REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE TECHNOLOGIE

DEPARTEMENT D'ELECTRONIQUE

N° :……

DOMAINE : SCIENCE ET TECHNOLOGIE

FILIERE : ELECTRONIQUE

OPTION : ELECTRONIQUE

Mémoire présenté pour l'obtention du

Diplôme de Master Académique

Présenté par:

BELAADA Wahid Akram & AFAN Mohamed

Intitulé :

Advanced Machine Learning for Heart Disease Prediction

A Comparative Study of Ensemble and Traditional

Soutenu devant le jury composé de :

Université Mohamed Boudiaf M'sila	**Président**
Université Mohamed Boudiaf M'sila	**Rapporteur**
Université Mohamed Boudiaf M'sila	**Examinateur**

Année universitaire : 2024 / 2025

Abstract

Heart disease, a leading global cause of mortality, demands innovative early detection strategies. This study evaluates machine learning models (DNN, KNN, SVM, XGBoost, Random Forest) for predicting cardiovascular disease using 1,000 patient records with 16 clinical features. After rigorous preprocessing and validation, ensemble methods like XGBoost (100% accuracy) and Random Forest (99%) outperformed traditional models, highlighting their clinical potential. Challenges such as overfitting and interpretability were addressed, emphasizing the need for diverse datasets and explainable AI (XAI). Future integration with wearable technologies and interdisciplinary collaboration could enable proactive, personalized care, transforming cardiovascular health outcomes globally.

Résumé

Les maladies cardiaques, principale cause de mortalité mondiale, nécessitent des stratégies innovantes de détection précoce. Cette étude évalue des modèles d'apprentissage automatique (DNN, KNN, SVM, XGBoost, Random Forest) pour prédire les maladies cardiovasculaires à l'aide de 1 000 dossiers patients. Les méthodes d'ensemble, notamment XGBoost (exactitude 100%) et Random Forest (99%), ont surpassé les modèles traditionnels. Les défis comme le surapprentissage et l'interprétabilité ont été analysés, soulignant l'importance de jeux de données diversifiés et d'IA explicable (XAI). Une intégration future avec des technologies portables et une collaboration interdisciplinaire pourraient permettre des soins personnalisés et proactifs, transformant la santé cardiovasculaire.

الملخص

أمراض القلب، السبب الرئيسي للوفيات عالميًّا، تستدعي استراتيجيات مبتكرة للكشف المبكر. تقيّم هذه الدراسة نماذج التعلّم الآليDNN) ، KNN، SVM، XGBoost، Random Forest) للتنبؤ بأمراض القلب باستخدام بيانات 1000 مريض. تفوّقت نماذج التجميع، خاصة XGBoost (بدقة 100%) و(99%) Random Forest، على النماذج التقليدية. جرى معالجة تحديات مثل التكيّف الزائد وضمان تفسير النتائج، مع التركيز على أهمية تنويع البيانات وتقنيات الذكاء الاصطناعي القابل للتفسير .(XAI) قد تتيح التكامل مع الأجهزة القابلة للارتداء والتعاون بين التخصصات رعاية استباقية وشخصية، مما يُحسّن النتائج الصحية القلبية عالميًّا.

Dedication

To our beloved families,
For their endless sacrifices, unconditional love, steadfast support, and prayers that guided us throughout this journey.

To our siblings,
For standing by us in our academic and personal lives, and for the encouragement that fueled our perseverance.

May this work stand as a testament to the unwavering support and faith of those who believed in us.

And to our dear friends,
Thank you for being our pillars during both challenges and triumphs.

Acknowledgments

First and foremost, we extend our deepest gratitude to Almighty God, who blessed us with the strength and guidance to overcome challenges and complete this work.

We express our sincere appreciation to our supervisor, **Dr. TABAKH Moustapha**, for her unwavering patience, dedication, and invaluable advice, which enriched our research and refined our ideas.

We also extend our thanks to the esteemed jury members for their interest in our project, their thorough review of our work, and their constructive feedback that greatly enhanced its quality.

We are profoundly grateful to our families, the beacons of our moral and emotional support, who stood by us at every step and instilled in us the confidence and perseverance to succeed. We also acknowledge our friends, whose constant encouragement and assistance helped us navigate difficulties.

With your collective support, we were able to establish a solid foundation for this research and develop a clear methodology that guided us toward achieving our goals. To all of you, we owe our deepest respect and gratitude.

Table of Contents

Abstract
Dedication
Acknowledgments
Table of Contents
List of Figures
List of Tables
INTRODUCTION GENERAL 1
CHAPTER ONE : MACHINE LEARNING FOR HEART DISEASE PREDICTION 4
I.1 Introduction 5
I.2 Understanding Heart Disease 5
I.2.1 Types of Cardiovascular Diseases 5
I.2.2 Causes of Cardiovascular Diseases 6
I.2.3 Symptoms of Cardiovascular Diseases 6
I.3 Machine Learning for Medical Prediction 7
I.3.1 Overview of Supervised Learning 7
I.3.2 Key Supervised Learning Algorithms for Medical Diagnosis 7
I.3.3 Feature Engineering in Medical Data 7
I.3.4 Techniques in Feature Engineering 8
I.3.5 Impact of Feature Engineering on Prediction Accuracy 8
I.4 Dataset Description 8
I.4.1 Data Source 8
I.4.2 Features and Descriptions 9
I.4.3 Data Preprocessing and Cleaning 10
I.5 Overview of Selected Machine Learning Algorithms 10
I.5.1 Deep Neural Network (DNN) 10
I.5.2 K-Nearest Neighbors (KNN) 13

I.5.3 Support Vector Machine (SVM) 15
I.5.4 XGBoost 17
I.5.5 Random Forest 21
I.6 Performance Metrics for Evaluation 22
I.6.1 Confusion Matrix and Classification Outcomes 22
I.6.2 Explanation of Key Performance Metrics 23
I.7 Conclusion 24
CHAPTER TWO: 26
IMPLEMENTATION AND PERFORMANCE EVALUATION 26
II.1 Introduction 27
II.2 Data Preprocessing and Feature Engineering 27
II.2.1 Handling Missing Data 27
II.2.2 Normalization and Standardization 27
II.2.3 Feature Selection 28
II.3 Model Implementation and Training 28
II.3.1 Training Process for Each Algorithm 28
II.3.2 Hyperparameter Optimization 29
II.4 Performance Comparison of Models 30
II.4.1 Performance Metrics Used for Evaluation 30
II.4.2 Approach to Model Comparison 30
II.4.3 Importance of Interpretability in Healthcare 31
II.5 Comparative Analysis of Algorithm Strengths and Weaknesses 31
II.6 Conclusion 33
CHAPTER THREE 35
RESULTS, DISCUSSION, AND FUTURE PROSPECTS 35
III.1 Introduction 36
III.2 State of the Art 36
III.3 Experimental Results and Model Performance 37
III.3.1 Deep Neural Network (DNN) 37
III.3.2 K-Nearest Neighbors (KNN) 39
III.3.3 Support Vector Machine (SVM) 42
III.3.5 XGBoost Algorithm 45
III.3.5 Random Forest Classifier 49

III.4 Final Comparative Analysis....50
III.5 Challenges and Limitations....52
III.5.1 Data Quality and Imbalance Issues....52
III.5.2 Computational Constraints....52
III.5.3 Ethical Considerations in AI-Driven Diagnostics....53
III.6 Future Research Directions....53
III.6.1 Advancing Model Performance with Larger and More Diverse Data....54
III.6.2 Explainable AI for Medical Applications....54
III.6.3 Bridging the Gap Between Research and Clinical Implementation....54
III.6.4 Addressing Ethical and Bias Concerns....55
III.6.5 Leveraging Real-Time Monitoring and Wearable Technologies....55
GENERAL CONCLUSION....56
General Conclusion....57
Bibliography....60

List of Figures

Figure I.1. Deep Neural Network (DNN) architecture. 12
Figure I.2. Using distance in the KNN algorithm 14
Figure I.3. Support Vector Machine (SVM) algorithm 16
Figure I.4. Schematic illustration of the XGboost model. 20
Figure I.5. Random Forest Algorithm in Machine Learning 21
Figure III.1 Confusion Matrix for DNN. 38
Figure III.2 Confusion Matrix for KNN. 40
Figure III.3 KNN Accuracy vs. K-Value 41
Figure III.4 Precision-Recall Curve for KNN 42
Figure III.5 Confusion Matrix for SVM. 43
Figure III.6 SVM Decision Boundary 44
Figure III.7 Precision-Recall Curve for SVM 45
Figure III.8 Confusion Matrix for XGBoost. 46
Figure III.9 XGBoost Training & Validation Loss 47
Figure III.10 Feature Importance in XGBoost 48
Figure III.11 Confusion Matrix for Random Forest. 49
Figure III.11 Random Forest Feature Importance 50

List of Tables

Table I.1 Several hyperparameters used in the XGBoost method. 19
Table I.2 Matrice de confusion 23
Table III.1 Related work on heart disease prediction 37
Table III.2 Classification Report for DNN 38
Table III.3 Classification Report for KNN 40
Table III.4 Classification Report for SVM. 43
Table III.5 Classification Report for XGBoost. 46
Table III.6 Classification Report for Random Forest 50
Table III.7 Comparison of Five Models 51

INTRODUCTION GENERAL

Introduction General

Heart disease remains one of the leading causes of mortality worldwide, posing a significant challenge to public health. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) account for approximately 17.9 million deaths each year, making them the primary cause of global mortality. The early detection of heart disease is crucial in reducing fatal outcomes, as timely medical intervention can significantly improve a patient's prognosis. Traditional diagnostic methods, such as electrocardiograms (ECG), stress tests, and cholesterol screenings, provide valuable insights but often require extensive medical expertise and can be time-consuming and costly. To address these limitations, researchers have explored the potential of machine learning as a powerful tool for automated heart disease prediction.

Machine learning, a branch of artificial intelligence (AI), has revolutionized various fields, including healthcare, by enabling predictive analytics based on large datasets. By leveraging computational models, machine learning can identify hidden patterns in medical data, improving the accuracy of disease diagnosis and risk assessment. Recent advancements in supervised learning techniques, including **Deep Neural Networks (DNNs), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), XGBoost, and Random Forest**, have demonstrated significant potential in medical applications. These models analyze patient attributes such as age, cholesterol levels, blood pressure, and lifestyle factors to determine the likelihood of heart disease with high accuracy.

The primary objective of this study is to develop a machine learning-based predictive model for heart disease diagnosis. This research aims to evaluate multiple machine learning algorithms, compare their performance, and determine the most effective approach for heart disease classification. The study uses a publicly available dataset containing **1,000 patient records** and 16 clinical and demographic features, ensuring a robust evaluation framework. Data preprocessing techniques, such as feature selection, normalization, and missing data handling, are applied to enhance model performance.

This thesis is structured into three main chapters. **Chapter 1** provides a comprehensive overview of heart disease, including its causes, symptoms, and traditional diagnostic methods. It also introduces machine learning principles and discusses the algorithms selected for this study. **Chapter 2** focuses on the implementation phase, detailing data preprocessing, model training, and hyperparameter optimization techniques. It also presents the performance metrics used to assess the effectiveness of each algorithm. **Chapter 3** presents the experimental results, discussing model performance, comparative analysis, and practical

implications of this research. The study concludes with a discussion on future directions and potential improvements in machine learning-driven heart disease prediction.

By integrating machine learning techniques into medical diagnostics, this research contributes to the growing field of AI-driven healthcare solutions. The findings of this study may assist healthcare professionals in making more accurate and timely diagnoses, ultimately improving patient outcomes and reducing the burden of cardiovascular diseases on global health systems.

GENERAL CONCLUSION

General Conclusion

Heart disease remains a major global health concern, responsible for a significant percentage of morbidity and mortality worldwide. Early detection and accurate prediction of cardiovascular diseases are critical in reducing fatality rates and improving patient outcomes. Machine learning has emerged as a transformative tool in this field, offering promising results in predictive analytics, risk assessment, and early diagnosis.

This thesis has explored the potential of machine learning algorithms in predicting heart disease, focusing on five key models: Deep Neural Networks (DNN), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), XGBoost, and Random Forest. Through rigorous implementation, training, and evaluation, we have assessed their performance using standard metrics such as accuracy, precision, recall, and F1-score. The results demonstrate that machine learning models, particularly ensemble methods and deep learning techniques, can achieve high predictive accuracy, supporting their applicability in real-world clinical settings.

Despite these advancements, several challenges persist. The study has highlighted issues related to data quality, feature selection, computational efficiency, and model interpretability. A key limitation remains the availability of high-quality, diverse datasets that adequately represent different populations. Additionally, the black-box nature of deep learning models poses interpretability challenges, which may hinder their adoption in clinical practice. Ethical concerns, including data privacy, bias in model predictions, and regulatory compliance, must also be addressed to ensure fair and responsible AI deployment in healthcare.

Moving forward, several research directions can be pursued to enhance the robustness and clinical relevance of machine learning models in heart disease prediction. First, integrating larger and more diverse datasets can improve generalizability and reduce bias in model predictions. The application of explainable AI (XAI) techniques, such as SHAP values and Local Interpretable Model-Agnostic Explanations (LIME), can help bridge the gap between AI predictions and clinician trust. Additionally, the integration of real-time monitoring systems using wearable technologies can provide continuous patient data, allowing for proactive intervention and improved risk assessment. Finally, interdisciplinary collaborations between data scientists, healthcare professionals, and policymakers are essential in ensuring the successful implementation of AI-driven diagnostic tools in clinical environments.

Finally, interdisciplinary collaborations between data scientists, healthcare professionals, and policymakers are essential in ensuring the successful implementation of AI-driven diagnostic tools in clinical environments.

This thesis bridges the gap between machine learning innovation and clinical cardiology by delivering models that combine diagnostic precision with interpretability. By achieving near-perfect accuracy while addressing ethical and computational challenges, our work empowers healthcare providers to adopt AI-driven tools confidently. Future integration with wearable technologies and real-time data pipelines promises to revolutionize preventive cardiology, transforming cardiovascular care from reactive treatment to proactive, personalized medicine—ultimately saving lives on a global scale.

BIBLIOGRAPHY

Bibliography

[1] Animesh Hazra, S. K. M. G. M. M. "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review." Advances in Computational Sciences and Technology, 2137-2159, 2017.

[2] World Health Organization. "World Health Organization," February 9, 2022. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[3] Khanji, C. Évaluation de la qualité des soins et des services préventifs cardiovasculaires en première ligne. Montréal: Université de Montréal, 2018.

[4] Yadav, S. P. Dhyan Chandra. "Prediction of Heart Disease Using Feature Selection and Random Forest Ensemble Method." International Journal for Pharmaceutical Research Scholars, 56-66, 2020.

[5] Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer, 2009.

[6] Gupta, C. "Cardiac Disease Prediction using Supervised Machine Learning Techniques." Journal of Physics: Conference Series, 2022.

[7] Krishnan, J., and Gretha, S. "Prediction of Heart Disease Using Machine Learning Algorithms." 1st International Conference of Innovations in Information and Communication Technology (ICIIcT), 2019.

[8] Mammadov, R. "Heart disease prediction dataset." Kaggle, 2023. Retrieved from https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction.

[9] LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning." Nature, 521(7553), 436-444, 2015.

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. "Dropout: A simple way to prevent neural networks from overfitting." Journal of Machine Learning Research, 15(1), 1929-1958, 2014.

[11] Xia, Y., Wulan, N., Wang, K., and Zhang, H. "Detecting atrial fibrillation by deep convolutional neural networks." Computers in Biology and Medicine, 122, 103883, 2020.

[12] Imandoust, M. B. Sadegh Bafandeh. "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background." International Journal of Engineering Research and Applications, 605-610, 2013.

[13] Pouriyeh, S. V. G. S. G. D. P. H. A. J. G. "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease." 22nd IEEE Symposium on Computers and Communication (ISCC), 2017.

[14] Wan, Z., Dong, Y., Yu, Z., Lv, H., and Lv, Z. "Semi-supervised support vector machine for digital twins-based brain image fusion." Frontiers in Neuroscience, 15, 705323, 2021.

[15] Budholiya, K., Shrivastava, S. K., and Sharma, V. "An optimized XGBoost-based diagnostic system for effective prediction of heart disease." Journal of King Saud University-Computer and Information Sciences, 34(7), 4514-4523, 2022.

[16] Pal, M., and Parija, S. "Prediction of heart diseases using random forest." Journal of Physics: Conference Series, Vol. 1817, No. 1, p. 012009, 2021.

[17] Dalianis, Hercules. "Evaluation metrics and evaluation." Clinical Text Mining: Secondary Use of Electronic Patient Records, 45–53, 2018.

[18] Narin, A., Isler, Y., and Ozer, M. "Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability." Proceedings of the 2016

Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, October 27–29, 2016.
[19] Shah, D., Patel, S., and Bharti, S. K. "Heart Disease Prediction using Machine Learning Techniques." SN Computer Science, 1, 345, 2020.
[20] Drożdż, K., Nabrdalik, K., Kwiendacz, H., Hendel, M., Olejarz, A., Tomasik, A., Bartman, W., Nalepa, J., Gumprecht, J., and Lip, G. Y. H. "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach." Cardiovascular Diabetology, 21, 240, 2022.
[21] Alotaibi, F. S. "Implementation of Machine Learning Model to Predict Heart Failure Disease." International Journal of Advanced Computer Science and Applications, 10, 261–268, 2019.
[22] Hasan, N., and Bao, Y. "Comparing different feature selection algorithms for cardiovascular disease prediction." Health Technology, 11, 49–62, 2020.
[23] Muhammad, G., et al. "Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach." IEEE Access, vol. 11, 97879-97895, 2023.
[24] Bhatt, C. M., Patel, P., Ghetia, T., and Mazzeo, P. L. "Effective Heart Disease Prediction Using Machine Learning Techniques." Algorithms, 16, 88, 2023.