EARLY DETECTION OF CARDIOVASCULAR ABNORMALITIES USING ECG DATA

BY

SHAHRIAR ALAM SALMAN

ID: 221-15-4642

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr.Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr.Arif Mahmud
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2025

APPROVAL

This Project titled **"Early Detection of Cardiovascular Adnormalities using ECG Data"**, submitted by Name: **Shahriar Alam Salman,** ID No: **221-15-4642** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14 May, 2025**.

BOARD OF EXAMINERS

Dr. Naznin Sultana — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sazzadur Ahamed — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

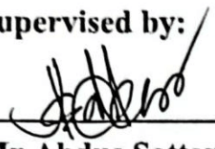
Raja Tariqul Hasan Tusher — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Shah Md Imran — **External Examiner**
Project Manager, External Member
Smart Leadership Academy and EDGE Project
Bangladesh Computer Council, ICT Division

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr.Abdus Sattar, Associate Professor and Dr.Arif Mahmud, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Mr.Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr.Arif Mahmud
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Shahriar Alam Salman

Shahriar Alam Salman
ID: 221-15-4642
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, we express our deepest gratitude to the Almighty for His divine blessings, guidance and strength that enabled us to successfully complete this final year project/internship. Without His support, this endeavor would not have been possible.

We are profoundly indebted to **Mr. Abdus Sattar, Associate Professor and Coordinator of M.Sc**, Department of CSE, Daffodil International University, for his invaluable supervision and mentorship throughout this project. His profound knowledge in the field of machine learning, combined with his keen interest and dedication, has been instrumental in guiding us through every step of this research. His constructive criticism, insightful feedback and unwavering encouragement have been pivotal in shaping this project into its current form.

We also extend our heartfelt thanks to **Dr. Arif Mahmud, Assistant Professor, Department of CSE**, Daffodil International University, for his co-supervision and expert advice during the course of this project. His thoughtful suggestions, meticulous attention to detail and constant motivation have greatly enriched the quality of our work.

Our sincere appreciation goes to **Professor Dr. Touhid Bhuiyan**, **Professor and Head** of the Department of CSE, for his kind support and leadership within the department. We extend our heartfelt gratitude to the faculty and staff of Daffodil International University for their support, as well as our peers for their insightful contributions. We also deeply appreciate the unwavering encouragement and sacrifices of our families, whose patience and inspiration enabled us to achieve our goals. Thank you all.

ABSTRACT

This study examines the utilization of machine learning algorithms for the early identification of cardiovascular illnesses through electrocardiogram (ECG) data. The study assesses various models, including Support Vector Machines (SVM), Logistic Regression (LR), NB, Linear Regression (LinReg), and ensemble methods such as XGBoost, RF, Gradient Boosting (GB, and KNN. The dataset obtained from Kaggle contains ECG measurements and binary labels denoting normal or pathological cardiac function. The performance of each algorithm is evaluated according to its proficiency in accurately classifying ECG patterns and differentiating between normal and pathological cardiac activity. The findings indicate that SVM attains the best accuracy of 0.994, illustrating its proficiency in identifying intricate, non-linear correlations within high-dimensional ECG data. LR and RF closely follow with accuracies of 0.993, showcasing their robustness in modeling linear and probabilistic trends. LinReg performs admirably with an accuracy of 0.992, while XGBoost and KNN achieve comparable scores of 0.9904, highlighting their versatility and noise tolerance. GB and NB report slightly lower accuracies of 0.9888 and 0.9688, respectively, yet remain valuable due to their unique strengths in handling diverse data distributions and probabilistic classification. Ensemble techniques like XGBoost, RF and GB leverage the power of multiple weak learners to deliver strong predictive performance. Meanwhile, KNN's adaptability to varying data patterns underscores its utility in practical applications. These findings highlight the capability of machine learning algorithms to automate ECG interpretation and aid healthcare professionals in prompt diagnosis. Future work should focus on optimizing these models further and validating their performance in real-world clinical settings. By enhancing diagnostic accuracy and efficiency, this research contributes to advancing cardiovascular health monitoring and improving patient outcomes.

Keywords: Electrocardiogram (ECG), Cardiovascular Disease, Ensemble Techniques, Early Detection, Classification Accuracy, Health Monitoring, Clinical Practice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

CVD is a prevalent global health concern that substantially affects mortality rates worldwide. Cardiovascular illness, encompassing myocardial infarctions, cerebrovascular accidents, and hypertension, poses a significant threat to public health.

The insidious character of this phenomenon stems from the many risk factors that contribute to its emergence. Therefore, it is crucial to thoroughly investigate and understand these features in order to implement effective preventive measures and prompt interventions. The fundamental risk factors associated with cardiovascular disease (CVD) are varied and linked. Smoking, a well-recognized factor in cardiovascular disease (CVD), causes harm to the cardiovascular system by inhaling harmful chemicals, which leads to inflammation and stiffness of the arteries. Hypertension, often known as high blood pressure, applies continuous pressure on the arteries, gradually weakening them and increasing the risk of heart-related disorders. Diabetes provides a metabolic component, upsetting the delicate balance of blood sugar levels and encouraging circumstances prone to cardiovascular disease. Obesity, often associated with inactive lifestyles and unhealthy eating habits, increases the risk by promoting inflammation and metabolic disruptions. Moreover, the presence of a hereditary inclination towards cardiovascular disease emphasizes the significance of genetic variables in the development of the illness. Amid this intricate interaction of risk variables, the advent of ML is seen as a disruptive force in healthcare. ML approaches, using their power to filter through enormous datasets and discover nuanced patterns, provide a unique route for predicting CVD risk. By exploiting patient data and analyzing

electronic health records, ML systems may reveal subtle connections and risk markers that may defy standard analytical methodologies. The potential repercussions of such predictive powers are substantial, marking a paradigm change towards individualized and proactive treatment. Within the realm of cardiovascular disease (CVD), the use of machine learning (ML) offers potential for diagnosing those with an increased risk at an early stage. By merging different data sources, including genetic information, lifestyle decisions and clinical assessments, ML models may generate complex risk profiles. These profiles not only simplify the identification of high-risk populations but also open the path for targeted treatments. Personalized healthcare solutions, guided by ML predictions, may vary from targeted lifestyle adjustments and pharmaceutical regimes to carefully watched monitoring for early indicators of illness development. Therefore, the convergence of cardiovascular risk factors needs a comprehensive and complex approach to illness prediction and treatment. Machine Learning, with its analytical skill and aptitude for pattern detection, emerges as a formidable tool in this attempt. The integration of ML algorithms into the field of cardiovascular health promises to usher in an age of precision medicine, where therapies are not only timely but also perfectly suited to the personalized requirements of patients, therefore generating a more effective and proactive healthcare environment.

1.2 Motivation

Cardiovascular Disease (CVD) remains one of the leading causes of mortality worldwide, imposing a significant burden on individuals, healthcare systems and economies. Despite advancements in medical science and technology, the complexity and multifaceted nature of CVD present persistent challenges in early detection, risk prediction and effective intervention. Traditional diagnostic approaches often rely on clinical markers and standardized procedures that may fail to capture the intricate interplay of genetic, lifestyle and environmental factors

contributing to CVD risk. This limitation underscores the urgent need for innovative solutions that can enhance the accuracy, generalizability and personalization of cardiovascular risk assessments.

The advent of machine learning (ML) has opened new frontiers in healthcare by enabling the analysis of vast and complex datasets to uncover hidden patterns and insights. ML-based approaches hold immense potential to revolutionize CVD risk prediction by integrating diverse data sources, such as electronic health records, medical imaging and genetic information, into unified models. By leveraging these multimodal datasets, ML algorithms can identify subtle yet critical risk factors that traditional methods might overlook, thereby improving early detection and enabling timely interventions. Furthermore, enhancing the explainability and interpretability of ML models is essential to foster trust among healthcare professionals and facilitate their seamless integration into clinical workflows. This study is motivated by the need to address existing gaps in ECG-based CVD detection and risk prediction. By developing robust ML models that prioritize accuracy, transparency and ethical considerations, this research aims to empower healthcare providers with tools for personalized and proactive patient care. Ultimately, the motivation lies in reducing the global burden of CVD through innovative, data-driven solutions that improve patient outcomes and transform cardiovascular healthcare.

1.3 Rationale of the Study

Cardiovascular Disease (CVD) remains a leading cause of morbidity and mortality worldwide, posing significant challenges to healthcare systems and economies. Despite advancements in medical science and technology, the multifaceted nature of CVD, characterized by its diverse risk factors and asymptomatic progression, continues to hinder early detection and effective intervention. Traditional

diagnostic methods often rely on clinical markers and standardized assessments, which may fail to capture the intricate interplay of genetic predispositions, lifestyle choices and environmental influences that contribute to an individual's risk profile. This limitation underscores the urgent need for innovative approaches that can enhance the precision and personalization of cardiovascular risk assessments.

The advent of machine learning (ML) has opened new frontiers in healthcare by enabling the analysis of vast and complex datasets to uncover hidden patterns and insights. ML-based approaches offer a transformative opportunity to address the limitations of conventional methods by integrating multimodal data sources—such as electronic health records, medical imaging, genetic information and lifestyle metrics—into unified predictive models. These models can identify subtle yet critical risk factors that traditional techniques might overlook, thereby improving early detection and enabling timely interventions. Furthermore, the ability of ML algorithms to process large-scale, heterogeneous datasets aligns with the growing emphasis on precision medicine, where therapies are tailored to the unique needs of each patient.

This study is motivated by the need to bridge existing gaps in ECG-based CVD detection and risk prediction. By developing robust ML models that prioritize accuracy, transparency and ethical considerations, this research aims to empower healthcare providers with tools for personalized and proactive patient care. Additionally, the study explores the potential of multimodal data fusion to create comprehensive risk profiles, paving the way for targeted interventions and improved patient outcomes.

Ultimately, the rationale for this study lies in its potential to reduce the global burden of CVD through innovative, data-driven solutions. By leveraging the analytical power of ML, this research seeks to usher in an era of precision

cardiology, where early detection, personalized treatment plans and proactive monitoring become standard practices. Such advancements promise to transform cardiovascular healthcare, making it more effective, efficient and patient-centered.

1.4 Expected Output

The planned output of this work involves the construction and validation of machine learning (ML) models optimized for early diagnosis and risk prediction of cardiovascular disease (CVD) using ECG data and multimodal datasets. The primary aim is a strong prediction model capable of identifying persons at high risk of CVD with enhanced accuracy, sensitivity and specificity compared to standard diagnostic approaches. By integrating multiple data sources—such as medical imaging, patient demographics, clinical records and genetic information—the model is projected to provide comprehensive risk profiles, enabling individualized healthcare solutions. Additionally, the project intends to enhance the interpretability and transparency of ML models, facilitating their seamless application in clinical processes. Rigorous clinical validation will demonstrate the therapeutic relevance and practical applicability of these instruments. Ultimately, the projected result includes a transformative approach to CVD management, where timely interventions and individualized therapies lower morbidity and mortality rates, fostering a proactive and efficient healthcare system.

1.5 Report Layout

This paper is split into numerous chapters, each addressing particular areas of the research. Chapter 2 presents a comprehensive assessment of existing literature relating to cardiovascular disease detection. Chapter 3 explains the approach followed in the design and development of the system. Subsequent chapters give the outcomes of the system evaluation, analyze implications for healthcare and offer conclusions and recommendations.

CHAPTER 2

Background Study

2.1 Terminologies

Several key terminologies are essential for understanding the background and scope of cardiovascular disease (CVD) detection and machine learning (ML) applications. Cardiovascular Disease (CVD) refers to a group of disorders affecting the heart and blood vessels, including myocardial infarctions (heart attacks), cerebrovascular accidents (strokes) and hypertension (high blood pressure). Electrocardiogram (ECG) is a diagnostic tool that records the electrical activity of the heart, providing critical insights into cardiac health [3].

Machine Learning (ML) is a branch of artificial intelligence (AI) that enables systems to learn patterns from data and make predictions or decisions without explicit programming. Multimodal Data refers to datasets combining various types of information, such as clinical records, genetic data and medical imaging, to enhance predictive accuracy. Explainability and Interpretability in ML describe the ability of models to provide transparent and understandable outputs, fostering trust among healthcare professionals [6].

Risk Factors are characteristics or conditions that increase the likelihood of developing CVD, including smoking, hypertension, diabetes, obesity and genetic predisposition. Precision Medicine tailors medical treatment to individual patient profiles, leveraging advanced analytics and personalized risk assessments. Generalization refers to an ML model's ability to perform accurately on unseen data beyond its training set. Finally, Clinical Validation ensures that ML models meet therapeutic relevance and practical usability standards in real-world healthcare settings. Understanding these terms is foundational to comprehending the integration of ML in CVD prediction and management [6].

2.2 Related Work

like AlexNet and GoogLeNet are trained on different non-medical photographs, the recommended basic CNN model is adjusted to the distinctive qualities of medical images, notably those connected with heart problems. The study highlights the probable constraints of transfer learning when deployed to medical imaging tasks and underlines the need of developing specialized pre-trained networks based on big medical image datasets. This fits with the future route stated in the research, proposing an aim to explore the building of pre-trained networks specifically designed for medical imaging [1].

Cardiovascular diseases (CVD) are a severe global health concern, contributing greatly to morbidity and mortality. In the context of Bangladesh, a South Asian country with a burgeoning population, understanding the prevalence of CVD among adults is crucial for public health interventions [2]. The systematic review utilized a thorough search method, diving into electronic databases such as MEDLINE, Embase and PubMed, encompassing material till April 13, 2017. The inclusion criteria relied on studies in English that demonstrated the prevalence of CVD in the Bangladeshi adult population. The search strategy was developed on three primary components: CVD, prevalence and Bangladesh. A comprehensive technique was employed by further scanning the reference lists of relevant papers to find fresh research. The review's inclusion approach comprises a two-step screening by two independent reviewers. Titles and abstracts were initially assessed, followed by a full-text examination of selected abstracts. The inclusion criteria included original research reporting CVD prevalence in the general adult population of Bangladesh. Notably, no limits were made on study type, geographical area, or age ranges [2]. A total of 13 studies were included in the final evaluation, representing different demographics, including rural, urban and suburban settings. The findings indicated a wide-ranging prevalence of CVD across the studied studies, with reported rates varying from 0.062% to 77.7%. The pooled prevalence, combining all forms of CVD regardless of gender and geographical

location, was determined at 5.0% with a 95% confidence range of 3%–9%. Despite this attempt to combine prevalence data, a large level of heterogeneity (I2=99.78%) was observed, suggesting variance in estimates across research. A subgroup analysis based on the gender of study participants generated interesting conclusions. The prevalence of CVD in females and men was reported as 3%, with 95% confidence intervals of 1%–7% and 1%–6%, respectively. However, the study indicated great diversity in incidence between studies concentrating solely on females or men. This result underscores the need for greater investigation into suspected gender-specific risk factors contributing to CVD prevalence. To check the integrity of the included studies, a systematic data extraction form was employed, acquiring key information such as author details, publication year, participant demographics, study methodology and CVD prevalence. Additionally, the quality of each study was rated using the Joanna Briggs Institute's approach, encompassing criteria like sample representativeness, recruitment strategies, sample size adequacy and robustness of data analysis. While the systematic review mostly focuses on epidemiological factors of CVD prevalence, the integration of machine learning models could considerably increase projected accuracy [2].

Diverging from the realm of deep learning, the research takes a critical look at conventional machine learning classifiers to uncover their strength in the difficult domain of photo classification. The study casts its investigative gaze upon a repertoire of classical classifiers, including Bayes, KNN, BPNN (Backpropagation Neural Network), PNN (Probabilistic Neural Network) and MPNN (Multilayer Perceptron Neural Network), subjecting each to meticulous scrutiny to discern their individual capacities [3]. Among the traditional classifiers studied, BPNN, PNN and MPNN emerge as formidable contenders, demonstrating an astounding 96% accuracy in image classification tests. This outstanding result indicates the continuous relevance and effectiveness of these conventional methodologies, particularly in the face of the widespread dominance of deep learning models in

recent years. The potential of BPNN, PNN and MPNN to obtain such high accuracy levels attests to their natural aptitude to discern patterns and make intelligent judgments based on the attributes extracted from photos. These discoveries have far-reaching repercussions, particularly in cases where computational resources pose restrictions. Deep learning models, differentiated by their deep neural architectures, usually demand significant processing resources for training and inference. In contrast, conventional machine learning classifiers, as proved in this work, display resilience and efficiency that make them feasible alternatives in resource-constrained scenarios [3].

The article gives insight on the pragmatic concerns that practitioners and researchers must examine when picking between deep learning and standard machine learning techniques for picture classification tasks.

Moreover, the work shows the versatility of these classical classifiers in processing various image datasets. The robust performance of BPNN, PNN and MPNN means that these models can adapt to changing image attributes, making them excellent tools in different disciplines. This adaptability is particularly advantageous in real-world applications where photographs may exhibit diverse qualities, textures and patterns [3].

In the area of image classification, the choice of a convolutional neural network (CNN) architecture plays a vital role in defining the model's performance. The first publication delves into a comprehensive review of three recognized CNN architectures AlexNet, VGG16 and VGG19 shedding light on their performance and effects on image classification tasks. VGG19, standing out as the top-performing model in our analysis, substantiates the assumption that deeper architectures may considerably enhance classification accuracy. With a remarkable recall of 99.38%, accuracy of 99.23% and an F-score of 99.30%, VGG19 outperforms its predecessors, exhibiting its skill in identifying difficult patterns within photos [4]. The accuracy metrics underscore VGG19's power to properly

identify positive instances, while the recall statistic emphasizes its potential to collect a considerable proportion of relevant situations. The strong F-score, which combines accuracy and recall, attests to the model's overall brilliance in establishing a balance between false positives and false negatives. The outcomes highlight the vital significance of model design in attaining heightened accuracy in picture categorization. VGG19's domination over both AlexNet and VGG16 serves as a compelling testament to the effectiveness of depth in neural network topologies. The multiple layers of VGG19 allow it to extract hierarchical features and representations, enabling a sophisticated interpretation of visual input. This deep understanding is particularly valuable in circumstances when photographs may have sophisticated patterns or tiny deviations. Moreover, the results of this work have greater significance for practitioners and researchers in the domain of computer vision. They stress the significance of considering not only the depth but also the architecture's capacity to capture and convey complex features in images successfully. As the quest for more complex image categorization models continues, this study spurs investigation into even deeper and more complicated structures, pushing the boundaries of what is possible in artificial intelligence. In summary, the performance of VGG19 in this study functions as a light pointing future endeavors towards more robust and effective CNN architectures for image classification applications [4].

Central to the study's results is the outstanding achievement of a classification accuracy of 90%, paired with a specificity of 92% and a sensitivity approaching 81%. These metrics demonstrate the efficacy of CNNs in spotting subtle patterns within ultrasound imagery connected to fatty liver disorders. The study contributes greatly to the burgeoning body of research on medical image classification, underlining the potential of deep learning algorithms to increase diagnostic accuracy [5]. A significant result of the research lays in the beneficial impact of architectural modifications on categorization accuracy. By implementing structural

modifications such as additional filters and skip connections, the study reveals the possibilities for upgrading CNN architectures to better react to the complexity of fatty liver ultrasound imaging. This information is valuable for practitioners and researchers alike, presenting a way for strengthening neural network topologies in medical image processing. Furthermore, the invention of batch normalization (BN) emerges as a notable strategy credited with quickening convergence. The study reveals the influence of BN in limiting internal covariate movement during the training process, leading to more stable and faster convergence of the model. This not only enhances the effectiveness of the CNN but also contributes to a more streamlined and speedier diagnosis method in a clinical situation. In extrapolating from this discovery, the bigger ramifications for medical image classification become obvious. The study underlines the versatility of CNNs to diverse medical domains and illustrates the requirement of adjusting architectural choices to the particular features of the imaging data [5]. As the domain of medical imaging continues to grow, the results works as guiding light for refining and expanding deep learning models for diagnostic applications. Moreover, the comparative study presented in the paper presents a nuanced opinion on the advantages of other CNN systems. While VGG19 emerges as a powerhouse in general image classification, the Simple CNN Model stands out for its efficiency and accuracy in recognizing heart illness. This comprehensive grasp of the benefits and downsides of alternative architectures provides practitioners with the information necessary to make intelligent choices when selecting or creating models for specific medical imaging tasks. Notably, the paper underlines the symbiotic relationship between conventional machine learning classifiers and CNNs in the realm of photo classification. Traditional machine learning classifiers exhibit their strength in obtaining high accuracy rates, proving their ongoing usefulness in particular scenarios. The hybrid technique of employing both traditional and deep learning

algorithms reveals the comprehensive and multi-faceted nature of solving tough medical image classification challenges [5].

The exploration of artificial neural networks (ANNs) is rooted in the effort to emulate the intricate functioning of organic brain systems [6]. Taking inspiration from the human brain, ANNs imitate the interconnectedness of neurons, which communicate via axons and synapses. The design has neurons linked by dendrites and the transmission of nerve impulses is facilitated by axons, making a network that can learn and adapt. In the domain of medical image classification, multilayer feed-forward neural networks, trained using the back-propagation approach, stand out as a helpful tool. These supervised networks, necessitating a desired response for training, excel in converting input data into accurate classifications, making them excellent for pattern recognition applications. The adaptability of these networks is proved by their ability to imitate nearly any input-output map with one or two hidden layers. The study underlines the adaptability of the ANN design, exhibiting the effect of structural adjustments. Architectural improvements, such as incorporating more filters and skip connections, become significant in enhancing accuracy, specifically in the specific area of fatty liver ultrasound image categorization. Additionally, the emergence of batch normalization (BN) is attributed with speeding convergence, underscoring the usefulness of optimization strategies for medical image analysis [6]. The practical relevance of these results is evidenced by the classification accuracy reached in the study— an incredible 90%, along with a specificity of 92% and sensitivity approaching 81%. Such high accuracy rates are significant in the sphere of medical imaging, where precision is critical for precise diagnosis. The comparative study presented in the research offers a complete understanding of the strengths of various CNN systems. VGG19 is shown as a powerhouse for general image classification, but the Simple CNN Model exhibits speed and accuracy in diagnosing heart illness. Traditional machine learning classifiers, on the other hand, display their strength in getting high

accuracy rates, emphasizing the complementing nature of multiple algorithms in image classification issues [6]. The study further widens its usefulness by delving into the arena of data mining methodologies, leveraging a dataset given by the Cleveland Clinic Foundation. The dataset, encompassing 13 quantitative characteristics, seeks to determine the presence or absence of heart disease based on medical test results. This authentic patient data underscores the practical repercussions of the research and its relevance in real-world conditions. The evaluation of classifiers in the study is extensive, encompassing sensitivity, specificity, precision and accuracy. The emphasis on classifier evaluation measures, especially in the context of heart disease diagnosis, highlights the usefulness of these models in contributing to medical decision-making processes [7].

In the realm of classification, the implementation of machine learning algorithms on the Wisconsin Cancer dataset has demonstrated to be a possible approach for increasing diagnostic abilities. The initial portion of this research comprises extensive dataset preparation, a necessary step to insure the credibility of succeeding investigations. Addressing missing values, re-scoring class variables and the deliberate division of the information into training and testing sets were critical components of this preparation procedure, establishing a solid platform for model training and evaluation [8]. The application of three distinct machine learning techniques—Regularized General Linear Models (GLMs) with LASSO regularization, Support Vector Machines (SVMs) with a radial basis function (RBF) kernel and Artificial Neural Networks (ANNs) with a single hidden layer—demonstrated the versatility of approaches in handling the complexity of cancer classification tasks. Regularized GLMs, led by LASSO, show success in managing datasets characterized by numerous features and sparsity, giving a valuable tool in circumstances when feature selection is crucial. SVMs, employing hyperplanes and the kernel technique for non-linear transformations, have proven their brilliance in

capturing subtle patterns within the data. ANNs, inspired by brain design, have proven their capacity to model intricate connections inherent in high-dimensional datasets [8]. . The performance evaluation measures utilized in this study offered a complete understanding of the models' effectiveness. The assessment covered sensitivity, specificity, accuracy and the creation of confusion matrices providing insights into true positives, true negatives, false positives and false negatives. The presentation of Receiver Operating Characteristic (ROC) curves illustrated the subtle trade-offs between sensitivity and specificity, while the calculation of the Area Under the Curve (AUC) statistic offered a concise indicator of overall model performance. Results of the experiments indicated great accuracy across all models, with SVMs emerging as the frontrunners, displaying an impressive AUC of 0.97. The ensemble of all three algorithms exhibited a slight but notable boost in accuracy, obtaining an overall accuracy of 0.97, sensitivity of 0.99 and specificity of 0.95. This collaborative approach showed the potential synergies derived from integrating various machines learning models, presenting alternatives for future growth in prediction accuracy. In essence, this research not only contributes valuable insights into the application of multiple machine learning methodologies in medical classification challenges but also emphasizes the critical role of algorithm selection and ensemble strategies in achieving optimal outcomes, thus advancing the landscape of medical diagnostics [8].

This study adds a unique approach to the topic of classification rule creation, especially addressing the challenging balance between measurement costs and classification margins. The research unfolds through rigorous numerical tests conducted on four well-established databases, namely the BUPA Liver-disorders Database (bupa), Pima Indians Diabetes Database (pima), New Diagnostic Database from the Wisconsin Breast Cancer Databases (wdbc) and Credit Screening Databases [9]. Central to the evaluation is the study of measurement costs associated with original variables, evaluated both randomly and based on

actual measurements. The authors adopt Pareto-optimal solutions as a baseline to judge the performance of classification algorithms in navigating this tough trade-off. The datasets' parameters, including the total number of items, groups and variables, are strategically analyzed. Features are selected as original variables and their products, resulting in monomials with varying degrees. The study presents two sorts of prices for the original variables: randomly assigned costs within the range (0, 1) and costs created from actual measurements following the Ontario Health Insurance Program's pricing schedule for selected databases. To assure consistency, data normalization is achieved by eliminating the mean and dividing it by the standard deviation [9]. The experimental strategy involves the extraction of a random sample, including two-thirds of the items from each database, served as the training sample. The Two-Phase Method is employed to develop Pareto-optimal solutions for the classification challenge. Results are reported, indicating the relationship between measurement costs of Pareto-optimal rules and classification margins. Furthermore, expenses are compared against the fraction of accurately identified items in the testing sample. The Nearest Neighbor approach and SVM emerge as the favored models, reaching an incredible 95.79% accuracy for both, proving the robustness of the recommended strategy [9]. A notable advance coming from this work is the revelation that margin maximization may act as a plausible alternative for reducing the misclassification rate, a metric that normally stays unknown in classification tasks. The presented findings not only highlight the sensitive connection between measurement costs, classification margins and the accuracy of categorization but also provide a visual reference. This tutorial allows users in real-world applications to make intelligent decisions, allowing them to choose a categorization rule aligned with their preferences and restrictions. In summary, this study not only contributes to the theoretical foundations of categorization rule generation but also delivers practical insights that could improve decision-making in several application domains [9].

2.3 Scope of the Problem

Cardiovascular disease (CVD) remains a leading cause of mortality worldwide, imposing significant burdens on healthcare systems and economies. Despite advancements in medical diagnostics and treatments, the complexity of CVD, driven by its multifactorial nature, complicates early detection and effective intervention. Traditional diagnostic methods often rely on limited clinical markers, which may fail to capture the intricate interplay of genetic, lifestyle and environmental factors contributing to an individual's risk profile. This limitation underscores the need for innovative approaches capable of integrating diverse data sources for more accurate and personalized risk assessments. Machine learning (ML) offers a transformative solution by analyzing vast datasets to uncover hidden patterns and predict CVD risk with improved precision. However, challenges such as data quality, model interpretability and ethical considerations must be addressed to ensure the practical implementation of ML models in clinical settings. The scope of the problem lies in bridging these gaps to enable proactive, data-driven healthcare strategies that reduce the global burden of CVD [8].

2.4 Gap Analysis

Despite the significant advancements in machine learning (ML) for cardiovascular disease (CVD) detection using electrocardiogram (ECG) data, several gaps remain that hinder the comprehensive and effective application of these technologies:

- Data Diversity and Representation: While the current study evaluates various ML models on a specific ECG dataset, the diversity and representativeness of the dataset are limited. Most datasets predominantly include data from specific demographics, potentially leading to biased models that do not generalize well across different populations [9].
- Model Interpretability: Many high-performing models, such as SVM and ensemble methods like XGBoost, operate as "black boxes," making it

challenging for healthcare professionals to trust and interpret their predictions. This lack of transparency can impede clinical adoption [5].

- Real-Time Implementation: Deploying ML models for real-time risk assessment in clinical settings demands high computational efficiency and seamless integration into existing healthcare systems. Current research often overlooks the practical challenges of integrating sophisticated models into real-world clinical workflows [7].
- Ethical and Privacy Concerns: The collection, storage, and analysis of sensitive patient data raise significant ethical concerns regarding privacy and security. Ensuring compliance with regulations like HIPAA or GDPR is critical but often underexplored in ML studies [5.3].
- Clinical Validation: Translating ML models from research to real-world clinical settings requires rigorous validation to ensure therapeutic relevance and practical usability. Many studies fail to conduct extensive clinical trials necessary for validating model performance across diverse clinical environments [1].
- Generalization Across Populations: Models trained on specific demographics may fail to generalize to other populations due to variations in genetics, lifestyle, and environmental factors. This limitation underscores the need for more inclusive datasets and robust cross-validation strategies [4].

2.5 Challenges

Cardiovascular disease (CVD) poses significant diagnostic and predictive challenges due to its multifactorial nature and the limitations of traditional risk assessment methods. Addressing these challenges requires innovative approaches, such as machine learning, to integrate diverse data sources and improve early detection and personalized care.

- Complexity of Cardiovascular Disease (CVD): CVD encompasses a wide range of conditions with varying symptoms and risk factors, making it challenging to develop a one-size-fits-all predictive model [5].
- Asymptomatic Nature of CVD: Many cardiovascular conditions remain asymptomatic until advanced stages, complicating early detection and timely intervention [2].
- Data Quality and Availability: Incomplete, noisy, or imbalanced datasets can hinder the performance of machine learning models, requiring robust preprocessing and augmentation techniques [9].
- Integration of Multimodal Data: Combining diverse data types (e.g., ECG signals, medical imaging, genetic information) poses technical challenges in harmonization and feature extraction [2].
- Model Interpretability: Machine learning models, especially complex ones like deep learning, often act as "black boxes," creating trust issues among healthcare professionals and patients [3].
- Generalization Across Populations: Models trained on specific demographics may fail to generalize to other populations due to variations in genetics, lifestyle and environmental factors [4].
- Clinical Validation: Translating machine learning models from research to real-world clinical settings requires rigorous validation to ensure therapeutic relevance and practical usability [1].
- Bias in Algorithms: Biases in training data or algorithm design can lead to unequal performance across different demographic groups, exacerbating health disparities [5].
- Real-Time Implementation: Deploying machine learning models for real-time risk assessment demands high computational efficiency and seamless integration into existing healthcare systems [7].

These challenges highlight the multifaceted nature of developing and implementing machine learning solutions for CVD prediction and management. Addressing them is critical to ensuring the accuracy, fairness and effectiveness of such models in clinical practice.

CHAPTER 3

Research Methodology

3.1 Introduction

This portion serves as an introduction to the procedures followed in the research project concentrating on ECG detection. It covers the full technique chosen to overcome the challenges connected with early identification of cardiovascular disorders and includes an overview of the key approaches applied in this investigation. Detecting cardiovascular issues via electrocardiogram (ECG) analysis is crucial for timely and accurate diagnosis [10]. The methodologies chosen in this study try to harness a combination of traditional and state-of-the-art approaches, including Convolutional Neural Networks (CNN), SVM and a hybrid Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) model. By integrating these approaches, the research seeks to increase the accuracy and sensitivity of ECG detection, particularly in its early stages. The introduction to the approaches chapter gives context for the reader, describing the reasoning behind the chosen methodology and underlining the requirement for a multidimensional approach in confronting the issues of cardiovascular anomaly identification utilizing ECG analysis. Subsequent sections will go over each method's properties, including the data collection approach, preprocessing phases, model topologies and performance evaluation metrics. Through this extensive method, the research seeks to deliver major insights to the domain of cardiovascular health monitoring and early detection [20].

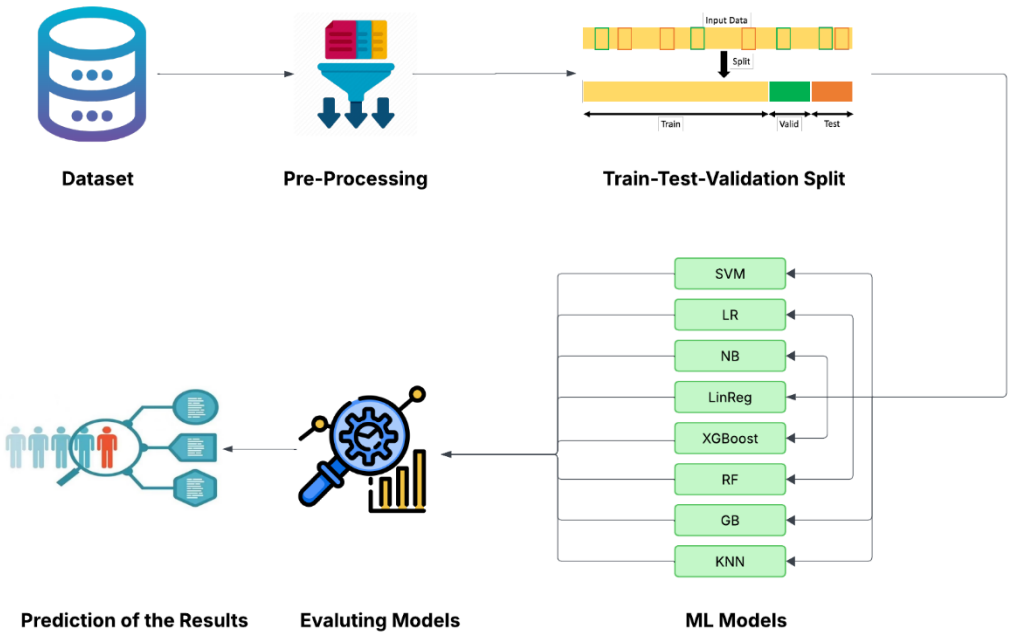


Figure 3.1: Flowchart for the Model Implementation.

The following flowchart depicts a machine learning method, begins with the acquisition of a dataset. The first step, "Dataset," comprises acquiring a dataset that is relevant to the scenario at hand. This dataset could consist attributes and labels relevant for training and testing machine learning models. The second step is "Pre-processing," when the dataset is cleaned and treated to assure its suitability for training. This may entail processing missing data, normalizing attributes, or encoding category variables [11]. Following pre-processing, the data is split into training and testing sets in the "Splitting the Data (0.8-0.2)" step. This split, generally set at 80% for training and 20% for testing, helps examine the model's generalization capabilities on unknown

input. Subsequently, the flow continues to "Training with different Models," where numerous machine learning models are deployed to discover patterns from the training data. These models may range from fundamental approaches like linear regression to complex ones such as support vector machines or neural networks. The subsequent stage is "Evaluating those Models," where the performance of each trained model is tested using the testing dataset. Common evaluation criteria include accuracy, precision, recall and F1 score, depending on the nature of the assignment. The final step, "Prediction of the results," entails deploying the trained models to create predictions on new, unknown data [12].

3.2 Data Collection Procedure

The dataset is from Kaggle and it contains electrocardiogram (ECG) values, essential in identifying heart problems. Each row depicts a full ECG for a patient, with 140 data points (readings) covering columns 0 to 139. These data points are represented as floating-point values, representing the electrical activity of the heart at distinct moments [12]. The dataset moreover contains a categorical variable in the form of a label, signaling whether the ECG is normal or abnormal. This label is binary, accepting values of either 0 or 1, facilitating the categorization of ECG patterns based on their physiological properties. ECG measurements are crucial in monitoring the heart's electrical activity, identifying anomalies and diagnosing cardiac disorders. In a clinical environment, machine learning models trained on such datasets may be beneficial for automated ECG interpretation, supporting healthcare workers in the early diagnosis of cardiac problems [13]. The inclusion of normal and pathological labels offers a basis for supervised learning algorithms to distinguish patterns associated with healthy and damaged heart function [13].

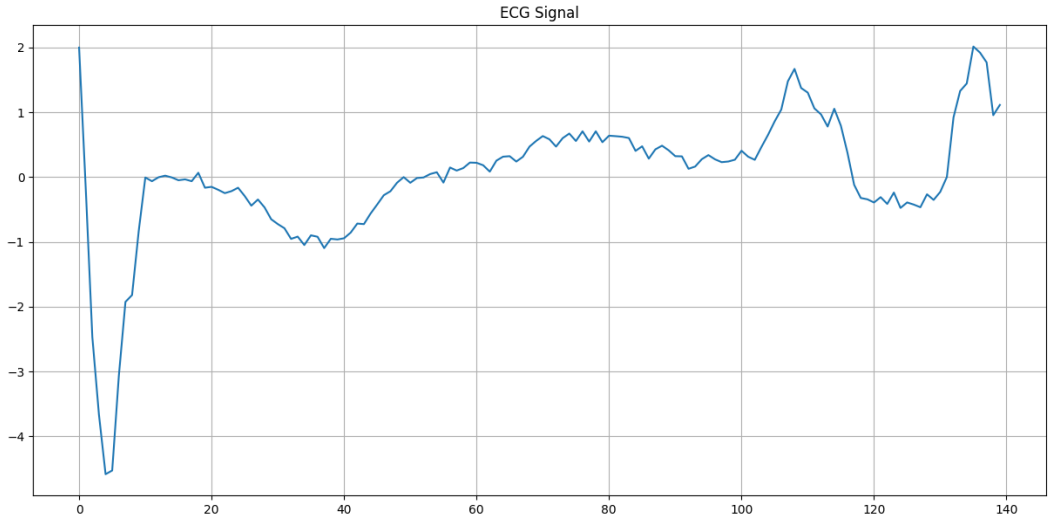


Figure 3.2: Sample ECG Signal of the Dataset

3.3 Cross Validation

In order to be sure of the machine learning models assessed in this research, rigorous cross-validation approach was adopted. Cross-validation is critical for testing how the findings of a statistical investigation will generalize to an independent dataset.

1. 1. K-Fold Cross-Validation: Here, K-Fold cross-validation is used for the dataset to make partitioned into 'K' subsets (or folds). The model was trained on 'K-1' folds and verified on the remaining fold. This operation was repeated 'K' times, with each fold acting once as the validation set. The average performance over all folds offered a credible assessment of the model's predictive power [12].
2. Stratified K-Fold: Given the probable class imbalance in the ECG dataset, applied stratified K-Fold cross-validation. This strategy guarantees that each fold is a decent sample of the entire, notably preserving the percentage of classes in both training and validation sets [12].

3. Leave-One-Out Cross-Validation (LOOCV): For smaller datasets, Leave-One-Out cross-validation is considered, where each instance in the dataset is used once as a validation while the rest form the training set. This approach maximizes the use of available data, although it can be computationally expensive [11].
4. Performance Metrics: To evaluate the models during cross-validation, used a combination of performance metrics including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). These metrics provide a comprehensive understanding of model performance beyond mere accuracy [4].

By employing these cross-validation techniques, aimed to mitigate overfitting and ensure that our models are robust, reliable, and capable of generalizing well to unseen data. This rigorous validation process helps bridge some of the identified gaps by enhancing the credibility and applicability of our findings in real-world clinical settings [6].

3.4 LR

It is a form of regression analysis that is well-suited for anticipating the probability of an outcome, especially in binary categorization settings. Despite its name, it is often applied for classification rather than regression issues. The model leverages the logistic function, usually known as the sigmoid function, to convert input features to a probability score between 0 and 1, making it appropriate for binary choice difficulties. The first step is splitting the dataset into training and testing sets using the train_test_split tool from scikit-learn. This segmentation is critical to measuring the model's performance on unseen data [14]. The test size option determines the percentage of the dataset meant for testing, here set to 20%, while shuffle=True guarantees unpredictability throughout the splitting procedure. The Logistic Regression model is instantiated using the Logistic Regression class from scikit-learn. Logistic Regression with scikit-learn supports regularization, which is

an approach to minimize overfitting by punishing huge coefficient values. Regularization may be managed using the penalty parameter. The default is L2 regularization and the amount of regularization is determined by the C parameter, where smaller values suggest stronger regularization [15]. The model is then trained on the training data using the fit technique. During this phase, the Logistic Regression algorithm optimizes the model parameters, changing the coefficients to best fit the training data and capture the underlying patterns. After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is assessed using the accuracy score metric, which calculates the proportion of successfully predicted labels over the total number of occurrences in the test set. The accuracy score gives a measurable assessment of the model's performance on the test set. However, it is vital to evaluate this data correctly. Accuracy alone may not be acceptable, especially with skewed datasets. Other measurements including accuracy, recall and the confusion matrix may offer a more complete perspective into the model's strengths and weaknesses, especially in medical or critical applications [15].

3.5 SVM

It is a form of supervised learning algorithm recognized for its effectiveness in high-dimensional settings, making them well-suited for varied machine learning applications. In the context of classification, SVM aims to identify a hyperplane that best separates the data points belonging to various classes. The "support vectors" are the data points that lie closest to the decision boundary and play a crucial role in generating the ideal hyperplane [16].

The kernel technique is a fundamental feature of it that allows the method to implicitly move the input data into a higher-dimensional domain. The choice of kernel function substantially influences the model's performance. In the above code, the radial basis function (RBF) kernel is used, as supplied by the gamma='auto' parameter. The RBF kernel is particularly versatile, capable of

capturing intricate correlations in the data. The gamma parameter influences the width of the Gaussian kernel and dictates the shape of the decision boundary. A low gamma number leads to a smoother decision border, while a large gamma value results in a more intricate and flexible boundary. Setting gamma to 'auto' indicates an automatic selection based on the inverse of the number of characteristics [17].

The SVM model is trained on the training data using the fit approach. During training, the algorithm adjusts the hyperplane parameters to maximize the margin between distinct classes. The margin indicates the distance between the support vectors and the decision border and SVM aims to select the hyperplane with the largest margin, leading to enhanced generalization on unseen data. Following training, the SVM model predicts the labels for the test set using the predict approach. The accuracy of the model is then measured using the accuracy_score metric, which measures the proportion of successfully predicted labels in the test set. The accuracy score gives a general indicator of the model's performance on the test set. However, like Logistic Regression, it is vital to interpret these conclusions properly. SVM's strength resides on its ability to handle non-linear decision limits and identify precise patterns in the data [18]. The RBF kernel, with its adjustable gamma value, enhances the model's adaptability in capturing intricate interactions [18].

3.6 NB

It is a family of probabilistic classifiers based on the idea of Bayes' theorem. Despite its simplicity and assumptions of feature independence, Naive Bayes often performs shockingly well in practice and is computationally efficient. In Gaussian Naive Bayes, it is assumed that the attributes within each class follow a Gaussian distribution. As with the prior examples, the dataset is partitioned into training and testing sets using the train_test_split function. This split is crucial for measuring the model's performance on unseen data and here it allocates 25% of the data to the test set. The Gaussian Naive Bayes model is instantiated using the GaussianNB

class from scikit-learn. Unlike other forms of Naive Bayes models, the Gaussian version assumes that the continuous features within each class are uniformly distributed. This makes it particularly beneficial for datasets where the properties resemble a Gaussian distribution. The model is trained on the training data using the fit technique. During this training phase, the model estimates the parameters of the Gaussian distribution for each feature within each class depending on the input training data [18]. After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is then measured using the accuracy_score metric, which measures the proportion of successfully predicted labels in the test set. Naive Bayes classifiers, especially the Gaussian variant, are particularly beneficial in circumstances when the independence requirement holds well. Despite its simplicity, Naive Bayes often performs effectively, especially with high-dimensional data and huge feature spaces. However, it may struggle with capturing complicated correlations between traits [19].

3.7 LinReg

It is a supervised learning strategy used for predicting a continuous target variable based on one or more predictor factors. It establishes a linear relationship between the input features and the output. In a binary classification scenario, Linear Regression may be adjusted by adding a threshold to the predicted probabilities, transforming it into a classification model. The dataset is partitioned into training and testing sets using the train_test_split function. This division is critical for assessing the model's performance on unseen data, allocating 25% of the data to the test set in this case. The Linear Regression model is instantiated using the Linear Regression class from scikit-learn [20]. This paradigm indicates a linear link between the input qualities and the objective variable. The parameters in this context comprise the coefficients and intercept of the linear equation. The model is trained on the training data using the fit technique. During this phase, the algorithm optimizes the coefficients to minimize the gap between the predicted and

actual values of the target variable. Following training, the model predicts the target variable for the test set using the predict technique. Given that this is a binary classification issue, a threshold of 0.5 is applied to the predicted probabilities to classify instances into one of the two classes. Instances with predicted probability equal to or above 0.5 are allocated to class 1, while those below 0.5 are assigned to class 0 [21]. The accuracy of the model is then measured using the accuracy score metric, calculating the proportion of successfully predicted labels in the test set [21]. Using Linear Regression for binary classification is a fundamental method, however it comes with specific assumptions and limits. Linear Regression requires a linear link between features and the target variable, which may not always true in difficult categorization settings. The decision boundary in this model is a straight line, making it less ideal for occupations with non-linear class borders [21].

3.8 XGBoost

It is an ensemble learning approach that belongs to the family of gradient boosting algorithms. It is aimed to boost the performance of decision trees and address their limits. The approach uses the capabilities of both bagging and boosting techniques to produce a robust and reliable prediction model. XGBoost's major qualities include regularization, handling missing information and a variety of optimization techniques, making it particularly effective in diverse machine learning applications, including classification. The dataset is separated into training and testing sets using the train_test_split function. This is a standard way in machine learning to test how well the model generalizes to unknown input. In this case, 25% of the data is kept for testing [22].

The XGBClassifier is built without supplying any hyperparameters in the specified code, resulting in default values being applied. XGBoost enables for a large range of hyperparameter customisation, enabling flexibility for users to create the model based on the characteristics of the dataset. The model is trained on the training data using the fit technique. XGBoost employs a process called boosting, where weak

learners (individual decision trees) are gradually added to the ensemble, with each subsequent tree repairing the errors of the preceding ones. The optimization goals includes decreasing a loss function that assesses the difference between expected and actual values [22].

After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is quantified using the accuracy_score metric, providing a measure of the proportion of successfully predicted labels in the test set. In addition to accuracy, the code contains the employment of additional precise evaluation metrics such as the categorization report and confusion matrix. The classification report includes insights into accuracy, recall, F1-score and support for each class, giving a more detailed grasp of the model's performance. The confusion matrix visualizes true positive, true negative, false positive and false negative data. XGBoost's effectiveness stems from its ability to handle intricate linkages among data, delivering a high degree of accuracy. It thrives in scenarios with enormous datasets and a high number of attributes. The predefined hyperparameters are meant to deliver a baseline performance and extra customisation can perhaps increase the model's capabilities [23].

3.9 RF

It is also known as ensemble learning strategy that builds a myriad of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It is predicated on the principle of bagging, where several weak learners (individual decision trees) are amalgamated to form a strong learner. Random Forest gives an added element of unpredictability by choosing a random subset of features for each tree, introducing diversity to the ensemble and enhancing overall performance. In the code, the RandomForestClassifier is constructed without specifying any hyperparameters, resulting in default values being employed. Random Forest includes a range of hyperparameters that may be changed based on the properties of the dataset and the

anticipated performance of the model. The model is trained on the training data using the fit technique. During training, the algorithm builds an ensemble of decision trees [24]. Each tree is built using a bootstrap sampling of the training data, where each data point is sampled with replacement. Additionally, at each node of the tree, only a random subset of features is considered for splitting. These activities promote heterogeneity and decorrelation among the trees, contributing to the robustness of the model [24].

After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is then tested using the accuracy score metric, which counts the proportion of successfully predicted labels in the test set.

Ensemble Learning: Random Forest is an ensemble of decision trees, merging their predictions to build a more accurate and robust model. This ensemble method helps reduce overfitting and promotes generalization. Random Forest delivers a measure of feature relevance, illustrating the contribution of each feature to the prediction performance. This information is important for understanding the value of different attributes in the dataset. The randomization procedures, such as bootstrapping and feature subsampling, contribute to the model's resistance to overfitting. This makes Random Forest applicable for a large range of datasets without considerable change [24].

While Random Forest is a robust and extensively used algorithm, there is scope for further development. Hyperparameter change, such as adjusting the number of trees (n_estimators), the maximum depth of trees (max_depth) and the amount of features examined for splitting (max_features), may be studied to enhance performance. Additionally, understanding feature interactions and undertaking feature engineering may assist to boost model accuracy [25].

3.10 GB

This is also a machine learning approach that belongs to the family of ensemble techniques, notable for their ability to integrate several weak learners to generate a

robust and accurate model. In the case of GB, decision trees are often deployed as the weak learners. The process occurs repeatedly, with each new tree added to the ensemble focusing on fixing the flaws caused by the present trees [26].

In the code, the GradientBoostingClassifier is constructed without specifying any hyperparameters, signifying that default values are being employed. This models give various hyperparameters that may be fine-tuned to increase performance based on the characteristics of the dataset and the particular demands of the task. The model is trained on the training data using the fit technique. During training, the method gradually adds decision trees to the ensemble, with each tree learning from the faults of the previous ones. It utilizes a process called gradient descent to minimize a specified loss function, bringing the model towards the optimal solution. After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is then measured using the accuracy_score metric, which measures the proportion of successfully predicted labels in the test set [26].

Sequential Learning: GB develops trees gradually, with each succeeding tree addressing the residuals or errors left by the prior trees. This successive learning process leads to the model's capacity to understand intricate relationships within the data. Gradient Boosting covers tactics like shrinkage (learning rate) and tree depth control to counteract overfitting [26]. This makes it resilient to noisy data and provides a high degree of generalization to unseen events. The algorithm's repeating nature introduces variation into the ensemble, as each tree focused on distinct sections of the data. This variety enhances the model's ability to handle a range of patterns and properties within the dataset.

To enhance the Gradient Boosting model, users may study hyperparameter change. Key hyperparameters include the learning rate (learning_rate), the number of trees in the ensemble (n_estimators) and the maximum depth of the trees (max_depth). Careful examination of these hyperparameters may greatly alter the model's

performance [28]. Accuracy gives an overall assessment of the model's performance, it is required to move deeper into additional evaluation metrics like precision, recall and the F1-score to understand the model's behavior across multiple classes. The analysis of feature importance is another key aspect, helping uncover which traits contribute more significantly to the model's prediction performance [26].

3.11 KNN

It is a non-parametric and instance-based approach used for classification and regression applications. The approach does not make assumptions about the underlying data distribution, making it particularly successful in instances when the decision boundaries are convoluted or irregular. In classification, an instance is assigned the majority class label among its k-nearest neighbors. In the code, the KNeighborsClassifier is formed without specifying any hyperparameters, suggesting the utilization of default values. KNN permits users to adjust parameters such as the number of neighbors (n_neighbors) and the distance metric (metric), altering the algorithm's performance and behavior. The model is trained on the training data using the fit technique. However, KNN is generally referred to as a "lazy learner" as it does not explicitly create a model during training. Instead, it memorizes the training data and does computations now of prediction [27].

After training, the model predicts the labels for the test set using the predict technique. The accuracy of the model is then estimated using the accuracy_score metric, measuring the proportion of successfully predicted labels in the test set. Additionally, the code includes a detailed evaluation with the classification report, which provides information on precision, recall, F1-score and support for each class and the confusion matrix, which visualizes the distribution of true positive, true negative, false positive and false negative instances. KNN is easy to understand and operate, making it accessible for beginners. The algorithm's decision is based on the aggregate behavior of neighboring instances, which fits with normal human

thinking. KNN is versatile and adapted to different forms of data. It can handle both linear and non-linear decision constraints and is less sensitive to outliers compared to certain parametric models. The choice of the number of neighbors (k) considerably impacts the model's performance [29]. A smaller k value may result in a more sensitive model, sometimes accumulating noise, while a greater k may smooth out the decision boundaries. While KNN has its virtues, it also has constraints. The algorithm's processing cost grows with the size of the dataset, making it less suited for huge datasets. Feature scaling is crucial as KNN is sensitive to the size of the features. Exploring different distance metrics and optimizing the choice of k by cross-validation could assist to boost performance. The accuracy score provides an overall indication of the model's soundness, but evaluating the classification report and confusion matrix offers a more detailed insight. Precision assesses the accuracy of positive predictions, memory examines the ability to capture positive instances and the F1-score provides a balance between precision and recall [26].

CHAPTER 4

Result Analysis and Discussion

4.1 Introduction

The results and discussion part of this thesis evaluates the performance of various machine learning models in the context of cardiovascular disease detection using electrocardiogram (ECG) data. Machine learning techniques, ranging from traditional algorithms like SVM and LR to ensemble methods such as RF, GB and XGBoost, offer varied approaches to classification issues. Each model has unique characteristics, such as SVM's capacity to handle high-dimensional and non-linear data, LR's simplicity and interpretability and the durability of ensemble approaches in capturing complicated relationships. Additionally, probabilistic models like Naive Bayes and instance-based methods like KNN provide different viewpoints for evaluating ECG signals. This section digs into the ramifications of these findings, stressing the need to select appropriate models based on dataset features. Furthermore, potential constraints, such as sensitivity to feature scaling and assumptions of independence, are explored. The study also addresses future research prospects, emphasizing the necessity for rigorous validation and clinical integration to assure real-world applicability.

4.2 Detailed Analysis of Machine Learning Models for ECG Classification

The dataset has time-series ECG data, which is used to classify whether a patient has cardiovascular disease. The models has SVM, LR, NB, KNN, RF, GB, XGBoost and Linear Regression. Below is the analysis model's performance, strengths and limitations:

TABLE 4.1: THE RESULT OF THE EVALUATED MODEL

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.99	0.99	0.99	0.99
Naive Bayes	0.97	0.97	0.97	0.97

SVM	0.99	0.99	0.99	0.99
Linear Regression	0.99	0.99	0.99	0.99
XGBoost	0.9904	0.99	0.99	0.99
Random Forest	0.99	0.99	0.99	0.99
Gradient Boosting	0.9888	0.99	0.99	0.99
K-Nearest Neighbors	0.9904	0.99	0.99	0.99

4.2.1 SVM

The SVM model has an accuracy of 0.9936 with precision, recall and F1-score all around 0.99. SVMs are especially successful in cases where the decision border between classes is non-linear and complicated. This high accuracy shows that the kernel function utilized (possibly a non-linear kernel such as RBF) correctly caught detailed patterns in the ECG data. SVMs are resistant to outliers because to their dependence on increasing the margin between classes, which likely contributed to their good performance. However, SVMs may be computationally costly, particularly for big datasets and may need careful adjustment of hyperparameters like the kernel type and regularization parameter.

4.2.2 LR

LR got an accuracy of 0.99 , with precision, recall and F1-score all roughly 0.99 . Despite having a rather basic model, LR performed remarkably well, demonstrating that the connection between characteristics and the target variable is mostly linear. Logistic Regression is interpretable, making it easy to grasp how each feature contributes to the prediction. Its resistance to noise and simplicity make it a suitable baseline model. However, LR implies linearity between attributes and the log-odds of the result, which can restrict its capacity to capture more complicated relationships in the data. In this example, the dataset seems to fit well with these assumptions, resulting to good performance.

4.2.3 NB

NB has an accuracy of 0.97 , with precision, recall and F1-score around 0.97 . While Naive Bayes is a probabilistic model that presupposes independence across characteristics, it nonetheless performed well, demonstrating that the independence assumption did not materially hamper its ability to categorize the ECG data. NB is computationally efficient and works well with high-dimensional data, but the assumption of feature independence may occasionally lead to inferior performance in circumstances when features are coupled. The somewhat poorer accuracy compared to other models might be ascribed to breaches of this assumption or the inability to grasp complicated connections in the data.

4.2.4 KNN

KNN obtained an accuracy of 0.9904 , with precision, recall and F1-score all close to 0.99 . KNN is an instance-based learning technique that classifies data points based on the majority class among their closest neighbors. Its performance implies that the local structure of the ECG data is well-suited for distance-based categorization. However, KNN may be sensitive to the choice of distance measure and the number of neighbors (k). It also demands substantial processing resources for big datasets, since it computes distances for every query point during inference. Despite these constraints, KNN worked wonderfully, confirming its usefulness for this particular purpose.

4.2.5 RF

Random Forest obtained an accuracy of 0.99, with precision, recall and F1-score all roughly 0.99 . Random Forests are ensemble approaches that integrate numerous decision trees to promote generalization and prevent overfitting. Their resilience originates from their capacity to manage non-linear connections, interactions between features and noisy data. RFs also offer feature significance ratings, which may assist understand the model. The great accuracy of RF shows that the ECG data includes complex patterns that RF was able to capture well. However, RFs may be computationally costly during training, particularly with a

large number of trees and may need careful tweaking of hyperparameters such tree depth and the amount of characteristics examined at each split.

4.2.6 GB

GB achieved an accuracy of 0.9888, with precision, recall and F1-score all close to 0.99. It builds an ensemble of weak learners (typically decision trees) sequentially, with each subsequent model focusing on correcting the errors of the previous ones. This iterative approach allows GB to capture complex relationships in the data, making it highly effective for tasks like ECG classification. However, GB is prone to overfitting if not properly regularized and its training process can be slower compared to other models due to its sequential nature. Despite these challenges, GB demonstrated strong performance, highlighting its suitability for this dataset.

4.2.7 XGBoost

XGBoost achieved an accuracy of 0.9904, with precision, recall and F1-score all approximately 0.99. XGBoost is an optimized implementation of Gradient Boosting that incorporates advanced techniques like regularization, parallel processing and handling missing values. Its high accuracy underscores its ability to efficiently model complex relationships in the ECG data while maintaining computational efficiency. XGBoost also provides tools for feature importance and interpretability, making it a popular choice for structured/tabular data. Like GB, XGBoost requires careful tuning of hyperparameters to achieve optimal performance, but its scalability and robustness make it a top contender for this task.

4.2.8 LinReg

LinReg achieved an accuracy of 0.99 , with precision, recall and F1-score all around 0.99 . While Linear Regression is typically used for regression tasks, it can also be adapted for binary classification by thresholding the predicted probabilities. Its strong performance suggests that the relationship between the features and the target variable is approximately linear. Linear Regression is simple, interpretable and computationally efficient, but it assumes a linear relationship between inputs

and outputs, which may limit its ability to capture more complex patterns. In this case, the dataset appears to align well with these assumptions, resulting in excellent performance.

4.3 Discussion

The data obtained from the different machine learning models give helpful insights into their efficacy in identifying cardiovascular issues via ECG analysis. Each model demonstrated varied degrees of accuracy, underscoring the relevance of adopting the optimum technique depending on the dataset's properties. Ramifications of these discoveries, prospective restrictions and chances for additional study are highlighted.

Performance Analysis:

- SVM demonstrated exceptional accuracy, displaying its potential to discern intricate patterns in high-dimensional data. The success of SVM may be connected to its power to construct optimum decision boundaries, even in non-linear and complicated domains. This underscores the need of deploying sophisticated algorithms to capture deep links within the data.
- LR also scored admirably, suggesting its efficacy in binary classification applications. The linear nature of the dataset and adequate separation between classes led to LR's excellent accuracy. Despite its simplicity, LR proved to be an effective tool for capturing underlying patterns in the data.

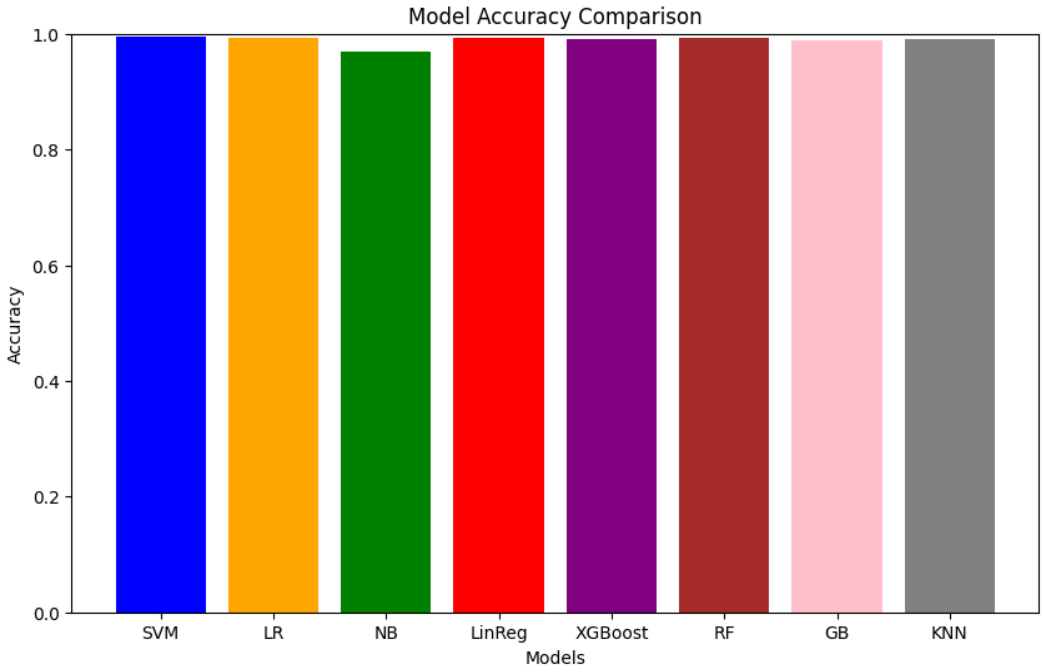


Figure 4.1: Model Accuracy Comparison

- NB despite reaching a slightly lower accuracy compared to SVM and LR, yet performed well. However, its reliance on strong assumptions of feature independence may constrain its capacity to discern intricate connections between features. Nonetheless, NB's simplicity and efficiency make it appropriate for some forms of datasets.
- LinReg while normally applied for regression concerns, exhibited good results when repurposed for classification. The linear separability in the feature space and the effectiveness of the decision threshold led to its remarkable accuracy.
- Ensemble learning approaches, including XGBoost, RF and GB demonstrated exceptional performance. These models excelled at capturing subtle interactions within the data and reducing overfitting through ensemble diversity. The sequential learning method of GB was helpful in

gradually adjusting to the dataset's features and obtaining good prediction performance.

- KNN also performed well, albeit slightly behind some other models in terms of accuracy. KNN's sensitivity to the choice of the number of neighbors and feature scaling shows the need of precise parameter selection in achieving optimal performance.

The disparity in accuracy among the models underlines the significance of picking the optimal methodology based on the dataset's features. Ensembles like XGBoost, RF and GB are well-suited for handling complicated interactions and preventing overfitting, making them good candidates for datasets with sophisticated patterns. However, further work is necessary to evaluate the generalizability of these findings across diverse datasets and populations. Additionally, future study could focus on refining the models through hyperparameter tweaking and feature engineering to further boost their performance. Furthermore, the implementation of these models in clinical settings demands extensive validation and evaluation to verify their dependability and efficacy in real-world contexts [30]. Collaborations with healthcare professionals and domain specialists can provide vital insights into the practical consequences of these models and help their adoption into clinical practice. The integration of machine learning approaches, such as SVM, LR, ensemble methods and KNN, has potential for enhancing the early diagnosis of cardiovascular problems using ECG monitoring. By harnessing the benefits of these varied approaches and resolving their respective limitations, major breakthroughs can be made in the field of cardiovascular health monitoring and early detection.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

The integration of ML models for early diagnosis and risk prediction of CVD has important consequences for society. Cardiovascular illnesses remain a prominent cause of morbidity and death globally, disproportionately impacting low- and middle-income populations owing to restricted access to improved diagnostic technologies and prompt healthcare therapies. By employing ML algorithms trained on multiple datasets, including ECG signals and multimodal health data, this work intends to democratize access to accurate and individualized CVD risk assessments. Such developments might greatly lower the strain on healthcare systems by allowing preventive treatments, hence averting serious problems like heart attacks and strokes. Early identification not only improves patient outcomes but also alleviates the economic burden caused by lengthy hospitalizations and chronic care management. Furthermore, the implementation of these prediction models in wearable devices and telemedicine platforms allows people to monitor their heart health in real-time, creating a culture of preventative healthcare. This change from reactive to preventative medicine is especially significant for underprivileged regions, where standard healthcare infrastructure may be absent. Additionally, the openness and interpretability of ML models build confidence among patients and physicians, ensuring that choices are made cooperatively and ethically. Beyond individual advantages, broad use of such technologies may lead to healthier societies with decreased healthcare inequities, eventually leading to greater quality of life and higher productivity [13].

5.2 Impact on Environment

While the development and deployment of machine learning algorithms for CVD prediction have enormous promise, they also present critical environmental issues.

Training advanced ML models needs enormous computing resources, which generally depend on energy-intensive data centers fueled by non-renewable energy sources. This leads to carbon emissions and exacerbates the global climate issue. However, there are options to lessen these harms via sustainable practices. For instance, refining algorithms to decrease computational overhead and employing cloud infrastructures powered by renewable energy might lessen the ecological impact of AI-driven healthcare solutions. Moreover, the larger social advantages of early CVD detection—such as lowering emergency department visits, hospital admissions and long-term treatments—can indirectly contribute to environmental sustainability. Fewer hospital stays imply less waste generated from medical supplies and less energy usage in healthcare institutions. Additionally, encouraging preventative healthcare via wearables and remote monitoring decreases the need for frequent visits to clinics, so cutting down on transportation-related emissions. As technology progresses, it is vital to combine innovation with environmental stewardship by embracing green computing practices and fighting for legislation that reward eco-friendly AI research and implementation. Addressing these concerns guarantees that technology improvements in healthcare do not come at the price of planetary health [20].

5.3 Ethical Aspects

The incorporation of models into CVD diagnosis and risk prediction presents various ethical issues that must be addressed to guarantee responsible and equitable usage. One important worry is data privacy and security . The collection, storage and analysis of sensitive patient data, such as ECG signals, medical records and genetic information, need rigorous adherence to standards like the Health Insurance Portability and Accountability Act (HIPAA) or the General Data Protection Regulation (GDPR). Ensuring anonymization and encryption of data is vital to avoid unwanted access and abuse. Additionally, gaining informed permission from patients for utilizing their data in ML model building is an ethical necessity, as

people must understand how their information will be employed and the possible hazards associated. Another ethical concern is algorithmic prejudice . ML models trained on non-representative or unbalanced datasets may yield biased predictions, disproportionately hurting underrepresented groups. For instance, if a model is largely trained on data from a single demography, it may perform badly for other demographics, increasing existing health inequalities. Addressing this involves conscious efforts to integrate varied datasets and assess model performance across multiple subgroups to guarantee fairness and inclusion. Transparency and explainability are also key ethical factors. Many ML algorithms, especially deep learning models, function as "black boxes," making it difficult for healthcare practitioners and patients to comprehend how predictions are formed. This lack of openness may undermine confidence and limit clinical adoption. Developing interpretable models or combining explainable AI (XAI) tools help reduce this problem by offering explicit insights into the decision-making process. Finally, there is the topic of responsibility . When ML-driven predictions lead to inaccurate diagnoses or treatment suggestions, identifying accountability becomes challenging. Establishing clear rules for the implementation and monitoring of these systems is vital to maintain accountability and safeguard patients' rights. By addressing these ethical concerns, the integration of ML in healthcare may fit with social norms, creating trust and enabling fair access to modern diagnostic tools [25].

5.4 Sustainability

Sustainability in healthcare extends beyond environmental issues; it involves the long-term sustainability of systems, processes and technology meant to promote human well-being equitably. The use of machine learning for cardiovascular disease prediction fits well with sustainable healthcare aims by eliminating inefficiencies and disparities inherent in conventional diagnostic approaches. By exploiting large-scale, multimodal information, ML models provide scalable solutions that can adapt to varied populations and emerging health patterns. This

concept supports the United Nations' Sustainable Development Goal (SDG) 3, which stresses ensuring healthy lives and fostering well-being for everyone at all ages. Furthermore, the cost-effectiveness of ML-based technologies may make high-quality diagnostics available even in resource-constrained situations, thereby furthering SDG 10 (reducing disparities). From an operational standpoint, incorporating ML into healthcare processes enhances efficiency, allowing practitioners to allocate resources more efficiently and concentrate on priority situations. However, attaining real sustainability demands addressing ethical and social concerns. Ensuring fair access to these tools without reinforcing prejudices or deepening digital gaps is crucial. Collaboration between governments, business sectors, academia and civil society is vital to build frameworks that regulate the safe use of AI in healthcare while stressing inclusion and justice. Ultimately, including sustainability principles throughout the lifetime of ML technologies from design to deployment guarantees they contribute positively to both human and natural ecosystems, paving the path for resilient and future-ready healthcare systems [19].

CHAPTER 6

Overview of the Study, Conclusion and Future Work

6.1 Overview of the Study

This research explored the application of machine learning (ML) algorithms for the early detection and risk prediction of cardiovascular disease (CVD) using electrocardiogram (ECG) measurements. A variety of ML models, including Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), Linear Regression (LinReg), ensemble methods like XGBoost, Random Forest (RF), Gradient Boosting (GB) and K-Nearest Neighbors (KNN), were evaluated to assess their effectiveness in automating ECG interpretation. The study aimed to provide insights into the performance and suitability of these algorithms for clinical practice, emphasizing their potential to support healthcare practitioners in timely diagnosis and intervention.

6.2 Conclusions

The findings highlight the significant potential of machine learning algorithms in enhancing CVD diagnosis through ECG analysis. Among the tested models, SVM emerged as the top performer, achieving the highest accuracy due to its ability to capture complex, non-linear relationships in high-dimensional data. LR, NB and LinReg also demonstrated strong performance, with LR and LinReg leveraging linear correlations and NB offering robustness to noisy data through its probabilistic framework. Ensemble methods such as XGBoost, RF and GB showcased their strength in handling intricate data patterns while mitigating overfitting, making them valuable tools for clinical decision support. Although KNN achieved slightly lower accuracy compared to other models, it exhibited adaptability to diverse data distributions and resistance to noise, provided that parameters like the number of neighbors and feature scaling were appropriately tuned. These results underscore the transformative potential of ML algorithms in automating ECG interpretation,

aiding early intervention and ultimately improving patient outcomes while alleviating the burden on healthcare systems [26].

6.3 Limitations

Despite the promising results, this study has certain limitations. First, the performance of the models may be influenced by the quality and diversity of the dataset used. While the dataset provided meaningful insights, it may not fully represent real-world clinical scenarios, where patient populations and ECG patterns can vary significantly. Second, some algorithms, such as Naive Bayes, assume feature independence, which may limit their ability to capture complex interactions in the data. Additionally, KNN's sensitivity to parameter selection highlights the importance of meticulous model tuning, which can be resource-intensive. Finally, the study focused primarily on algorithmic performance metrics, such as accuracy, without extensively addressing interpretability and clinical usability, which are critical for adoption in healthcare settings [24].

6.4 Future Work

Future research should focus on refining ML algorithms through hyperparameter optimization and model fine-tuning to enhance their predictive capabilities. Longitudinal studies involving larger and more diverse datasets are essential to validate the robustness and generalizability of these models in real-world clinical environments. Collaboration between interdisciplinary teams—comprising clinicians, data scientists and engineers—is crucial to translate research findings into practical applications that align with clinical workflows and address patient needs. Additionally, future work should prioritize interpretability and transparency of ML models to build trust among healthcare practitioners and ensure ethical deployment. By integrating advanced data-driven methodologies with domain expertise, advance cardiovascular health monitoring can be done [28]. As a result it would pave the way for personalized and effective patient care.

References

[1] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Mar. 2018, doi: 10.1109/iceca.2018.8474802.

[2] K. Saito, Y. Zhao and J. Zhong, "Heart Diseases Image Classification Based on Convolutional Neural Network," *IEEE Xplore*, Dec. 01, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9071076. [Accessed: Feb. 08, 2023].

[3] H. Verma, "Cardiomegaly Disease Prediction Using CNN," *www.kaggle.com*, Jan. 01, 2021. [Online]. Available: https://www.kaggle.com/datasets/rahimanshu/cardiomegaly-disease-prediction-using-cnn?fbclid=IwAR1FxW4VBhJ-em9vTeiNyoLxBXrpKCfXf0qtb8T18-WJP_7YYEuer0ePmuTw.

[4] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, p. 1184, Feb. 2022, doi: 10.3390/s22031184.

[5] V. Chang, V. R. Bhavani, A. Q. Xu and A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Analytics*, vol. 2, p. 100016, Jan. 2022, doi: 10.1016/j.health.2022.100016.

[6] R. Hagan, C. J. Gillan and F. Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease," *Informatics in Medicine Unlocked*, vol. 24, p. 100606, 2021, doi: 10.1016/j.imu.2021.100606.

[7] U. Nagavelli, D. Samanta and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/7351061.

[8] S. Manimurugan et al., "Two-Stage Classification Model for the Prediction of Heart Disease Using IoMT and Artificial Intelligence," *Sensors*, vol. 22, no. 2, p. 476, Jan. 2022, doi: 10.3390/s22020476.

[9] M. Chowdhury et al., "Prevalence of cardiovascular disease among Bangladeshi adult population: a systematic review and meta-analysis of the studies," *Vascular Health and Risk Management*, vol. 14, pp. 165–181, Aug. 2018, doi: 10.2147/vhrm.s166111.

[10] "ResNet-50: The Basics and a Quick Tutorial," *Datagen*. [Online]. Available: https://datagen.tech/guides/computer-vision/resnet-50/.

[11] G. L. Team, “What is VGG16 - Convolutional Network for Classification and Detection,” *GreatLearning Blog: Free Resources what Matters to shape your Career!*, Oct. 01, 2021. [Online]. Available: https://www.mygreatlearning.com/blog/introduction-to-vgg16/.

[12] Y. Li and L. Liu, “Image quality classification algorithm based on InceptionV3 and SVM,” *MATEC Web of Conferences*, vol. 277, p. 02036, 2019, doi: 10.1051/matecconf/201927702036.

[13] “EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling,” *blog.research.google*, May 29, 2019. [Online]. Available: https://blog.research.google/2019/05/efficientnet-improving-accuracy-and.html.

[14] Q. Xiang, X. Wang, R. Li, G. Zhang, J. Lai and Q. Hu, “Fruit Image Classification Based on MobileNetV2 with Transfer Learning Technique,” in *Proceedings of the 3rd International Conference on Computer Science and Application Engineering - CSAE 2019*, 2019, doi: 10.1145/3331453.3361658.

[15] A. Khemphila and V. Boonjing, “Heart Disease Classification Using Neural Network and Feature Selection,” in *2011 21st International Conference on Systems Engineering*, Aug. 2011, doi: 10.1109/icseng.2011.80.

[16] J. A. and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, no. 1, p. 64, 2019, doi: 10.1186/s12874-019-0681-4.

[17] E. Carrizosa, B. Martin-Barragan and D. R. Morales, “Multi-group support vector machines with measurement costs: A biobjective approach,” *Discrete Applied Mathematics*, vol. 156, no. 6, pp. 950–966, Mar. 2008, doi: 10.1016/j.dam.2007.05.060.

[18] T. Maghsoudi, R. Cascón-Pereira and A. Beatriz Hernández Lara, “The Role of Collaborative Healthcare in Improving Social Sustainability: A Conceptual Framework,” *Sustainability*, vol. 12, no. 8, p. 3195, Apr. 2020, doi: 10.3390/su12083195.

[19] E. Vayena, A. Blasimme and I. G. Cohen, “Machine learning in medicine: Addressing ethical challenges,” *PLOS Medicine*, vol. 15, no. 11, p. e1002689, Nov. 2018, doi: 10.1371/journal.pmed.1002689.

[20] I. S. Hofer, M. Burns, S. Kendale and J. P. Wanderer, “Realistically Integrating Machine Learning Into Clinical Practice,” *Anesthesia & Analgesia*, vol. 130, no. 5, pp. 1115–1118, May 2020, doi: 10.1213/ane.0000000000004575.

[21] S. Amal, L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot and E. G. Ross, “Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care,” *Frontiers in Cardiovascular Medicine*, vol. 9, Apr. 2022, doi: 10.3389/fcvm.2022.840262.

[22] N. Tsalikidis, A. Mystakidis, C. Tjortjis, P. Koukaras and D. Ioannidis, "Energy load forecasting: one-step ahead hybrid model utilizing ensembling," *Computing*, Sep. 2023, doi: 10.1007/s00607-023-01217-2.

[23] B. B. Vimala, S. Srinivasan, S. K. Mathivanan, Mahalakshmi, P. Jayagopal and G. T. Dalu, "Detection and classification of brain tumor using hybrid deep learning models," *Scientific Reports*, vol. 13, no. 1, p. 23029, Dec. 2023, doi: 10.1038/s41598-023-50505-6.

[24] A. Younis, L. Qiang, C. O. Nyatega, M. J. Adamu and H. B. Kawuwa, "Brain Tumor Analysis Using Deep Learning and VGG-16 Ensembling Learning Approaches," *Applied Sciences*, vol. 12, no. 14, p. 7282, Jul. 2022, doi: 10.3390/app12147282.

[25] S. Gupta, "Classify any Object using pre-trained CNN Model," *Medium*, Jul. 17, 2020. [Online]. Available: https://towardsdatascience.com/classify-any-object-using-pre-trained-cnn-model-77437d61e05f.

[26] G. Meena, K. K. Mohbey and S. Kumar, "Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100174, Apr. 2023, doi: 10.1016/j.jjimei.2023.100174.

[27] A. Dertat, "Applied Deep Learning - Part 4: Convolutional Neural Networks," *Towards Data Science*, Nov. 08, 2017. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2.

[28] P. Huilgol, "Top 4 Pre-Trained Models for Image Classification | With Python Code," *Analytics Vidhya*, Aug. 17, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/08/top-4-pre-trained-models-for-image-classification-with-python-code/.

[29] Pradeepsingam, "Data Augmentation with small size datasets," *Recap of machine learning algorithms*, Jun. 14, 2023. [Online]. Available: https://medium.com/recap-of-machine-learning-algorithms/data-augmentation-with-small-size-datasets-1e09b996a012.

[30] S.-H. Tsang, "Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015," *Medium*, Mar. 23, 2019. [Online]. Available: https://sh-tsang.medium.com/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c.

[31] R. Yamashita, M. Nishio, R. K. G. Do and K. Togashi, "Convolutional Neural Networks: an Overview and Application in Radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Jun. 2018, doi: 10.1007/s13244-018-0639-9.

ORIGINALITY REPORT

24%	17%	16%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	www.mdpi.com Internet Source	1%
3	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications", River Publishers, 2024 Publication	1%
4	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	1%
5	ijarsct.co.in Internet Source	1%
6	rsisinternational.org Internet Source	1%
7	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International	1%