 %, attestant d'une performance fiable et équilibrée dans la prédiction de la maladie. Le modèle Bagging, avec un score de 94,5 %, démontre également une performance prometteuse, bien que légèrement inférieure.

Ces résultats montrent que la méthode de sélection LASSO, en conservant des variables hautement significatives, permet aux modèles d'atteindre une bonne synergie entre précision et sensibilité, condition indispensable à une application clinique efficace dans le cadre du diagnostic précoce de la MRC.

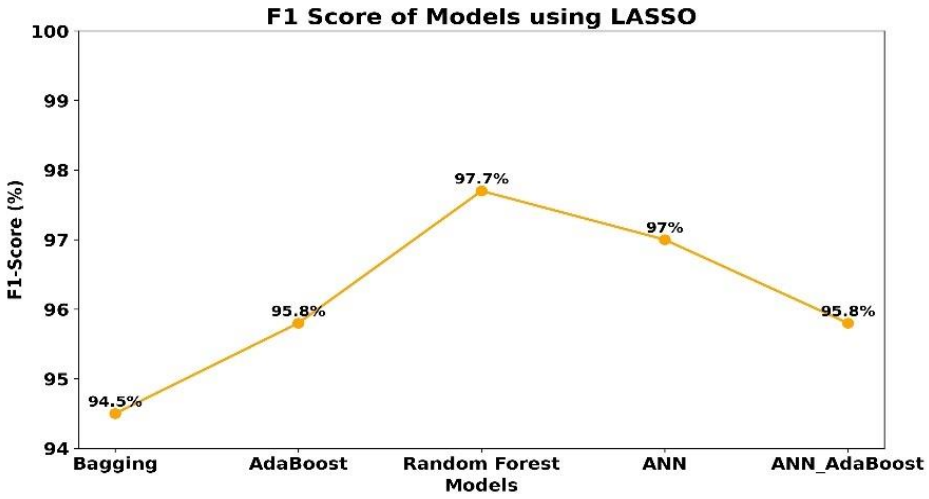


Figure 5.18: Résultats des scores de F1-Score à l'aide de LASSO

5.7.2.10 Analyse comparative des performances de l'Accuracy avec les sélections des caractéristiques LASSO et PCA

Dans le cadre de la prédiction de la MRC, le choix de la méthode de sélection des caractéristiques peut avoir un impact notable sur l'efficacité des algorithmes. Cette section compare les résultats obtenus avec LASSO et PCA, deux approches complémentaires de réduction de dimensionnalité.

Avec la méthode LASSO, le modèle Random Forest affiche une performance stable et élevée avec une exactitude de 98,3 %, résultat qui reste identique avec ACP. Cela témoigne de la robustesse de Random Forest, peu sensible au type de sélection appliquée. En revanche, AdaBoost connaît une nette amélioration, passant de 96,6 % avec LASSO à 99,1 % avec ACP, suggérant que les composantes principales extraites par ACP sont mieux adaptées à son architecture. Bagging progresse également (de 95,8 % à 97,5 %) et le modèle ANN conserve une exactitude stable de 96,67 %. Notre modèle hybride ANN_AdaBoost se démarque avec une augmentation significative de sa performance, atteignant 99,5 % avec ACP contre 96,6 % avec LASSO.

Ces résultats soulignent l'importance de choisir une technique de sélection des caractéristiques en cohérence avec le type de modèle utilisé. Alors que LASSO favorise une sélection directe des variables les plus discriminantes, ACP projette les données dans un nouvel espace optimisé, ce qui semble particulièrement avantageux pour les modèles complexes ou hybrides. Cette complémentarité méthodologique offre des perspectives intéressantes pour le développement de systèmes prédictifs robustes et adaptés aux contextes cliniques réels.

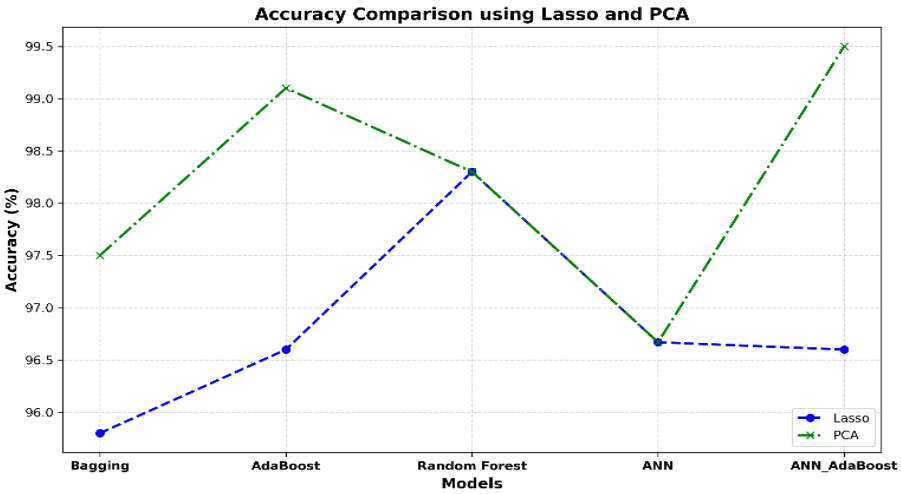


Figure 5.19: Résultats de l'Accuracy des modèles avec LASSO et PCA

5.7.3 Comparaison des performances de l'Accuracy des modèles sur l'ensembles de données de test

Pour évaluer la capacité réelle de généralisation de nos modèles prédictifs dans un contexte clinique, nous avons testé leur performance sur un second jeu de données indépendant. Cette approche méthodologique rigoureuse vise à vérifier que les modèles ne sont pas uniquement adaptés aux données d'entraînement, mais qu'ils conservent également une précision élevée lorsqu'ils sont confrontés à de nouvelles observations.

Les résultats, illustrés à la figure suivante, confirment la robustesse de notre stratégie. Le modèle AdaBoost s'illustre par une excellente performance, atteignant une exactitude de 99,3 %, soulignant sa capacité remarquable à identifier correctement les cas de MRC. Notre modèle hybride ANN_AdaBoost surpasse même ce résultat, avec une performance maximale de 99,6 %, démontrant la synergie efficace entre l'apprentissage neuronal profond et les techniques d'ensembles. Le modèle Random Forest maintient un haut niveau de précision avec un score de 98,3 %, confirmant une fois de plus sa fiabilité. Le modèle Bagging suit de près avec une exactitude de 98 %, prouvant son efficacité dans la gestion de l'hétérogénéité des données. Le modèle ANN, malgré une architecture plus complexe et sensible aux variations des données, obtient un résultat satisfaisant de 97,1 %.

Ce qu'il convient de souligner ici, c'est l'apport déterminant de l'utilisation d'un jeu de données externe. Cette démarche a permis d'évaluer la robustesse des modèles dans un environnement différent, en simulant des conditions réelles d'application. Grâce à cette validation externe, nous pouvons affirmer que nos modèles conservent une capacité de prédiction élevée en dehors du cadre strict de l'apprentissage initial, ce qui en fait des candidats crédibles pour une mise en œuvre dans des systèmes cliniques d'aide à la décision. Ainsi, notre approche globale renforce la fiabilité des prédictions et garantit une applicabilité concrète dans des contextes médicaux variés.

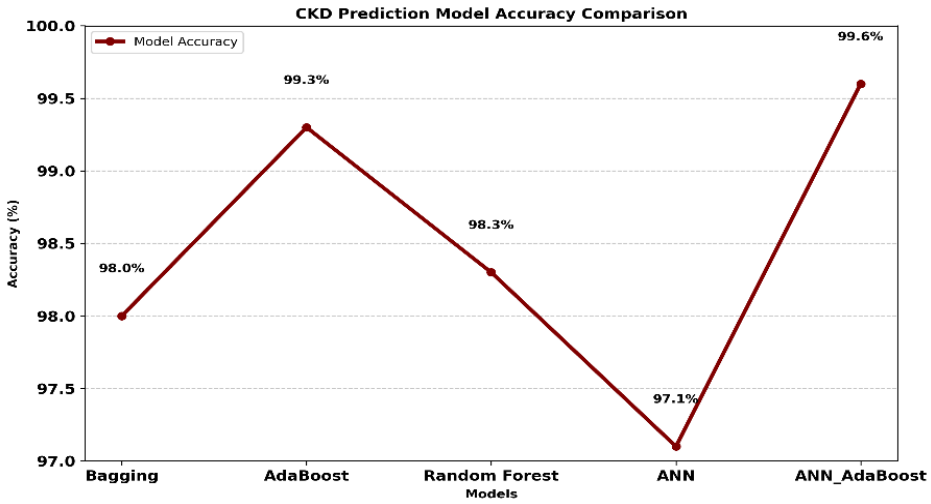


Figure 5.20: Résultats de l'Accuracy de nos modèles sur l'ensembles de données de test

5.7.4 Comparaison des performances de notre approche avec des travaux connexes

Dans cette étude, nous avons exploité deux ensembles de données distincts pour l'entraînement, la validation et les tests prédictifs de la MRC. Travailler sur des données médicales réelles implique souvent de faire face à des valeurs manquantes, causées par des erreurs de collecte, des difficultés d'extraction ou des absences d'information. La gestion de ces valeurs manquantes est donc une étape incontournable pour garantir la cohérence et la qualité des analyses prédictives. Dans notre démarche, un soin particulier a été apporté à la qualité de la phase de prétraitement. Plutôt que d'adopter des méthodes d'imputation basiques, nous avons opté pour des techniques plus rigoureuses, adaptées à la nature des variables (catégorielles ou numériques), afin de limiter au maximum les biais et préserver la structure informative du jeu de données. Cette stratégie de traitement nous a permis d'alimenter les algorithmes d'apprentissage avec des données plus fiables, condition indispensable pour développer des modèles robustes et généralisables.

Un autre élément distinctif de notre approche réside dans l'utilisation d'un second jeu de données entièrement indépendant pour l'étape de validation. Cette validation externe constitue un gage de crédibilité, car elle permet de tester les modèles sur des cas jamais vus auparavant, offrant ainsi une évaluation réaliste de leur comportement dans un cadre clinique réel.

Enfin, notre démarche s'est distinguée par l'enrichissement du jeu de données original grâce à l'intégration d'un biomarqueur essentiel : le DFG (eGFR). Cet indicateur, encore jamais intégré dans les études antérieures, joue un rôle fondamental dans l'évaluation de la fonction rénale. Son inclusion a permis d'améliorer significativement la pertinence médicale des prédictions, en renforçant la capacité des modèles à détecter les stades précoces et avancés de la MRC.

Les résultats obtenus dans notre étude dépassent globalement ceux rapportés dans la littérature, que ce soit avec les approches classiques ou les techniques récentes. Comme le montre le Tableau suivant, les performances atteintes par nos modèles attestent de la pertinence de notre méthodologie, aussi bien dans le choix des variables que dans l'architecture algorithmique mise en œuvre. Ces constats soulignent l'intérêt de notre contribution pour le développement d'outils intelligents d'aide à la décision dans la prédiction précoce de la MRC.

Tableau 5.2: Comparaison de l'Accuracy de notre étude avec les travaux existants

Travail	Modèles	Accuracy	Dataset 2 pour valider les résultats	Caractéristique eGFR
Xioa et al[169]	RF ANN	80% 80%	–	–
Hassan et al [170]	ANN RF	97.5% 98.75%	–	–
Islam et al [172]	AdaBoost RF	94% 95%	–	–
Chittora et al [188]	ANN	90.2%	–	–
Yashfi et al [173]	ANN RF	94,5% 97,12%	–	–
Pal [155]	RF ANN	93% 89%	–	–
Mustafizur Rahman et al [163]	RF Bagging AdaBoost	96,15% 98,07% 100%	Oui	_
Ghosh et al [162]	AdaBoost	97,91%,	–	–
Ebiaredoh-Mienye et al [174]	AdaBoost RF	93% 95,2%,	–	–
A. Islam et al [189]	AdaBoost RF ANN	98,3% 97,5% 60%	–	–
Notre proposition	**AdaBoost** **RF** **Bagging** **ANN** **Modèle_Hybride**	**97,5%** **99,1%** **98,3%** **96,6%** **99,5%**	**Oui**	**Oui**

5.8 Contribution 4 : Approche d'augmentation intelligente des données

Dans le cadre de cette quatrième contribution, nous avons poursuivi nos travaux sur la prédiction de la maladie rénale chronique (MRC) en utilisant le même jeu de données que dans les contributions précédentes, enrichi par le taux de filtration glomérulaire estimé (DFG ou eGFR). Ce biomarqueur, considéré comme une référence en néphrologie, ne se limite pas à l'évaluation de la fonction rénale globale : il constitue également la base clinique de classification des stades de la MRC (de 1 à 5). Son intégration dans notre dataset présente donc un double avantage. D'une part, elle renforce la pertinence clinique des variables, en fournissant un indicateur directement exploité par les praticiens. D'autre part, elle améliore la capacité de nos modèles à prédire la maladie dès ses premiers stades silencieux, où les symptômes sont absents mais où une prise en charge précoce est cruciale pour ralentir la progression. En intégrant le eGFR, nous avons ainsi transformé un simple jeu de données diagnostique en un

outil prédictif capable de soutenir le dépistage, le suivi et la stratification des patients selon la gravité de la maladie.

Cependant, la taille initiale du jeu de données (400 instances) demeurait limitée, ce qui constituait un frein majeur à l'entraînement de modèles d'apprentissage capables de généraliser correctement. À cette contrainte s'ajoutait un fort déséquilibre des classes, les cas de CKD étant proportionnellement surreprésentés.

Pour dépasser ces limites, nous avons mis en œuvre une stratégie d'augmentation de données reposant sur SMOTENC, une méthode spécifiquement conçue pour les ensembles de données mixtes (numériques et catégoriels). L'avantage de SMOTENC réside dans sa capacité à générer des exemples synthétiques plausibles tout en respectant la structure et la distribution des variables d'origine, évitant ainsi la création de valeurs incohérentes d'un point de vue médical. Afin de garantir un équilibre parfait entre les classes (CKD et non-CKD), nous avons complété cette étape par un RandomOverSampler, permettant de renforcer la représentativité des cas minoritaires.

Grâce à cette double stratégie, notre ensemble de données a été enrichi tout en préservant la cohérence clinique des variables. Les statistiques descriptives avant et après augmentation confirment que cette opération n'a introduit ni valeurs aberrantes ni distorsions significatives dans les distributions, validant ainsi la qualité et la plausibilité des nouvelles données générées.

Ensuite, pour assurer la robustesse et la généralisation de nos modèles, nous avons effectué un partitionnement stratifié du dataset augmenté, en conservant 80 % des données pour l'entraînement et 20 % pour le test. Ce choix garantit une répartition proportionnelle et équilibrée des classes dans les deux sous-ensembles, notamment les biais liés à une distribution aléatoire simple.

Pour l'évaluation, quatre modèles d'apprentissage supervisé ont été mobilisés : Random Forest, Decision Tree, AdaBoost et Gaussian Naive Bayes. Le recours à cette diversité d'algorithmes a permis de comparer leurs performances respectives, de valider la pertinence des variables sélectionnées et de renforcer la robustesse des résultats obtenus.

5.8.1 Augmentation des données

La taille restreinte du jeu de données initial (400 instances) constituait une contrainte majeure, insuffisante pour l'entraînement de modèles prédictifs robustes et généralisables. Par ailleurs, le déséquilibre marqué entre les classes CKD et non-CKD accentuait le risque de biais en faveur de la classe majoritaire, limitant la capacité des modèles à identifier correctement les cas minoritaires, notamment les patients aux stades précoces de la maladie. Ces profils, pourtant essentiels pour un diagnostic précoce, risquaient ainsi d'être sous-représentés lors de l'apprentissage.

Afin de remédier à ces limitations, nous avons adopté une stratégie d'augmentation raisonnée des données reposant sur la méthode SMOTENC (Synthetic Minority Oversampling Technique for Nominal and Continuous variables). Contrairement à SMOTE classique, initialement conçu pour des variables purement numériques, SMOTENC est spécifiquement adapté aux ensembles

de données mixtes intégrant à la fois des variables continues (âge, créatinine sérique, eGFR) et des variables catégorielles (hypertension, diabète). Ce choix méthodologique est particulièrement pertinent dans un contexte médical, car il permet de générer des échantillons synthétiques cliniquement plausibles tout en respectant la structure et la cohérence des variables cliniques.

Il est important de souligner que l'augmentation des données a été appliquée exclusivement après la phase de séparation du jeu de données, et uniquement sur le sous-ensemble d'entraînement. Le jeu de test a été conservé dans son état réel, sans aucune modification ni génération artificielle d'instances. Ce choix vise à prévenir toute fuite d'information (data leakage) et à garantir une évaluation objective et réaliste des performances des modèles, reflétant leur capacité de généralisation sur des données cliniques authentiques et non vues lors de l'apprentissage.

Afin de compléter le processus et d'atteindre un équilibre strict entre les classes CKD et non-CKD au sein du jeu d'entraînement, un RandomOverSampler a ensuite été appliqué. Ce double mécanisme a permis d'augmenter le volume global du dataset à 1000 instances équilibrées, améliorant ainsi la stabilité de l'apprentissage et la robustesse statistique des modèles développés.

Cette approche est particulièrement justifiée dans le domaine biomédical, où les données sont à la fois coûteuses à collecter, sensibles sur le plan éthique et souvent naturellement déséquilibrées. Certaines catégories de patients, telles que les sujets non-CKD, les patients aux stades précoces de la maladie rénale chronique ou encore les profils cliniques atypiques, sont généralement sous-représentées dans les bases de données existantes. Or, ces cas minoritaires sont précisément ceux qui présentent un intérêt clinique majeur, notamment pour la détection précoce d'une pathologie silencieuse comme la MRC.

Ainsi, l'augmentation des données par SMOTENC et RandomOverSampler ne vise pas uniquement à accroître la taille du jeu de données, mais surtout à renforcer la représentativité des profils cliniquement critiques au sein du jeu d'entraînement. Cette stratégie permet de concevoir des modèles plus équilibrés, capables de généraliser efficacement et de fournir des prédictions fiables, y compris pour les situations les moins fréquentes mais les plus déterminantes sur le plan clinique.

Enfin, une analyse descriptive réalisée avant et après l'augmentation a confirmé la préservation des distributions des variables clés (âge, créatinine sérique, eGFR). L'absence de distorsions majeures ou de valeurs aberrantes atteste de la qualité du processus d'augmentation et garantit la cohérence clinique du jeu de données enrichi.

5.8.2 Distribution des caractéristiques après l'augmentation des données

Afin d'assurer la validité scientifique des données enrichies, nous avons évalué la distribution des principales variables biomédicales avant et après l'augmentation réalisée à l'aide de SMOTENC et du RandomOverSampler. Les comparaisons montrent que les distributions des variables continues (telles que l'âge, la créatinine sérique, la pression artérielle ou le DFG) et catégorielles (par exemple le diabète sucré) demeurent globalement stables. Cette cohérence

indique que les méthodes utilisées n'ont pas introduit de valeurs aberrantes ou médicalement incohérentes. Plus encore, la combinaison de SMOTENC et du RandomOverSampler a permis de corriger efficacement le déséquilibre des classes tout en préservant l'intégrité clinique et la représentativité des données initiales. Ainsi, le jeu de données augmenté conserve sa pertinence biomédicale et constitue une base fiable pour les étapes ultérieures de modélisation prédictive et d'analyse des stades précoces de la MRC.

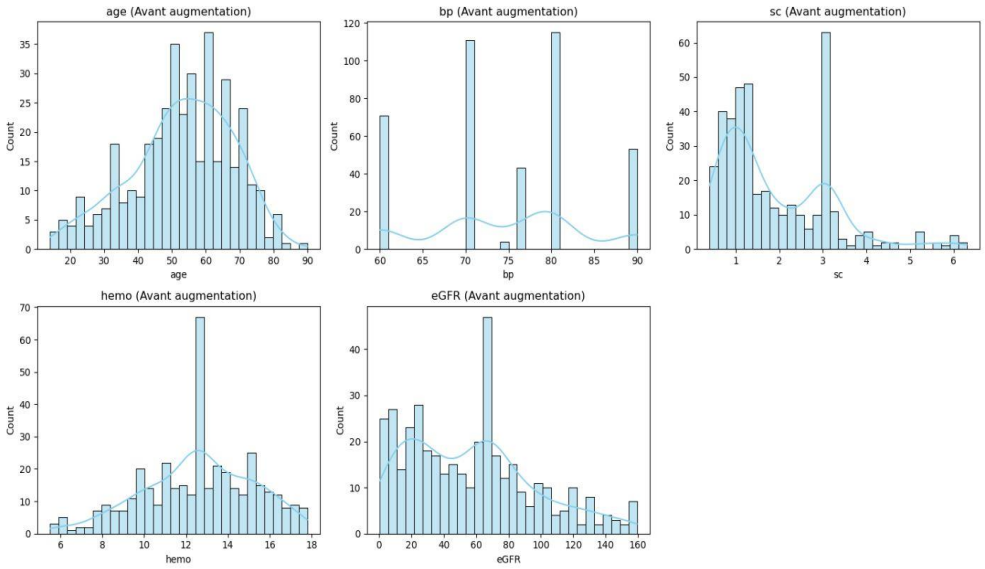


Figure 5.21: Distribution des caractéristiques avant augmentation

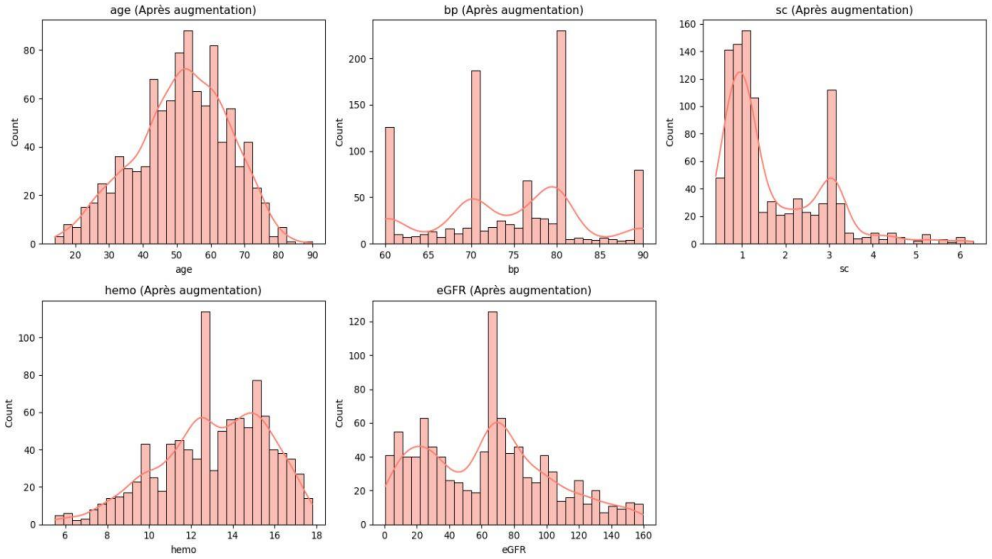


Figure 5.22: Distribution des caractéristiques apprès augmentation

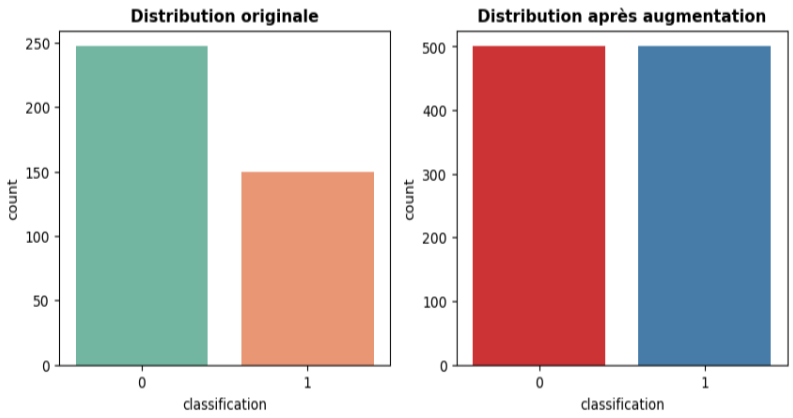


Figure 5.23: Distribution de la classe cible avant après augmentation

5.8.3 Validation de l'intégrité et de la cohérence des données augmentées

Afin de garantir que l'augmentation des données n'introduise ni erreurs ni biais, nous avons procédé à une validation rigoureuse de l'intégrité du jeu de données. Nous avons d'abord vérifié l'absence de valeurs nulles et de doublons, confirmant que le suréchantillonnage n'a pas compromis la qualité des données originales. Ces contrôles confirment que chaque variable numérique et chaque indicateur clinique restent cohérents, et que les nouvelles instances générées s'intègrent de manière cohérente dans la distribution initiale. L'application combinée de SMOTENC et RandomOverSampler a ainsi permis d'augmenter significativement la taille et l'équilibre du jeu de données tout en préservant sa fiabilité biomédicale et statistique, condition indispensable pour entraîner des modèles prédictifs robustes et cliniquement pertinents en néphrologie.

```
=== Doublons et valeurs nulles avant/après augmentation ===
          Variable  Nulls Before  Nulls After
               age             0            0
                bp             0            0
                sg             0            0
                al             0            0
                su             0            0
               bgr             0            0
                bu             0            0
                sc             0            0
               sod             0            0
               pot             0            0
              hemo             0            0
               pcv             0            0
                wc             0            0
                rc             0            0
               rbc             0            0
                pc             0            0
               pcc             0            0
                ba             0            0
               htn             0            0
                dm             0            0
               cad             0            0
             appet             0            0
                pe             0            0
               ane             0            0
              eGFR             0            0
    classification             0            0
Global Duplicates              0            0
```

Figure 5.24: Vérification de l'intégrité des données avant et après augmentation

5.8.4 Résultats expérimentaux et discussion

Cette section analyse de manière approfondie les performances des modèles de classification proposés pour la prédiction de la maladie rénale chronique (MRC), en s'appuyant sur une procédure de validation croisée avec différents nombres de plis (K = 10, 20 et 30). Ce choix méthodologique vise à évaluer la stabilité des modèles face aux variations des sous-ensembles d'apprentissage et de test, et à vérifier leur capacité de généralisation dans un contexte de données médicales limitées et hétérogènes. Les performances ont été évaluées à l'aide de métriques standard et cliniquement pertinentes, à savoir l'accuracy, la precision, le recall et le F1-score, afin d'obtenir une évaluation globale et équilibrée du comportement des classificateurs.

Les résultats obtenus mettent en évidence une faible sensibilité des performances au nombre de folds, traduisant une robustesse globale de l'approche proposée. Le modèle Random Forest présente des performances remarquablement stables, avec une accuracy constante de 99,4 % pour les trois configurations de validation croisée, une precision atteignant 100 % et un F1-score de 99,48 %. Cette constance indique une forte capacité de généralisation et suggère que le modèle parvient à exploiter efficacement les interactions complexes entre les variables cliniques, tout en limitant le risque de sur-apprentissage.

Le classificateur Decision Tree affiche également des performances élevées, avec une accuracy comprise entre 98,2 % et 98,9 %. Son recall élevé et invariant (99,4 %) souligne une excellente capacité à identifier les patients atteints de MRC, ce qui est particulièrement important dans un contexte clinique. Toutefois, de légères fluctuations de la precision selon le nombre de folds indiquent une sensibilité accrue aux variations des données d'apprentissage, un comportement attendu pour ce type de modèle.

Le modèle AdaBoost confirme l'intérêt des techniques d'ensembles basées sur le boosting, avec des valeurs d'accuracy proches de 99 % et un F1-score compris entre 98,7 % et 98,9 %. La stabilité de ces résultats pour différents nombres de plis met en évidence l'apport d'AdaBoost dans l'amélioration de la robustesse prédictive, notamment en réduisant l'impact des erreurs commises par les classificateurs faibles.

Le classificateur Gaussian Naive Bayes, bien que présentant des performances globales légèrement inférieures (accuracy de 96,1 % et precision de 92,9 %), se distingue par un recall de 100 % pour l'ensembles des configurations de validation croisée. Cette caractéristique indique une excellente capacité de détection des patients atteints de MRC, au prix d'un taux plus élevé de faux positifs. Dans une perspective clinique, ce compromis reste acceptable, la priorité étant accordée à la minimisation des faux négatifs afin d'éviter la non-détection de cas pathologiques.

Dans l'ensembles, la comparaison des performances obtenues pour différents nombres de folds confirme que l'enrichissement et l'augmentation raisonnée du jeu de données, combinés à une stratégie de validation croisée rigoureuse, permettent de développer des modèles à la fois performants, stables et généralisables. Ces résultats soutiennent la pertinence de l'approche proposée pour une prédiction précoce et fiable de la MRC, et renforcent son potentiel d'intégration dans des systèmes d'aide à la décision clinique. Les résultats quantitatifs obtenus

pour chaque modèle et pour les différentes configurations de validation croisée sont synthétisés dans le tableau suivant, tandis que les figures associées illustrent de manière comparative l'évolution de chaque métrique de performance (accuracy, precision, recall, F1-score et AUC).

Tableau 5.3: Évaluation de différents classificateurs pour différents ratios de division basés sur une validation croisée k-fold

Métrique	K-Folds	Modèles			
		RF	DT	GNB	AdaBoost
Accuracy %	10	99.4	98.2	96.1	98.9
	20	99.4	98.9	96.1	98.7
	30	99.4	98.4	96.1	98.9
Précision %	10	100	97.02	92.9	98.01
	20	100	98.4	92.9	98
	30	100	97.5	92.9	98.01
Recall %	10	98.98	99.4	100	100
	20	98.98	99.4	100	99.49
	30	98.98	99.4	100	99.8
F1-score %	10	99.48	98.2	96.33	98.9
	20	99.48	98.9	96.3	98.7
	30	99.48	98.4	96.3	98.9
Taux d'erreurs	10	0.003	0.01	0.05	0.003
	20	0.003	0.01	0.05	0.003
	30	0.003	0.01	0.05	0.005
AUC %	10	99.98	98.22	98.17	99.98
	20	99.97	98.98	98.15	99.97
	30	99.97	98.47	98.14	99.98

5.8.4.1. Analyse des courbes ROC et des valeurs d'AUC

L'évaluation des performances des modèles a été complétée par l'analyse des courbes ROC (Receiver Operating Characteristic) et des valeurs de l'aire sous la courbe (AUC), un indicateur largement utilisé pour mesurer la capacité discriminante d'un classificateur indépendamment du seuil de décision. Cette métrique est particulièrement pertinente dans un contexte médical, car elle permet d'évaluer le compromis entre le taux de vrais positifs et le taux de faux positifs sur l'ensembles des seuils possibles.

Les résultats obtenus montrent des valeurs d'AUC très élevées pour l'ensembles des modèles étudiés, traduisant une excellente capacité de discrimination entre les patients atteints de MRC et les sujets non-CKD. Les modèles Random Forest et AdaBoost se distinguent par des performances quasi parfaites, avec une AUC de 0,999, indiquant une séparation presque optimale entre les deux classes. Cette performance confirme l'efficacité des méthodes d'ensembles dans la capture des relations complexes et non linéaires présentes dans les données cliniques.

Le modèle Decision Tree présente également une AUC élevée de 0,989, témoignant d'une très bonne capacité discriminante, bien que légèrement inférieure à celle des modèles d'ensembles. Le classificateur Gaussian Naive Bayes affiche une AUC de 0,981, ce qui reste très satisfaisant et confirme sa pertinence pour la détection de la maladie, malgré les hypothèses simplificatrices sur lesquelles il repose.

L'analyse des courbes ROC confirme les résultats observés à travers les métriques classiques de performance et met en évidence la robustesse et la fiabilité des modèles proposés, en particulier des approches basées sur les méthodes d'ensembles. Ces résultats renforcent la pertinence de l'approche développée pour une intégration potentielle dans des systèmes d'aide à la décision clinique dédiés à la prédiction précoce de la maladie rénale chronique.

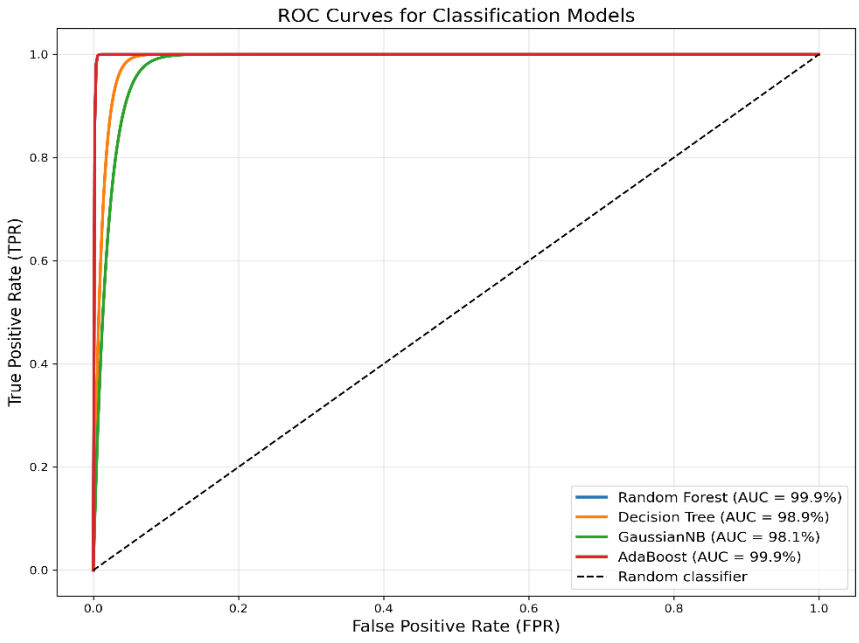


Figure 5.25: Courbe ROC des modèles RF, DT, GNB et AdaBoost

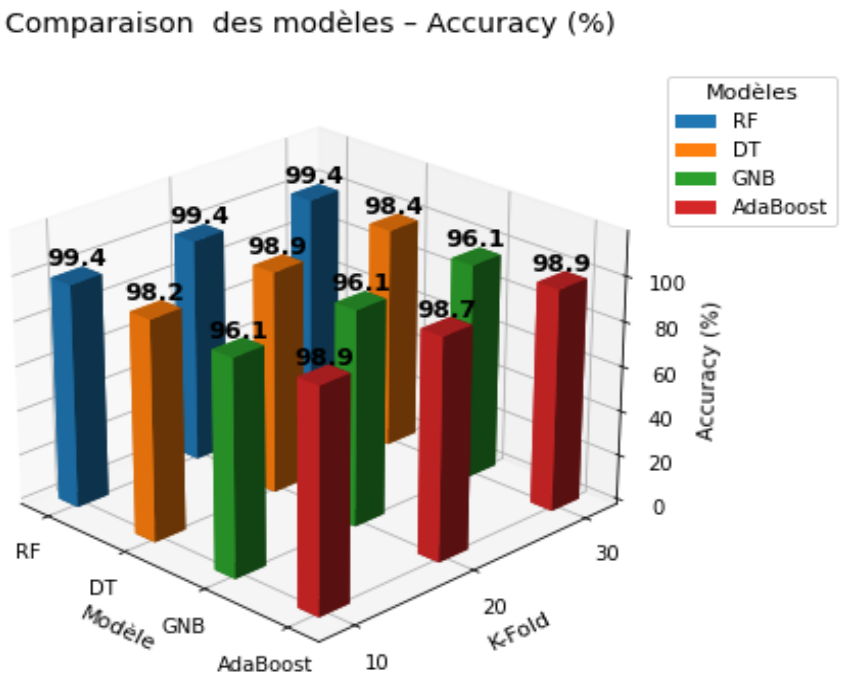


Figure 5.26: Performance en Accuracy des classificateurs selon plusieurs ratios de validation croisée k-fold

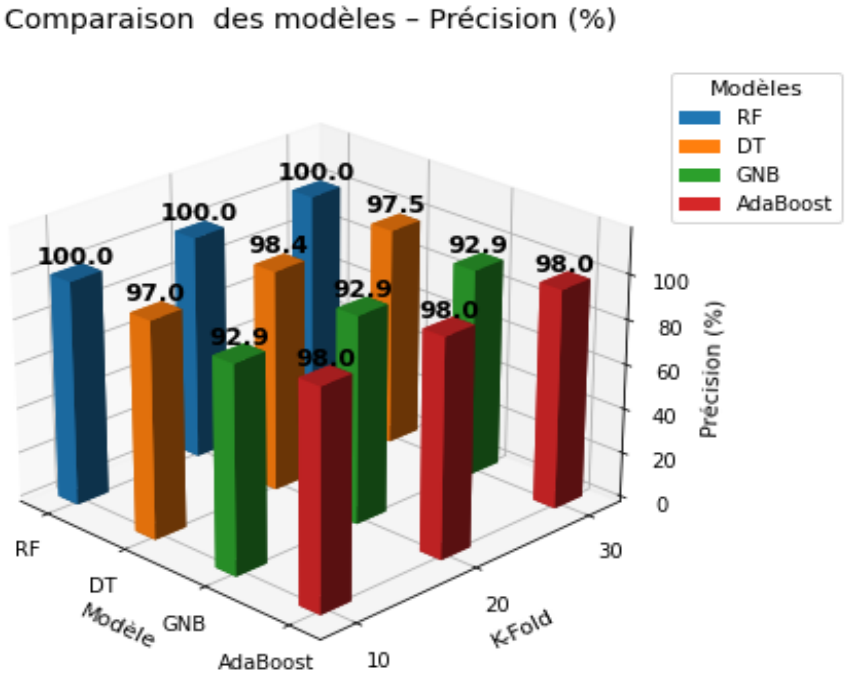


Figure 5.27: Performance en Précision des classificateurs selon plusieurs ratios de validation croisée k-fold

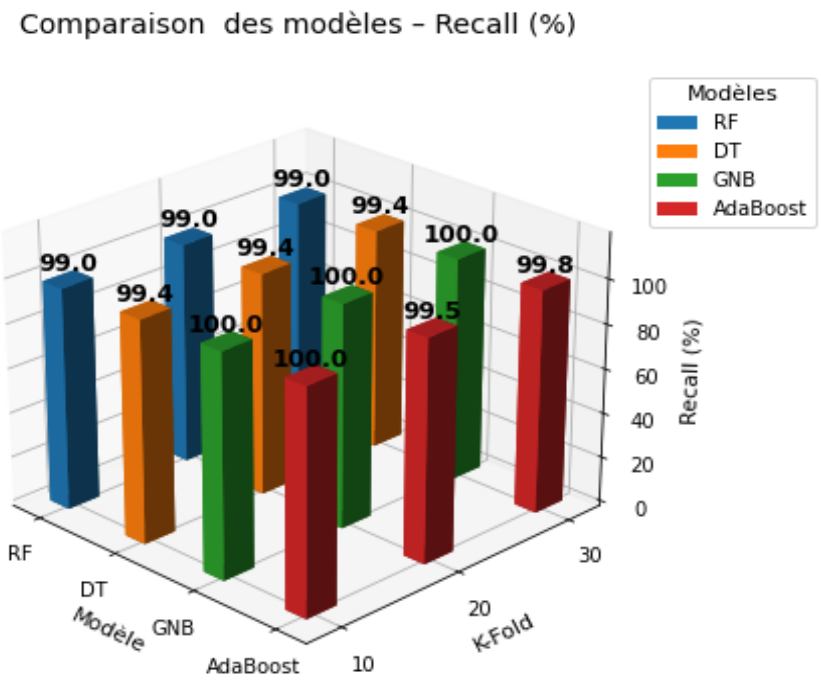


Figure 5.28: Performance en Recall des classificateurs selon plusieurs ratios de validation croisée k-fold

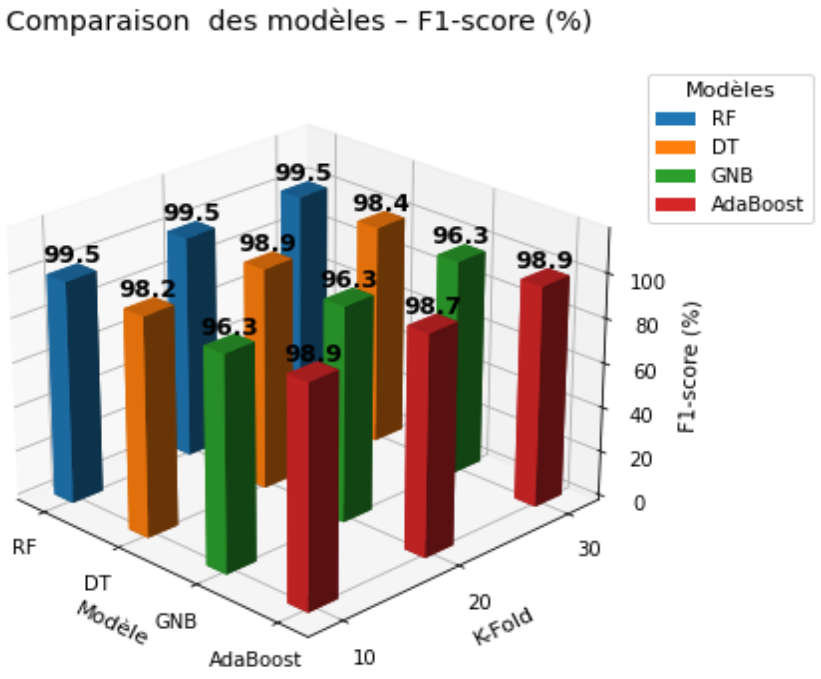


Figure 5.29: Performance en F1-Score des classificateurs selon plusieurs ratios de validation croisée k-fold

5.8.4.2. Évaluation des modèles sur l'ensembles de test

Afin d'évaluer la capacité réelle de généralisation des modèles proposés, une phase d'évaluation finale a été conduite sur un ensembles de test indépendant composé de 200 instances, strictement séparé des données d'entraînement et non soumis à aucun processus d'augmentation. Cette étape est essentielle pour mesurer les performances des modèles dans des conditions proches d'un usage clinique réel, où les données sont entièrement nouvelles et non vues lors de l'apprentissage.

Les résultats obtenus confirment les tendances observées lors de la validation croisée. Le modèle Random Forest affiche des performances parfaites sur l'ensembles des métriques évaluées, avec une accuracy, une precision, un recall et un F1-score atteignant 100 %. Ces résultats traduisent une excellente capacité de discrimination entre les patients CKD et non-CKD, ainsi qu'une absence de faux positifs et de faux négatifs sur l'échantillon de test considéré.

Le modèle Decision Tree présente également des performances très élevées, avec une accuracy de 99,5 %, une precision de 98,67 %, un recall de 100 % et un F1-score de 99,33 %. Ce comportement met en évidence une capacité optimale à détecter l'ensembles des cas pathologiques, au prix d'un très faible nombre de faux positifs, ce qui demeure acceptable dans un contexte clinique.

Le classificateur AdaBoost obtient une accuracy de 99 %, associée à une precision parfaite de 100 % et un recall de 97,22 %, conduisant à un F1-score de 98,59 %. Ces résultats indiquent que le modèle est particulièrement fiable lorsqu'il prédit la présence de la maladie, tout en générant un nombre limité de faux négatifs, ce qui confirme l'apport du boosting pour améliorer la robustesse de la classification

Enfin, le modèle Gaussian Naive Bayes, bien que légèrement moins performant que les autres classificateurs, conserve des résultats satisfaisants avec une accuracy de 97,5 %, une precision de 98,59 %, un recall de 95,59 % et un F1-score de 96,55 %. Cette diminution relative des performances peut s'expliquer par les hypothèses d'indépendance conditionnelle propres à ce modèle, moins adaptées à la complexité et aux corrélations présentes dans les données cliniques.

Dans l'ensembles, l'évaluation sur l'ensembles de test indépendant confirme la robustesse, la fiabilité et la capacité de généralisation des modèles proposés, en particulier des approches basées sur les méthodes d'ensembles. Ces résultats renforcent le potentiel de l'approche développée pour une application en aide à la décision clinique, notamment dans un cadre de prédiction précoce de la maladie rénale chronique. Les matrices de confusion associées aux résultats de test sont présentées par la suite afin d'illustrer de manière détaillée la répartition des vraies et fausses classifications pour chaque modèle.

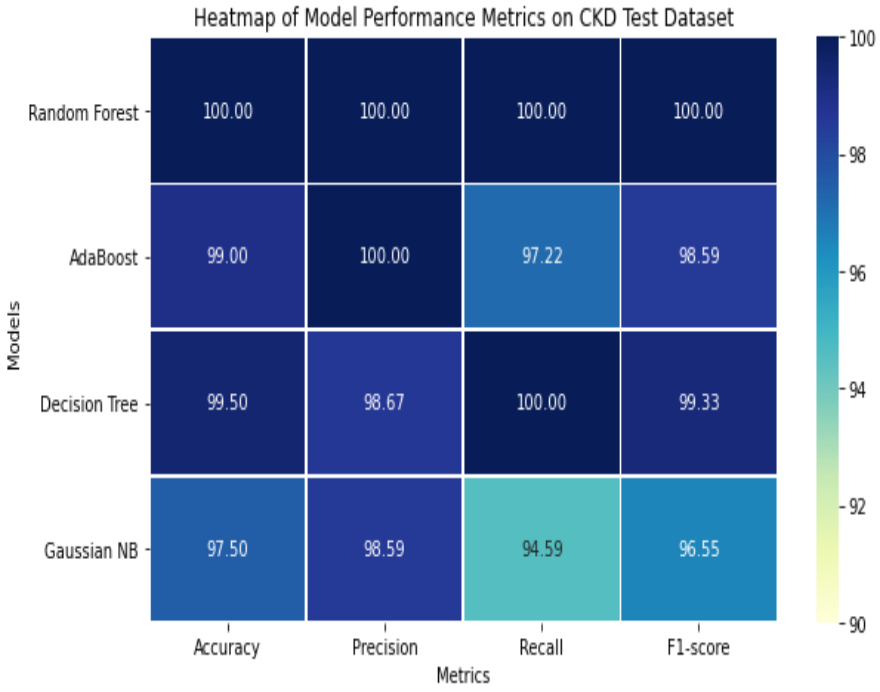


Figure 5.30: Evaluation des performances des modèles sur le datatset de test

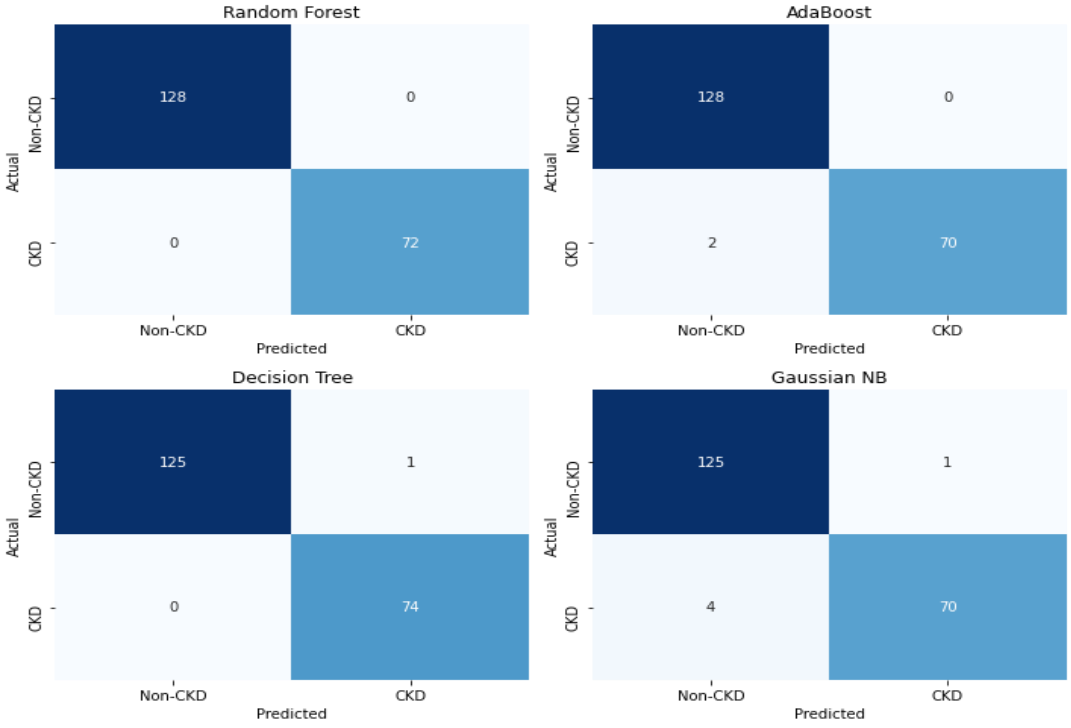


Figure 5.31: Matrices de confusion des modèles sur le datatset de test

5.8.5 Comparaison des résultats obtenus avec d'autres travaux connexes

Afin d'évaluer de manière rigoureuse la performance de notre approche, un tableau comparatif a été élaboré, regroupant les résultats obtenus dans cette étude et ceux rapportés dans des travaux connexes. Cette analyse comparative met en évidence des différences notables en termes de précision, de robustesse et de stabilité des performances, tout en tenant compte du contexte d'application et des caractéristiques spécifiques des ensembles de données utilisés.

Notre méthodologie se distingue des études antérieures par plusieurs choix méthodologiques clés. Tout d'abord, le jeu de données a été enrichi par le calcul du débit de filtration glomérulaire estimé (eGFR), permettant une meilleure caractérisation clinique de la fonction rénale et une distinction plus fine des stades de la maladie. Ensuite, afin de traiter le déséquilibre des classes, une stratégie d'augmentation raisonnée des données a été adoptée, reposant sur les techniques SMOTENC et RandomOverSampler, spécifiquement adaptées aux ensembles de données mixtes intégrant des variables numériques et catégorielles.

Il convient de préciser que ces techniques d'augmentation ont été appliquées exclusivement au sous-ensemble d'entraînement, après la phase de séparation des données, tandis que l'ensembles de test a été conservé dans son état réel. Ce choix méthodologique vise à prévenir toute fuite d'information et à garantir une évaluation objective et réaliste des performances des modèles. L'augmentation n'a ainsi pas pour objectif d'augmenter artificiellement la taille globale du jeu de données, mais de renforcer la représentativité des classes au sein du processus d'apprentissage, en particulier pour les profils cliniques minoritaires.

Par ailleurs, l'évaluation des modèles s'appuie sur une validation croisée réalisée pour trois valeurs de k (10, 20 et 30), permettant d'analyser la stabilité des performances face aux variations des sous-échantillons d'apprentissage. Cette procédure contribue à une estimation plus fiable de la capacité de généralisation des classificateurs.

L'ensembles de ces éléments souligne la valeur ajoutée de notre contribution par rapport aux travaux existants et démontre la compétitivité, la robustesse et la pertinence clinique de l'approche proposée pour la prédiction précoce de la maladie rénale chronique.

Tableau 5.4: Comparaison de l'approche proposée avec d'autres travaux

Réf	**Dataset**	**Méthode d'augmentation**	**Modèles**	**Validation**	**Accuracy (%)**	**Innovations**
[160]	400,24	–	Extra Tree	–	98	–
[190]	400,24	–	KNN, SVM	–	98.9	–
[191]	400,24	–	DT	15 folds	97.23	Validation sur k-fold,
[182]	400,24	–	GNB	–	85.5	–
Notre approche	**400,26** **200,28** **Dataset augmenté**	**SMOTENC, RandomOverSampler**	**RF, DT, GNB, AdaBoost, ANN, Hybride**	**10,20 et 30 folds** **+ test indépendant**	**99.7**	**Enrichissement clinique (DFG, stade), augmentation contrôlée, validation multi-k, test sur données réelles**

5.9 Conclusion

Dans ce chapitre a présenté une série d'expérimentations visant à améliorer la prédiction de la maladie rénale chronique. Nous avons d'abord enrichi le jeu de données initial en intégrant le eGFR, un biomarqueur clinique majeur, afin de renforcer la pertinence médicale des variables et de permettre la détection précoce des stades silencieux de la maladie. Ensuite, nous avons surmonté les limitations liées à la taille réduite et au déséquilibre des classes en augmentant le dataset initial grâce aux méthodes SMOTENC et RandomOverSampler, garantissant à la fois la cohérence clinique et l'équilibre des classes. Pour l'évaluation des modèles, nous avons employé plusieurs algorithmes d'apprentissage supervisé : Random Forest, Decision Tree, AdaBoost et Gaussian Naive Bayes et effectué une validation croisée sur plusieurs valeurs de k (10, 20 et 30) afin d'assurer la robustesse et la généralisabilité des résultats. Les performances finales, présentées sous forme de matrices de confusion et de métriques classiques, ont montré la supériorité des modèles d'ensembles et la pertinence des variables sélectionnées. Enfin, une comparaison avec les travaux existants a permis de situer notre contribution dans un contexte scientifique plus large et de démontrer la valeur ajoutée de notre approche en termes de précision, de fiabilité et de représentativité clinique.

Conclusion générale et perspectives

Conclusion générale et perspectives

La prédiction de la maladie rénale chronique (MRC) constitue un enjeu majeur de santé publique à l'échelle mondiale. Cette pathologie silencieuse, souvent observée à un stade avancé, entraîne une dégradation progressive de la fonction rénale pouvant aboutir à l'insuffisance rénale terminale. Dans ce contexte, le recours à l'intelligence artificielle, notamment aux algorithmes d'apprentissage automatique, représente une prometteuse avancée pour détecter précocement les cas à risque, optimiser les décisions cliniques et améliorer la qualité de vie des patients. C'est dans cette optique que s'inscrit notre travail de thèse, qui vise à développer des modèles intelligents, robustes et fiables pour la prédiction précoce de la MRC à partir de données médicales réelles.

Tout au long de ce travail, nous avons mené une série d'expérimentations rigoureuses, organisées en quatre contributions complémentaires. Dans la première, nous avons comparé différentes techniques de sélection de caractéristiques (LASSO, RFE, RELIEF) afin d'identifier les variables les plus significatives pour la prédiction de la MRC, tout en évaluant plusieurs classificateurs traditionnels et hybrides. Dans la deuxième contribution, nous avons enrichi notre protocole expérimental en intégrant un deuxième jeu de données indépendant, permettant de valider la capacité de généralisation de nos modèles sur des cas cliniques inédits. La troisième contribution a consisté à structurer une approche prédictive complète et rigoureuse, intégrant des étapes de nettoyage, transformation, normalisation, sélection des attributs et enrichissement du jeu de données par l'ajout du taux de filtration glomérulaire notable (eGFR), un indicateur clinique fondamental rarement, voire jamais, utilisé dans les travaux antérieurs. Enfin, dans la quatrième contribution, nous avons cherché à dépasser la limitation du faible volume de données initial en appliquant des techniques d'augmentation synthétique (SMOTENC et RandomOverSampler), en combinaison avec des modèles tels que RF, Decision Tree, GNB et AdaBoost.

Les résultats obtenus ont démontré une significative de la performance prédictive, notamment grâce à l'enrichissement du jeu de données et à l'équilibrage des classes. L'ensembles des modèles testés, a montré des taux élevés d'exactitude, de précision et de rappel, confirmant la pertinence de notre méthodologie. De plus, l'utilisation d'un jeu de test externe valide l'applicabilité de notre approche dans un contexte médical réel.

Afin de prolonger et valoriser cette recherche, plusieurs pistes d'approfondissement sont envisagées :

- **Création d'un jeu de données local propre à la population algérienne (en cours de réalisation) :** cette perspective vise à adapter les modèles aux réalités cliniques spécifiques à notre région, en tenant compte des particularités épidémiologiques, génétiques et environnementales.

- **Intégration de données longitudinales** : l'analyse de l'évolution des patients dans le temps permet de passer d'une prédiction statique à une modélisation dynamique de la progression de la MRC.

- **Développement de modèles explicables (XAI)** : il est fondamental que les professionnels de santé puissent comprendre et interpréter les décisions prises par les modèles. L'intelligibilité des résultats facilitera leur adoption clinique.

- **Déploiement d'un prototype d'outil décisionnel** : à terme, nous envisageons le développement d'une interface intelligente ou d'une application intégrant nos modèles, afin de les rendre accessibles aux praticiens hospitaliers et médecins généralistes.

- **Extension aux comorbidités fréquentes** : notre approche pourra être étendue à d'autres maladies chroniques associées à la MRC telles que l'hypertension artérielle, en vue d'une prise en charge globale et anticipée.

  En conclusion, cette thèse pose les fondements d'un système intelligent d'aide au diagnostic précoce de la MRC, fondé sur une rigueur méthodologique, une pertinence clinique, et une ouverture vers des applications concrètes au service de la médecine prédictive et personnalisée.

Références bibliographiques

Références bibliographiques

[1] "Your Kidneys & How They Work , National Institute of Diabetes and Digestive and Kidney Diseases," 2018. https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work (accessed Jan. 31, 2024).

[2] "The Kidneys and How They Work, National Kidney Foundation," 2024. https://www.kidney.org/about/kidney-disease-fact-sheet

[3] "Les fonctions du rein, France Rein," 2021. https://www.francerein.org/vivre-avec-la-maladie/maladies-et-traitements/les-fonctions-du-rein/

[4] "LE REIN : QU'EST-CE QUE C'EST ?,VIDAL," 2022. https://www.vidal.fr/agir-pour-ma-sante/don-rein/rein.htm

[5] "Maladie rénale chronique et insuffisance rénale chronique : définition et causes," 2025. https://www.ameli.fr/assure/sante/themes/maladie-renale-chronique/comprendre-maladie-renale-chronique

[6] "Comprendre l'Anatomie et les Fonctions des Reins : Prévention et Traitement des Maladies Rénales,DDG Gastro," 2024. http://ddg-gastro.be/organe-rein-guide-comple/

[7] "Le rein, IFSI-IFAS Saverne," 2022. [Online]. Available: https://www.ifsi-ifas-saverne.fr/wp-content/uploads/2020/02/UE-2.7-S4-Le-rein-4.02.19.pdf

[8] C. S. Charlène Le Neindre, Damien Bricard, M. L. Florian Bayer, Cécile Couchoud, C. Le Neindre1, D. Bricard1, and C. Sermet1, "ATLAS de l'insuffisance rénale chronique terminale en France," p. 10, 2018, [Online]. Available: https://www.irdes.fr/recherche/ouvrages/004-atlas-de-l-insuffisance-renale-chronique-terminale-en-france.pdf

[9] Z. A. Miller and K. Dwyer, "Artificial Intelligence to Predict Chronic Kidney Disease Progression to Kidney Failure: A Narrative Review," *Nephrology*, vol. 30, no. 1, Jan. 2025, doi: 10.1111/nep.14424.

[10] C. P. Kovesdy, "Epidemiology of chronic kidney disease: an update 2022," *Kidney Int. Suppl.*, vol. 12, no. 1, pp. 7–11, Apr. 2022, doi: 10.1016/j.kisu.2021.11.003.

[11] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," *J. Med. Syst.*, vol. 41, no. 4, p. 55, Apr. 2017, doi: 10.1007/s10916-017-0703-x.

[12] M.-H. Tsai *et al.*, "Incidence, Prevalence, and Duration of Chronic Kidney Disease in Taiwan: Results from a Community-Based Screening Program of 106,094 Individuals," *Nephron*, vol. 140, no. 3, pp. 175–184, 2018, doi: 10.1159/000491708.

[13] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.

[14] H. S. Shad, Z. Jamal, S. M. F. Ahmed, S. Momen, and N. Mansoor, "Early Prediction of Chronic Kidney Disease Using Data Mining Techniques," 2021, pp. 947–957. doi: 10.1007/978-3-030-90321-3_79.

[15] L. Zhang *et al.*, "Prevalence of chronic kidney disease in China: a cross-sectional survey," *Lancet*, vol. 379, no. 9818, pp. 815–822, Mar. 2012, doi: 10.1016/S0140-6736(12)60033-6.

[16] Z. Chen, X. Zhang, and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," *Int. Urol. Nephrol.*, vol. 48, no. 12, pp. 2069–2075, Dec. 2016, doi: 10.1007/s11255-016-1346-4.

[17] N. R. Hill *et al.*, "Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis," *PLoS One*, vol. 11, no. 7, p. e0158765, Jul. 2016, doi: 10.1371/journal.pone.0158765.

[18] "What is chronic kidney disease?," *What is chronic kidney disease?* https://fondation-du-rein.org/la-maladie-renale-chronique/

[19] N. Meguellati *et al.*, "Épidémiologie de la maladie rénale chronique stade V traitée dans le Sud-Est algérien : étude prospective de 2016 à 2018," *Néphrologie & Thérapeutique*, vol. 19, no. 1, pp. 45–57, Feb. 2023, doi: 10.1684/ndt.2023.4.

[20] R. Noble and M. W. Taal, "Epidemiology and causes of chronic kidney disease," *Medicine (Baltimore).*, vol. 47, no. 9, pp. 562–566, Sep. 2019, doi: 10.1016/j.mpmed.2019.06.010.

[21] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, "Diabetic Kidney Disease," *Clin. J. Am. Soc. Nephrol.*, vol. 12, no. 12, pp. 2032–2045, Dec. 2017, doi: 10.2215/CJN.11491116.

[22] P. D. Evans and M. W. Taal, "Epidemiology and causes of chronic kidney disease," *Medicine (Baltimore).*, vol. 43, no. 8, pp. 450–453, Aug. 2015, doi: 10.1016/j.mpmed.2015.05.005.

[23] P. D. Evans and M. W. Taal, "Epidemiology and causes of chronic kidney disease," *Medicine (Baltimore).*, vol. 39, no. 7, pp. 402–406, Jul. 2011, doi: 10.1016/j.mpmed.2011.04.007.

[24] V. et al Olié, "Prévalence de la maladie rénale chronique en France, Esteban 2014–2016. sante publique France," 2021.

[25] M. J. Klag *et al.*, "Blood Pressure and End-Stage Renal Disease in Men," *N. Engl. J. Med.*, vol. 334, no. 1, pp. 13–18, Jan. 1996, doi: 10.1056/NEJM199601043340103.

[26] J. Floege *et al.*, "Management and treatment of glomerular diseases (part 1): conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference," *Kidney Int.*, vol. 95, no. 2, pp. 268–280, Feb. 2019, doi: 10.1016/j.kint.2018.10.018.

[27] E. Cornec-Le Gall, A. Alam, and R. D. Perrone, "Autosomal dominant polycystic kidney disease," *Lancet*, vol. 393, no. 10174, pp. 919–935, Mar. 2019, doi: 10.1016/S0140-6736(18)32782-X.

[28] C. Bergmann, L. M. Guay-Woodford, P. C. Harris, S. Horie, D. J. M. Peters, and V. E. Torres, "Polycystic kidney disease," *Nat. Rev. Dis. Prim.*, vol. 4, no. 1, p. 50, Dec. 2018, doi: 10.1038/s41572-018-0047-y.

[29] S. Mäkimattila, V. Harjutsalo, C. Forsblom, and P.-H. Groop, "Response to Comment on Mäkimattila et al. Every Fifth Individual With Type 1 Diabetes Suffers From an Additional Autoimmune Disease: A Finnish Nationwide Study. Diabetes Care 2020;43:1041–1047," *Diabetes Care*, vol. 43, no. 8, pp. e106–e107, Aug. 2020, doi: 10.2337/dci20-0020.

[30] P. K. Whelton *et al.*, "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr," *Hypertension*, vol. 71, no. 6, Jun. 2018, doi: 10.1161/HYP.0000000000000065.

[31] D. Ettehad *et al.*, "Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis," *Lancet*, vol. 387, no. 10022, pp. 957–967, Mar. 2016, doi: 10.1016/S0140-6736(15)01225-8.

[32] A. K. Cheung *et al.*, "KDIGO 2021 Clinical Practice Guideline for the Management of Blood Pressure in Chronic Kidney Disease," *Kidney Int.*, vol. 99, no. 3, pp. S1–S87, Mar. 2021, doi: 10.1016/j.kint.2020.11.003.

[33] D. Zickler *et al.*, "Tumour necrosis factor-alpha in uraemic serum promotes osteoblastic transition and calcification of vascular smooth muscle cells via extracellular signal-regulated kinases and activator protein 1/c-FOS-mediated induction of interleukin 6 expression," *Nephrol. Dial. Transplant.*, vol. 33, no. 4, pp. 574–585, Apr. 2018, doi: 10.1093/ndt/gfx316.

[34] C. P. Kovesdy *et al.*, "Obesity and kidney disease: hidden consequences of the epidemic," *Kidney Int.*, vol. 91, no. 2, pp. 260–262, Feb. 2017, doi: 10.1016/j.kint.2016.10.019.

[35] H. Lin, D. Xiao, Z. Liu, Q. Shi, P. Hajek, and C. Wang, “National survey of smoking cessation provision in China,” *Tob. Induc. Dis.*, vol. 17, no. April, Apr. 2019, doi: 10.18332/tid/104726.

[36] T. Shafi *et al.*, “Estimating residual kidney function in dialysis patients without urine collection,” *Kidney Int.*, vol. 89, no. 5, pp. 1099–1110, May 2016, doi: 10.1016/j.kint.2015.10.011.

[37] E. D. O’Sullivan, J. Hughes, and D. A. Ferenbach, “Renal Aging: Causes and Consequences,” *J. Am. Soc. Nephrol.*, vol. 28, no. 2, pp. 407–420, Feb. 2017, doi: 10.1681/ASN.2015121308.

[38] C. E. Kashtan, “Alport syndrome and thin glomerular basement membrane disease,” *J. Am. Soc. Nephrol.*, vol. 11(9), pp. 1736–1750, 2000, doi: 10.1681/ASN.V1191736.

[39] J.-M. Krzesinski and A. J. Scheen, “La [b]maladie[/b] rénale diabétique : prise en charge actuelle et perspectives d’avenir,” *Rev. Med. Suisse*, vol. 11, no. 483, pp. 1534–1542, 2015, doi: 10.53738/REVMED.2015.11.483.1534.

[40] et al. Niewczas, M. A., “Circulating TNF receptors 1 and 2 predict ESRD in type 2 diabetes,” *J. Am. Soc. Nephrol.*, vol. 30(5), pp. 804–816, 2019, doi: 10.1681/ASN.2018070683.

[41] J. E. Flythe *et al.*, “Characteristics and Outcomes of Individuals With Pre-existing Kidney Disease and COVID-19 Admitted to Intensive Care Units in the United States,” *Am. J. Kidney Dis.*, vol. 77, no. 2, pp. 190-203.e1, Feb. 2021, doi: 10.1053/j.ajkd.2020.09.003.

[42] A. C. Webster, E. V Nagler, R. L. Morton, and P. Masson, “Chronic Kidney Disease,” *Lancet*, vol. 389, no. 10075, pp. 1238–1252, Mar. 2017, doi: 10.1016/S0140-6736(16)32064-5.

[43] Y. Zheng, S. H. Ley, and F. B. Hu, “Global aetiology and epidemiology of type 2 diabetes mellitus and its complications,” *Nat. Rev. Endocrinol.*, vol. 14, no. 2, pp. 88–98, Feb. 2018, doi: 10.1038/nrendo.2017.151.

[44] R. T. Gansevoort *et al.*, “Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention,” *Lancet*, vol. 382, no. 9889, pp. 339–352, Jul. 2013, doi: 10.1016/S0140-6736(13)60595-4.

[45] et al. Levey, A. S., “Chronic kidney disease,” *Lancet*, vol. 395, no. 10225, pp. 709–733, 2020, doi: 10.1016/S0140-6736(20)30043-6.

[46] M. Afkarian *et al.*, “Kidney Disease and Increased Mortality Risk in Type 2 Diabetes,” *J. Am. Soc. Nephrol.*, vol. 24, no. 2, pp. 302–308, Feb. 2013, doi: 10.1681/ASN.2012070718.

[47] I. H. de Boer, “Temporal Trends in the Prevalence of Diabetic Kidney Disease in the United States,” *JAMA*, vol. 305, no. 24, p. 2532, Jun. 2011, doi: 10.1001/jama.2011.861.

[48] M. E. Molitch *et al.*, “Diabetic kidney disease: a clinical update from Kidney Disease: Improving Global Outcomes,” *Kidney Int.*, vol. 87, no. 1, pp. 20–30, Jan. 2015, doi: 10.1038/ki.2014.128.

[49] C. S. Fox *et al.*, “Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis,” *Lancet*, vol. 380, no. 9854, pp. 1662–1673, Nov. 2012, doi: 10.1016/S0140-6736(12)61350-6.

[50] J. F. E. Mann *et al.*, “Liraglutide and Renal Outcomes in Type 2 Diabetes,” *N. Engl. J. Med.*, vol. 377, no. 9, pp. 839–848, Aug. 2017, doi: 10.1056/NEJMoa1616011.

[51] S. Halimi, “Épidémiologie des maladies rénales chez les patients diabétiques et place des marqueurs,” *Médecine des Mal. Métaboliques*, vol. 17, no. 8, pp. 614–626, Dec. 2023, doi: 10.1016/j.mmm.2023.10.011.

[52] V. Perkovic *et al.*, “Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy,” *N. Engl. J. Med.*, vol. 380, no. 24, pp. 2295–2306, Jun. 2019, doi: 10.1056/NEJMoa1811744.

[53] “Diabète et maladie rénale chronique, Maux Croisés.”

[54] I. H. de Boer *et al.*, “KDIGO 2020 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease,” *Kidney Int.*, vol. 98, no. 4, pp. S1–S115, Oct. 2020, doi: 10.1016/j.kint.2020.06.019.

[55] KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease, “Notice,” *Kidney Int. Suppl.*, vol. 3, no. 1, p. 1, Jan. 2013, doi: 10.1038/kisup.2012.73.

[56] A. S. Levey *et al.*, “A New Equation to Estimate Glomerular Filtration Rate,” *Ann. Intern. Med.*, vol. 150, no. 9, p. 604, May 2009, doi: 10.7326/0003-4819-150-9-200905050-00006.

[57] L. A. Inker *et al.*, “Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C,” *N. Engl. J. Med.*, vol. 367, no. 1, pp. 20–29, Jul. 2012, doi: 10.1056/NEJMoa1114248.

[58] “Notice,” *Kidney Int. Suppl.*, vol. 2, no. 1, p. 1, Mar. 2012, doi: 10.1038/kisup.2012.1.

[59] A. S. Levey *et al.*, “The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report,” *Kidney Int.*, vol. 80, no. 1, pp. 17–28, Jul. 2011, doi: 10.1038/ki.2010.483.

[60] L. A. Inker *et al.*, “New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race,” *N. Engl. J. Med.*, vol. 385, no. 19, pp. 1737–1749, Nov. 2021, doi: 10.1056/NEJMoa2102953.

[61] D. W. Cockcroft and H. Gault, “Prediction of Creatinine Clearance from Serum Creatinine,” *Nephron*, vol. 16, no. 1, pp. 31–41, 1976, doi: 10.1159/000180580.

[62] A. S. Levey, “A More Accurate Method To Estimate Glomerular Filtration Rate from Serum Creatinine: A New Prediction Equation,” *Ann. Intern. Med.*, vol. 130, no. 6, p. 461, Mar. 1999, doi: 10.7326/0003-4819-130-6-199903160-00002.

[63] Matthieu Mleczak, “Réalité du dépistage et du suivi de la maladie rénale chronique en médecine générale par recherche d’excrétion urinaire d’albumine : audit de pratique chez 10 médecins généralistes en secteur semi-rural sur 4 ans,” UNIVERSITE HENRI POINCARE, NANCY 1, 2023. doi: hal-04517133.

[64] “Évaluation du rapport albuminurie/créatininurie dans le diagnostic de la maladie rénale chronique chez l’adulte - Rapport d’évaluation,” *Haute Autorité de Santé*. https://www.has-sante.fr/jcms/c_1169049/fr/evaluation-du-rapport-albuminurie/creatininurie-dans-le-diagnostic-de-la-maladie-renale-chronique-chez-l-adulte-rapport-d-evaluation (accessed Feb. 07, 2025).

[65] “Précis de de biopathologie analyses médicales spécialisées . Microalbuminurie,” 2013. [Online]. Available: https://www.eurofins-biomnis.com/referentiel/liendoc/precis/MICROALBUMINE.pdf

[66] M. Haenlein and A. Kaplan, “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence,” *Calif. Manage. Rev.*, vol. 61, no. 4, pp. 5–14, Aug. 2019, doi: 10.1177/0008125619864925.

[67] S. D. Tagliaferri *et al.*, “Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews,” *npj Digit. Med.*, vol. 3, no. 1, p. 93, Jul. 2020, doi: 10.1038/s41746-020-0303-x.

[68] B. X. Tran *et al.*, “Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study,” *J. Clin. Med.*, vol. 8, no. 3, p. 360, Mar. 2019, doi: 10.3390/jcm8030360.

[69] R. Vinuesa *et al.*, “The role of artificial intelligence in achieving the Sustainable Development Goals,” *Nat. Commun.*, vol. 11, no. 1, p. 233, Jan. 2020, doi: 10.1038/s41467-019-14108-y.

[70] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, “Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery,” *Chem. Rev.*, vol. 119, no. 18, pp. 10520–10594, Sep. 2019, doi: 10.1021/acs.chemrev.8b00728.

[71] R. J. Burton, M. Albur, M. Eberl, and S. M. Cuff, “Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 171, Dec. 2019, doi: 10.1186/s12911-019-0878-9.

[72] B. Meskó, Z. Drobni, É. Bényei, B. Gergely, and Z. Győrffy, “Digital health is a cultural transformation of traditional healthcare,” *mHealth*, vol. 3, pp. 38–38, Sep. 2017, doi: 10.21037/mhealth.2017.08.07.

[73] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, “The role of artificial intelligence in healthcare: a structured literature review,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 125, Dec. 2021, doi: 10.1186/s12911-021-01488-9.

[74] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, “Artificial Intelligence in Healthcare: Review and Prediction Case Studies,” *Engineering*, vol. 6, no. 3, pp. 291–301, Mar. 2020, doi: 10.1016/j.eng.2019.08.015.

[75] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, “Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods,” *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 17, no. 1, p. 102690, Jan. 2023, doi: 10.1016/j.dsx.2022.102690.

[76] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, “Machine Learning in Healthcare: A Review,” in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Mar. 2018, pp. 910–914. doi: 10.1109/ICECA.2018.8474918.

[77] A. Callahan and N. H. Shah, “Machine Learning in Healthcare,” in *Key Advances in Clinical Informatics*, Elsevier, 2017, pp. 279–291. doi: 10.1016/B978-0-12-809523-2.00019-4.

[78] D. S. Char, M. D. Abràmoff, and C. Feudtner, “Identifying Ethical Considerations for Machine Learning Healthcare Applications,” *Am. J. Bioeth.*, vol. 20, no. 11, pp. 7–17, Nov. 2020, doi: 10.1080/15265161.2020.1819469.

[79] M. A. Ahmad, C. Eckert, and A. Teredesai, “Interpretable Machine Learning in Healthcare,” in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Aug. 2018, pp. 559–560. doi: 10.1145/3233547.3233667.

[80] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” *J. Phys. Conf. Ser.*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[81] P. Kaur, M. Sharma, and M. Mittal, “Big Data and Machine Learning Based Secure Healthcare Framework,” *Procedia Comput. Sci.*, vol. 132, pp. 1049–1059, 2018, doi: 10.1016/j.procs.2018.05.020.

[82] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” *Int. J. Intell. Networks*, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.

[83] “A Path for Translation of Machine Learning Products into Healthcare Delivery,” *EMJ Innov.*, Jan. 2020, doi: 10.33590/emjinnov/19-00172.

[84] A. Gupta and R. Katarya, “Social media based surveillance systems for healthcare using machine learning: A systematic review,” *J. Biomed. Inform.*, vol. 108, p. 103500, Aug. 2020, doi: 10.1016/j.jbi.2020.103500.

[85] A. O. Tucker and J. Janick, “An Introduction to the Codex Cruz-Badianus,” in *Flora of the Codex Cruz-Badianus*, Cham: Springer International Publishing, 2020, pp. 1–13. doi: 10.1007/978-3-030-46959-7_1.

[86] I. Y. Chen, S. Joshi, M. Ghassemi, and R. Ranganath, “Probabilistic Machine Learning for Healthcare,” *Annu. Rev. Biomed. Data Sci.*, vol. 4, no. 1, pp. 393–415, Jul. 2021, doi: 10.1146/annurev-biodatasci-092820-033938.

[87] S. Siddique and J. C. L. Chow, “Machine Learning in Healthcare Communication,” *Encyclopedia*, vol. 1, no. 1, pp. 220–239, Feb. 2021, doi: 10.3390/encyclopedia1010021.

[88] J. Waring, C. Lindvall, and R. Umeton, “Automated machine learning: Review of the state-of-the-art and opportunities for healthcare,” *Artif. Intell. Med.*, vol. 104, p. 101822, Apr. 2020, doi: 10.1016/j.artmed.2020.101822.

[89] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and*

Techniques to Build Intelligent Systems. 2019.

[90] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.

[91] I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane, *Big Data and Social Science*. Chapman and Hall/CRC, 2020. doi: 10.1201/9780429324383.

[92] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[93] A. DeMaris, "A Tutorial in Logistic Regression," *J. Marriage Fam.*, vol. 57, no. 4, p. 956, Nov. 1995, doi: 10.2307/353415.

[94] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.

[95] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behav. Ther.*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.

[96] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[97] B. M. G. Brad Boehmke, *Hands-On Machine Learning with R*. 2019.

[98] M. Usama *et al.*, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019, doi: 10.1109/ACCESS.2019.2916648.

[99] A. I. Károly, R. Fullér, and P. Galambos, "Unsupervised Clustering for Deep Learning: A tutorial survey," *Acta Polytech. Hungarica*, vol. 15, no. 8, pp. 29–53, 2018, doi: 10.12700/APH.15.8.2018.8.2.

[100] C. M. Parlett-Pelleriti, E. Stevens, D. Dixon, and E. J. Linstead, "Applications of Unsupervised Machine Learning in Autism Spectrum Disorder Research: a Review," *Rev. J. Autism Dev. Disord.*, vol. 10, no. 3, pp. 406–421, Sep. 2023, doi: 10.1007/s40489-021-00299-y.

[101] A. Adimabua Ojugo and E. Ekurume, "Predictive Intelligent Decision Support Model in Forecasting of the Diabetes Pandemic Using a Reinforcement Deep Learning Approach," *Int. J. Educ. Manag. Eng.*, vol. 11, no. 2, pp. 40–48, Apr. 2021, doi: 10.5815/ijeme.2021.02.05.

[102] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 3, pp. 1464–1468, Jun. 1997, doi: 10.1109/23.589532.

[103] B. M. K. P, S. P. R, N. R K, and A. K, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," *Int. J. Cogn. Comput. Eng.*, vol. 1, pp. 55–61, Jun. 2020, doi: 10.1016/j.ijcce.2020.10.002.

[104] S. Y. Rhee, J. M. Sung, S. Kim, I.-J. Cho, S.-E. Lee, and H.-J. Chang, "Development and Validation of a Deep Learning Based Diabetes Prediction System Using a Nationwide Population-Based Cohort," *Diabetes Metab. J.*, vol. 45, no. 4, pp. 515–525, Jul. 2021, doi: 10.4093/dmj.2020.0081.

[105] D. Yu and L. Deng, *Automatic Speech Recognition*. London: Springer London, 2015. doi: 10.1007/978-1-4471-5779-3.

[106] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[107] D. Ravi *et al.*, "Deep Learning for Health Informatics," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[108] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan, “Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014, doi: 10.1109/LGRS.2014.2309695.

[109] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, “A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues,” *J. Biomed. Inform.*, vol. 113, p. 103627, Jan. 2021, doi: 10.1016/j.jbi.2020.103627.

[110] M. Längkvist, L. Karlsson, and A. Loutfi, “A review of unsupervised feature learning and deep learning for time-series modeling,” *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, Jun. 2014, doi: 10.1016/j.patrec.2014.01.008.

[111] S. Nosratabadi *et al.*, “Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods,” *Mathematics*, vol. 8, no. 10, p. 1799, Oct. 2020, doi: 10.3390/math8101799.

[112] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” *Futur. Healthc. J.*, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: 10.7861/futurehosp.6-2-94.

[113] F. K. Dosilovic, M. Brcic, and N. Hlupic, “Explainable artificial intelligence: A survey,” in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215. doi: 10.23919/MIPRO.2018.8400040.

[114] S. Reddy, J. Fox, and M. P. Purohit, “Artificial intelligence-enabled healthcare delivery,” *J. R. Soc. Med.*, vol. 112, no. 1, pp. 22–28, Jan. 2019, doi: 10.1177/0141076818815510.

[115] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 310, Dec. 2020, doi: 10.1186/s12911-020-01332-6.

[116] S. Castagno and M. Khalifa, “Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study,” *Front. Artif. Intell.*, vol. 3, Oct. 2020, doi: 10.3389/frai.2020.578983.

[117] M.-C. Laï, M. Brian, and M.-F. Mamzer, “Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France,” *J. Transl. Med.*, vol. 18, no. 1, p. 14, Dec. 2020, doi: 10.1186/s12967-019-02204-y.

[118] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[119] F. Bellocchio and H. Zhang, “Editorial: Artificial intelligence in nephrology,” *Front. Nephrol.*, vol. 3, Sep. 2023, doi: 10.3389/fneph.2023.1270769.

[120] Q. Pan and M. Tong, “Artificial intelligence in predicting chronic kidney disease prognosis. A systematic review and meta-analysis,” *Ren. Fail.*, vol. 46, no. 2, Dec. 2024, doi: 10.1080/0886022X.2024.2435483.

[121] F. I. R. Md. Taufiqul Haque Khan Tusar, Md. Touhidul Islam, “Detecting Chronic Kidney Disease(CKD) at the Initial Stage: A Novel Hybrid Feature-selection Method and Robust Data Preparation Pipeline for Different ML Techniques,” in *the International Conference on Computing and Informatics (ICCI)*, 2022. doi: https://doi.org/10.48550/arXiv.2203.01394.

[122] T. Mathieu *et al.*, “Lexical fields of predictive and personalized medicine,” *Ann. Biol. Clin. (Paris).*, vol. 70, no. 6, pp. 651–658, Nov. 2012, doi: 10.1684/abc.2012.0767.

[123] I. Akhmetov and R. V. Bubnov, “Assessing value of innovative molecular diagnostic tests in the concept of predictive, preventive, and personalized medicine,” *EPMA J.*, vol. 6, no. 1, p. 19, Dec. 2015, doi: 10.1186/s13167-015-0041-3.

[124] C. Feng and F. Liu, “Artificial intelligence in renal pathology: Current status and future,” *Bosn. J. Basic Med. Sci.*, Nov. 2022, doi: 10.17305/bjbms.2022.8318.

[125] M. Hueso and A. Vellido, “How artificial intelligence is transforming nephrology,” *BMC Nephrol.*, vol. 25, no. 1, p. 276, Aug. 2024, doi: 10.1186/s12882-024-03724-6.

[126] C.-C. Wu, M. M. Islam, T. N. Poly, and Y.-C. Weng, “Artificial Intelligence in Kidney Disease: A Comprehensive Study and Directions for Future Research,” *Diagnostics*, vol. 14, no. 4, p. 397, Feb. 2024, doi: 10.3390/diagnostics14040397.

[127] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[128] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?,” *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[129] B. N. Lakshmi and G. H. Raghunandhan, “A conceptual overview of data mining,” in *2011 National Conference on Innovations in Emerging Technology*, Feb. 2011, pp. 27–32. doi: 10.1109/NCOIET.2011.5738828.

[130] “Advanced data mining techniques,” *Choice Rev. Online*, vol. 45, no. 12, pp. 45-6838-45–6838, Aug. 2008, doi: 10.5860/CHOICE.45-6838.

[131] V. Kaul, S. Enslin, and S. A. Gross, “History of artificial intelligence in medicine,” *Gastrointest. Endosc.*, vol. 92, no. 4, pp. 807–812, Oct. 2020, doi: 10.1016/j.gie.2020.06.040.

[132] G. E. A. P. A. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, May 2003, doi: 10.1080/713827181.

[133] F. Biessmann, D. Salinas, S. Schelter, P. Schmidt, and D. Lange, “‘Deep’ Learning for Missing Value Imputationin Tables with Non-Numerical Data,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Oct. 2018, pp. 2017–2025. doi: 10.1145/3269206.3272005.

[134] S. Jäger, A. Allhorn, and F. Bießmann, “A Benchmark for Data Imputation Methods,” *Front. Big Data*, vol. 4, Jul. 2021, doi: 10.3389/fdata.2021.693674.

[135] H. Jin, Q. Song, and X. Hu, “Auto-Keras: An Efficient Neural Architecture Search System,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019, pp. 1946–1956. doi: 10.1145/3292500.3330648.

[136] N. Mahesh, J. J. Babu, K. Nithya, and S. A. Arunmozhi, “Water quality prediction using LSTM with combined normalizer for efficient water management,” *Desalin. Water Treat.*, vol. 317, p. 100183, Jan. 2024, doi: 10.1016/j.dwt.2024.100183.

[137] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, “Expert cancer model using supervised algorithms with a LASSO selection approach,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, p. 2631, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2631-2639.

[138] M. M. Hassan *et al.*, “A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction,” *Decis. Anal. J.*, vol. 7, p. 100245, Jun. 2023, doi: 10.1016/j.dajour.2023.100245.

[139] B. Zhang, Y. Zhang, and X. Jiang, “Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm,” *Sci. Rep.*, vol. 12, no. 1, p. 9244, Jun. 2022, doi: 10.1038/s41598-022-13498-2.

[140] M. Y. Shams, Z. Tarek, and A. M. Elshewey, “A novel RFE-GRU model for diabetes classification using PIMA Indian dataset,” *Sci. Rep.*, vol. 15, no. 1, p. 982, Jan. 2025, doi: 10.1038/s41598-024-82420-9.

[141] Q. Zhou, H. Zhou, Q. Zhou, F. Yang, and L. Luo, "Structure damage detection based on random forest recursive feature elimination," *Mech. Syst. Signal Process.*, vol. 46, no. 1, pp. 82–90, May 2014, doi: 10.1016/j.ymssp.2013.12.013.

[142] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[143] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning Proceedings 1992*, Elsevier, 1992, pp. 249–256. doi: 10.1016/B978-1-55860-247-2.50037-1.

[144] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189–203, Sep. 2018, doi: 10.1016/j.jbi.2018.07.014.

[145] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[146] C. Ricciardi *et al.*, "Linear discriminant analysis and principal component analysis to predict coronary artery disease," *Health Informatics J.*, vol. 26, no. 3, pp. 2181–2192, Sep. 2020, doi: 10.1177/1460458219899210.

[147] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/3820360.

[148] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," *J. Stat. Softw.*, vol. 36, no. 11, 2010, doi: 10.18637/jss.v036.i11.

[149] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," *Informatics Med. Unlocked*, vol. 44, p. 101442, 2024, doi: 10.1016/j.imu.2023.101442.

[150] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensembles learning," *BMC Bioinformatics*, vol. 24, no. 1, p. 224, Jun. 2023, doi: 10.1186/s12859-023-05300-5.

[151] P. Chittora *et al.*, "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[152] P. Ghosh, S. Azam, M. Jonkman, S. Shultana, and A. R. Beeravolu, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3053759.

[153] P. Muthulakshmi, M. Parveen, and P. Rajeswari, "Prediction of Heart Disease using Ensembles Learning," *Indian J. Sci. Technol.*, vol. 16, no. 20, pp. 1469–1476, May 2023, doi: 10.17485/IJST/v16i20.2279.

[154] D. C. Yadav and S. Pal, "An Ensembles Approach for Classification and Prediction of Diabetes Mellitus Disease," 2021, pp. 225–235. doi: 10.1007/978-981-16-3915-9_18.

[155] S. Pal, "Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms," *Multimed. Tools Appl.*, Apr. 2023, doi: 10.1007/s11042-023-15188-1.

[156] R. Ani, G. Sasi, U. R. Sankar, and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 1287–1292. doi: 10.1109/ICACCI.2016.7732224.

[157] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney

disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.017.

[158] S. P. Vikas Chaurasia, "Data Mining Approach to Detect Heart Disease," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, pp. 56–66, 2014.

[159] Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Feb. 2019, pp. 1–6. doi: 10.1109/ECACE.2019.8679388.

[160] S. K. Dey, K. M. M. Uddin, H. M. H. Babu, M. M. Rahman, A. Howlader, and K. M. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," *Intell. Syst. with Appl.*, vol. 16, p. 200144, Nov. 2022, doi: 10.1016/j.iswa.2022.200144.

[161] J. Mishra and S. Tarar, "Chronic Disease Prediction Using Deep Learning," 2020, pp. 201–211. doi: 10.1007/978-981-15-6634-9_19.

[162] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Nov. 2020, pp. 1–6. doi: 10.1109/iSAI-NLP51646.2020.9376787.

[163] M. Mustafizur Rahman, M. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomed. Signal Process. Control*, vol. 87, p. 105368, Jan. 2024, doi: 10.1016/j.bspc.2023.105368.

[164] B. Navaneeth and M. Suchetha, "A dynamic pooling based convolutional neural network approach to detect chronic kidney disease," *Biomed. Signal Process. Control*, vol. 62, p. 102068, Sep. 2020, doi: 10.1016/j.bspc.2020.102068.

[165] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.

[166] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models," *Economies*, vol. 10, no. 8, p. 188, Jul. 2022, doi: 10.3390/economies10080188.

[167] A. Nishanth and T. Thiruvaran, "Identifying Important Attributes for Early Detection of Chronic Kidney Disease," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 208–216, 2018, doi: 10.1109/RBME.2017.2787480.

[168] Indira S. Fal Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," *",International J. Adv. Comput. Theory Eng.*, vol. 2, pp. 2319–2526, 2013.

[169] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, no. 1, p. 119, Dec. 2019, doi: 10.1186/s12967-019-1860-0.

[170] M. M. Hassan *et al.*, "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records," *Human-Centric Intell. Syst.*, vol. 3, no. 2, pp. 92–104, Feb. 2023, doi: 10.1007/s44230-023-00017-3.

[171] D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2022, pp. 1305–1310. doi: 10.1109/ICOEI53556.2022.9776876.

[172] D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," *2022 6th Int. Conf. Trends Electron. Informatics, ICOEI 2022 - Proc.*, no. Icoei, pp. 1305–1310, 2022, doi: 10.1109/ICOEI53556.2022.9776876.

[173] S. Y. Yashfi *et al.*, “Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms,” in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, pp. 1–5. doi: 10.1109/ICCCNT49239.2020.9225548.

[174] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, “A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease,” *Bioengineering*, vol. 9, no. 8, p. 350, Jul. 2022, doi: 10.3390/bioengineering9080350.

[175] S. Pal, “Chronic Kidney Disease Prediction Using Machine Learning Techniques,” *Biomed. Mater. Devices*, Aug. 2022, doi: 10.1007/s44174-022-00027-y.

[176] M. F. Hossain, S. T. Diya, and R. Khan, “ACD-ML: Advanced CKD detection using machine learning: A tri-phase ensembles and multi-layered stacking and blending approach,” *Comput. Methods Programs Biomed. Updat.*, vol. 7, p. 100173, 2025, doi: 10.1016/j.cmpbup.2024.100173.

[177] R. K. Halder *et al.*, “ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application,” *J. Pathol. Inform.*, vol. 15, p. 100371, Dec. 2024, doi: 10.1016/j.jpi.2024.100371.

[178] E. Dritsas and M. Trigka, “Machine Learning Techniques for Chronic Kidney Disease Risk Prediction,” *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 98, Sep. 2022, doi: 10.3390/bdcc6030098.

[179] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, “A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease,” *Decis. Anal. J.*, vol. 6, p. 100169, Mar. 2023, doi: 10.1016/j.dajour.2023.100169.

[180] M. S. Arif, A. Mukheimer, and D. Asif, “Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model,” *Big Data Cogn. Comput.*, vol. 7, no. 3, p. 144, Aug. 2023, doi: 10.3390/bdcc7030144.

[181] R. C. Poonia *et al.*, “Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease,” *Healthcare*, vol. 10, no. 2, p. 371, Feb. 2022, doi: 10.3390/healthcare10020371.

[182] H. Ilyas *et al.*, “Chronic kidney disease diagnosis using decision tree algorithms,” *BMC Nephrol.*, vol. 22, no. 1, p. 273, Dec. 2021, doi: 10.1186/s12882-021-02474-z.

[183] “UCI machine learning repository: Early stage of Chronic kidney disease dataset.” https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[184] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006, doi: 10.1016/j.chemolab.2006.01.007.

[185] S. Islam,Md. Ashiqul and Akter, “Risk Factor prediction of Chronic Kidney Disease,” *UCI machine learning repository*, 2023.

[186] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, “Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2019, pp. 679–684. doi: 10.1109/ICCMC.2019.8819654.

[187] K. Hema, K. Meena, and R. Pandian, “Analyze the impact of feature selection techniques in the early prediction of CKD,” *Int. J. Cogn. Comput. Eng.*, vol. 5, pp. 66–77, 2024, doi: 10.1016/j.ijcce.2023.12.002.

[188] P. Chittora *et al.*, “Prediction of Chronic Kidney Disease - A Machine Learning Perspective,” *IEEE Access*, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[189] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha, and S. Hossain, “Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020, pp. 952–957. doi: 10.1109/ICISS49785.2020.9315878.

[190] S. M. Awad Yousif, H. T. Halawani, G. Amoudi, F. M. Osman Birkea, A. M. R. Almunajam, and A. A. Elhag, “Early detection of chronic kidney disease using eurygasters optimization algorithm with ensembles deep learning approach,” *Alexandria Eng. J.*, vol. 100, pp. 220–231, Aug. 2024, doi: 10.1016/j.aej.2024.05.011.

[191] A. Bilal, A. Alzahrani, A. Almuhaimeed, A. H. Khan, Z. Ahmad, and H. Long, “Advanced CKD detection through optimized metaheuristic modeling in healthcare informatics,” *Sci. Rep.*, vol. 14, no. 1, p. 12601, Jun. 2024, doi: 10.1038/s41598-024-63292-5.

[192] S. Pal, “Chronic Kidney Disease Prediction Using Machine Learning Techniques,” *Biomed. Mater. Devices*, vol. 1, no. 1, pp. 534–540, Mar. 2023, doi: 10.1007/s44174-022-00027-y.