ASSISTED ARTIFICIAL INTELLIGENCE MEDICAL DIAGNOSIS SYSTEM FOR HEART DISEASE

BY

MWEEMBA MAAMBO

21000382

A DISSERTATION SUBMITTED IN PARTIAL FULFUILMENT OF THE REQUIREMENT OF A DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF ZAMBIA

SCHOOL OF NATURAL SCIENCES

LUSAKA

NOVEMBER 2023

COPYRIGHT

DECLARATION

I, Mweemba Maambo do hereby declare that this dissertation is my own original work and has not been submitted to any other college, institution, or university other than the University of Zambia.

Name: ……..

Sign: ……

Date: ……

APPROVAL

This dissertation, by Mweemba Maambo has been approved as partial fulfilment of the requirements for the award of Master of Science in Computer Science (Management Strategy by the University of Zambia.

Examiner 1

Name: …….

Signature: …….

Date: …….

Examiner 2

Name: …….

Signature: …….

Date: …….

Examiner 3

Name: …….

Signature: …….

Date: …….

Chairperson (Board of examiners)

Name: …….

Signature: …….

Date: …….

Supervisor

Name: …….

Signature: …….

Date: …….

ACKNOWLEDGEMENTS

This study would not have stood feasible minus the wealth of information and insights gathered from various researchers and sources who have delved extensively into the study of heart disease. The researcher acknowledges their valuable contributions to this endeavour. The University of Zambia (UNZA) provided support for this research, for which the researcher is grateful. Special appreciation is extended to Prof Jackson Phiri, the supervisor, whose patient review, constructive criticisms, and insightful suggestions helped maintain focus throughout the entire process. The researcher expresses deep gratitude to their parents for unwavering measurable and support, without which this study would not have been possible, as well as to their siblings and extended family for their encouragement , love, and prayers that provided strength at every stage.

Furthermore, heartfelt thanks are extended to friends whose discussions and ideas illuminated various aspects of the research journey. Lastly, the researcher acknowledges the grace of God, which sustained them intellectually and physically and provided the opportunity to see the research through to completion. To all others whose contributions may not have been explicitly mentioned, their support is recognized and appreciated.

DEDICATION

To my parents, Mr Milford Maambo and Mrs Florence Maambo, for their kindness and devotion, and for their endless support they rendered in seeing that I complete this research paper and obtain my masters' degree.

ABSTRACT

In recent years, the expanding array of innovative applications in the medical domain has been instrumental in propelling research forward. Among these advancements, Artificial Intelligence (AI) systems have emerged as influential tools, significantly contributing to the development of various medical applications and tools. However, heart disease remains a pressing health concern globally, highlighting the critical need for accurate diagnosis to enable effective regulation and intervention. This study focuses on harnessing the capabilities of an AI system to enhance the diagnostic process for heart disease. By leveraging input medical data sourced from a well-established dataset on Kaggle, the developed AI application is tailored specifically to cater to the demographic characteristics of Zambian patients. The primary objective is to evaluate the model's predictive accuracy when applied to medical data from the Zambian population. To facilitate this assessment, 80% of the collected dataset is allocated for training purposes, with the remaining 20% reserved for testing. Central to the prediction process is the utilization of a Bayesian data-mining algorithm, which plays a pivotal role in forecasting the risk level and likelihood of heart disease. An extensive array of medical parameters, including blood sugar levels, sex, heart rate, age, cholesterol levels, blood pressure, presence of exercise-induced angina, ST-slope, oldpeak, resting electrocardiogram results, and chest pain type, serves as the foundation for predicting heart disease in patients. Following a meticulous pre-processing phase, supervised learning techniques are employed to craft a robust prediction model. The outcomes of this process reveal a commendable prediction accuracy of 90.97%. Comparative analysis with established algorithms such as KNN, Random Forest, and Decision Tree algorithms further validates the efficacy of the proposed AI system in medical diagnosis. This study not only emphasizes the effectiveness of AI systems in medical diagnosis but also contributes valuable insights to the ongoing efforts to combat heart disease. The integration of data mining, artificial intelligence, and predictive modeling presents a promising avenue for advancing healthcare practices and outcomes, particularly in regions like Zambia grappling with cardiovascular health challenges. Achieving an 89% accuracy rate demonstrates the model's ability to adapt to Zambian patients' traits, while incorporating real-world data from the National Heart Hospital enhances its credibility. This validation underscores the potential of AI-driven diagnostic systems to improve healthcare and patient outcomes.

Keywords: Heart Disease, Artificial Intelligence, Bayesian Classification, Prediction Model, Supervised Learning Techniques, Data Mining Algorithms

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
GA	Genetic Algorithm
KNN	K-Nearest Neighbours Algorithm
MLP	Multi-Layer Perception
NB	Naïve Bayes
NHRA	National Health Research Authority
ROC	Receiver Operating Characteristic
SMO	Sequential Minimal Optimization
UCI	California University, Irvine
WEKA	Waikato Environment for Knowledge Analysis
WHF	world Health Federation

1 INTRODUCTION AND BACKGROUND

1.1 Introduction

This chapter serves as a foundational introduction to the research by offering a succinct background that delves into historical context and pertinent information. The historical perspective presented here contextualizes the evolution of the problem under investigation, providing a comprehensive understanding of its development over time. Following the background exposition, the problem statement is articulated to precisely define and outline the specific issues that the research endeavours to address. This section serves as a bridge between historical insights and the contemporary challenges, offering a clear perspective on the relevance and urgency of the identified problem.

The subsequent focus is on elucidating the aim of the study. This serves as a guiding beacon, pinpointing the overarching goal that the research seeks to achieve. The aim encapsulates the ultimate objective, guiding the entire trajectory of the study towards a specific outcome. In tandem with the aim, the research objectives are systematically presented. These objectives serve as targeted inquiries that structure the investigation and provide a framework for answering the research questions. Each objective is carefully designed to contribute to the overall understanding of the problem and collectively address the research questions posed at the outset.

The significance of the study is then expounded upon, shedding light on the real-world implications and potential contributions of the research. This section articulates who stands to benefit from the findings and insights generated by the study, thereby defining the broader impact and relevance of the research within the academic and practical spheres. To encapsulate the essence of this chapter, a summary is provided. This summary acts as a concise recapitulation of all discussions, encapsulating the background, problem statement, aim, objectives, and significance of the study. It serves as a concluding reflection, reinforcing the key aspects and setting the stage for the subsequent chapters of the research endeavour.

1.2 Background

Heart disease stands as a formidable health challenge in contemporary society, encapsulating a spectrum of medical illnesses that directly affect the heart and its different parts [1]. This pervasive health concern claims a staggering 17.5 million lives annually, ranking among the leading global causes of mortality. Projections indicate a further increase to 23 million deaths by 2030, underscoring the urgent need for effective interventions [1].

Within the Zambian context, heart disease emerges as a significant contributor to mortality, accounting for 10% of all deaths in individuals aged between 30 and 70, according to data compiled by the World Heart Federation's (WHF) CVD World Monitor [2]. This statistic emphasizes the criticality of accurate diagnosis and proactive measures to mitigate the impact of heart disease within the Zambian population.

Against this backdrop, the resurgence of Artificial Intelligence (AI) has become a transformative force in healthcare. The intersection of AI and health technologies has witnessed a surge in applications, including healthcare administration, predictive healthcare, and analysis of patient data, diagnostics, and clinical decision-making. Notably, AI has been increasingly deployed in addressing heart diseases, leveraging advanced computational techniques for enhanced diagnostic precision.

Prominent scholars, such as Mirzajani et al. [3], Kumar et al. [4], and Enriko et al. [5], have explored various computational intelligence methods for heart disease diagnosis. Techniques like Genetic Algorithm (GA) have been employed alongside classification algorithms such as KNN, j48 decision tree, SMO, and Naive Bayes (NB), demonstrating the versatility and efficacy of AI in predicting heart disease. In the Zambian context, where heart disease has exhibited a concerning upward trend, accounting for a 2% increase in death rates from 2017 to 2020 [2], the need for innovative approaches to diagnosis becomes imperative.

This study proposes a novel approach—an AI-assisted medical diagnosis system—designed to support healthcare practitioners in diagnosing heart disease. The intelligent diagnosis system utilizes a medical heart disease dataset sourced from Kaggle, serving as both testing and training data. By harnessing the power of AI and leveraging a robust dataset, this system aims to enhance diagnostic accuracy, ultimately aiding in the enhancement of patient outcomes and the management of healthcare related to heart disease in Zambia.

1.3 Statement of the Problem

In the face of technological progress, heart disease persists as a significant and pressing challenge in Zambia, a reality underscored by the considerable number of deaths attributed to this health condition. Despite strides in various domains of technology, a notable gap exists, particularly in the realm of software applications that could synergize with the medical practitioners efforts, thereby enhancing heart disease management and prediction. The gravity of this situation necessitates innovative solutions to augment existing healthcare practices and address the persistent threat posed by heart disease.

The apparent lack of advanced software tools capable of complementing the endeavours of medical professionals underscores the urgency for proactive measures to mitigate the impact of heart disease. Recognizing the potential of artificial intelligence techniques (AI), there is a compelling opportunity to leverage computer-assisted diagnosis systems to address this critical gap. By integrating AI methodologies, such systems have the potential to significantly contribute to the reduction of the death rate associated with heart disease.

The essence of employing AI in the context of heart disease lies in its capacity to offer decision support to medical practitioners. Through advanced algorithms and data analytics, these AI-driven systems can analyse complex medical data swiftly and accurately, facilitating early diagnosis and intervention. This proactive approach is pivotal in mitigating the severity of heart disease, as early detection allows for timely medical interventions and tailored treatment plans.

In essence, the implementation of computer-assisted diagnosis using AI techniques represents a transformative pathway towards enhancing the capabilities of healthcare professionals in Zambia. By providing nuanced decision support, these systems aim to bridge the existing gap in predicting and managing heart disease, ultimately contributing to a reduction in mortality rates and improved overall healthcare outcomes for individuals grappling with this pervasive health challenge.

1.4 Aim of the Study

The studies aim was to propose a Machine Learning system to assist medical practitioners in predicting heart disease in Zambia.

1.5 Research Objectives

i To develop a machine-learning designed to aid in the prediction of heart disease.
ii To utilize a machine learning classification model to predict heart disease.
iii To determine how accurate the model is at diagnosing heart disease on Zambian patients.

1.6 Research Questions

i How might we create a model for forecasting heart disease using machine learning classification?
ii How could we build a prototype utilizing the model from (i) to forecast heart disease using machine-learning classification?
iii How does the model perform at diagnosing heart disease on Zambian patients?

1.7 Significance of the Study

The study was of significance in that it aimed at assisting medical practitioners to detect heart disease in Zambia.

1.8 Scope of the study

This study was a comparative study of heart disease. This study attempted to tackle heart disease prediction based on the attributes recorded by medical specialists.

1.9 Structure of the Dissertation

The dissertation comprises five chapters, outlined as follows.

Chapter one initiates the dissertation by providing a comprehensive introduction, setting the stage for the subsequent exploration of the research. The background of the study is expounded, offering historical insights and contextual information that lays the groundwork for understanding the subject matter. Within this chapter, the problem statement is articulated, delineating the specific issues and challenges that the research endeavours to address. Furthermore, the aim and objectives of the study are clearly defined, serving as guiding beacons

for the entire research endeavour. Research questions, the scope of the study, and its significance are also intricately covered, establishing a robust foundation for the subsequent chapters.

Chapter Two delves into an extensive review of existing literature conducted by diverse scholars on the subject matter. This chapter meticulously identifies key findings, thematic trends, and crucial gaps in the current body of knowledge, providing a comprehensive overview of the scholarly landscape related to the research topic. By synthesizing and critically evaluating existing literature, Chapter Two sets the stage for the original contributions and insights that the current research aims to make.

Chapter Three strategically outlines the methodology employed to conduct the study. It delves into discussions on research design, population selection, data collection methods, techniques, and analysis. This chapter also elaborates on the proposed research method, formulates hypotheses, and delves into ethical considerations that underpin the research approach. By transparently detailing the methodology, Chapter Three offers readers a clear understanding of how the research was conducted and how the data was collected and analysed.

Chapter Four is dedicated to the analysis of the collected data. This pivotal chapter not only presents the results but also rigorously tests the hypotheses formulated in Chapter Three. It goes beyond mere presentation by interpreting, discussing, and drawing conclusions from the research findings. This chapter serves as the crucible where raw data transforms into meaningful insights, providing a comprehensive understanding of the implications and significance of the study.

Chapter Five serves as the culmination of the research journey, answering the study questions posited in Chapter One. This chapter synthesizes the accumulated knowledge, offering conclusions and recommendations based on the empirical findings of the study. By aligning the research outcomes with the initially stated objectives, Chapter Five provides a valuable synthesis and a forward-looking perspective, contributing to both academic scholarship and practical applications in the field.

1.10 Chapter Summary

Within this chapter, a comprehensive background has been provided, shedding light on the critical necessity for an AI-powered medical diagnostic assistance system tailored for heart disease. This contextualization serves to underscore the gravity of the existing challenges and gaps in the current healthcare landscape, particularly in the realm of heart disease diagnosis and management.

The problem statement articulated in this chapter further delineates the intricacies and nuances of the identified issues. By clearly articulating the challenges faced in the current medical diagnostic paradigm, the chapter sets the stage for the exploration of innovative solutions, emphasizing the role of artificial intelligence as a potential game-changer in addressing these challenges.

The overarching aim of the study has been highlighted, offering a clear beacon that guides the entire research endeavour. This aim encapsulates the ultimate goal, emphasizing the transformative potential of an AI-powered medical diagnostic assistance system in the realm of heart disease. It acts as the lodestar that directs the research towards a tangible and impactful outcome.

To operationalize this aim, specific objectives have been delineated. These objectives serve as targeted inquiries, systematically guiding the research towards answering key research questions. By breaking down the overarching aim into measurable and achievable components, the objectives provide a roadmap for the systematic investigation of the identified problem.

The significance of the study has been elucidated within this chapter, aiming to define the beneficiaries of the research outcomes. By clearly articulating the potential impact and relevance of the study, the chapter establishes the broader implications for various stakeholders. These stakeholders may include healthcare practitioners, researchers, policymakers, and, most importantly, individuals at risk or affected by heart disease. The significance of the study thus extends beyond academic realms, offering practical insights and potential advancements for the betterment of healthcare practices and patient outcomes.

In essence, this chapter serves as the bedrock upon which the subsequent research unfolds. It not only lays the groundwork by providing a thorough background and problem statement but also articulates a clear aim, objectives, and significance, framing the research within a context

that emphasizes its relevance and potential contributions to the field of artificial intelligence in medical diagnosis, specifically for heart disease.

2 LITERATURE REVIEW

2.1 Introduction

This chapter serves as a comprehensive exploration into the realm of utilizing machine learning for heart disease prediction, aligning closely with the central focus and objectives outlined in Chapter One. The core aim here is to unravel the intricacies of machine learning applications, particularly within the context of heart disease prediction, by delving into the existing body of literature that informs and shapes this domain.

The chapter begins by dissecting the landscape of machine learning in the context of heart disease prediction. Various machine learning classification algorithms are scrutinized and analysed for their efficacy in predicting heart disease. This involves a meticulous examination of the underlying principles and methodologies employed by these algorithms, ranging from decision trees and support vector machines to neural networks and ensemble methods. By comprehensively understanding the intricacies of each algorithm, the chapter aims to establish a nuanced perspective on the diverse approaches employed in the field.

In tandem with explaining the intricacies of machine learning algorithms, the chapter synthesizes and presents major findings from previous studies in this domain. This involves a critical examination of research outcomes, highlighting successful applications, novel insights, and advancements achieved through the implementation of machine learning for heart disease prediction. The aim is to distil key knowledge from the existing literature, identifying trends and patterns that contribute to the overall understanding of the subject matter.

Simultaneously, the chapter meticulously identifies gaps within the existing body of literature. These gaps may manifest as unexplored avenues, methodological limitations, or areas where further research is warranted. By pinpointing these gaps, the chapter not only contributes to the scholarly discourse but also sets the stage for the current study to make meaningful contributions and advancements.

As the chapter draws to a close, a comprehensive summary encapsulates the key insights gleaned from the literature review. This summary serves as a concise synthesis of the major findings, existing gaps, and overarching trends identified in the literature. It acts as a crucial bridge, connecting the literature review to the subsequent chapters, providing a foundation upon which the original contributions and insights of the current study can be built. In essence,

this chapter lays the groundwork for a deeper understanding of the intricacies of machine learning in heart disease prediction and paves the way for the subsequent phases of the research.

2.2 Background to the Study

Accurate diagnosis is paramount for determining the nature and severity of heart disease, enabling timely and effective interventions. The integration of intelligent systems offers valuable support in the diagnostic process. As highlighted by Wiharto, Herianto, and Kusnanto [6], the implementation of systems aiding clinical decision-making has the potential to enhance medical practitioners' clinical practice, diminishing the likelihood of erroneous diagnoses. In this context, the development of an intelligent diagnosis system becomes imperative, leveraging existing health data, particularly datasets intricately linked to heart disease, as the foundational training data.

The importance of precise diagnosis cannot be overstated, as it forms the bedrock for informed decision-making in healthcare. An intelligent diagnosis system serves as a potent ally for medical professionals, providing them with nuanced insights and evidence-based recommendations. Drawing upon the wealth of existing health data, the system gains the capacity to discern patterns, correlations, and anomalies associated with heart disease, thereby contributing to more accurate and efficient diagnostic outcomes.

Clinical decision support systems, as elucidated by Wiharto, Kusnanto, and Herianto [6], have the potential not only to elevate the quality of clinical practice but also to mitigate the risks associated with diagnostic errors. By harnessing the power of artificial intelligence and data-driven methodologies, an intelligent diagnosis system can continuously learn and adapt, staying abreast of emerging trends and refining its diagnostic capabilities over time.

Crucially, the development of such a system necessitates access to comprehensive and relevant health data, specifically tailored to heart disease. This data serves as the backbone of the training process, enabling the intelligent diagnosis system to discern subtle nuances and variations within the realm of cardiac health. As technology continues to evolve, the integration of intelligent systems in healthcare not only augments diagnostic accuracy but also represents a paradigm shift toward personalized and data-driven medicine, ultimately contributing to improved patient outcomes and overall healthcare efficacy.

2.3 Types of AI Technology

2.3.1 Machine Learning Algorithms

Supervised learning is a type of machine learning where the algorithm is trained on a labelled dataset, meaning that it learns from input-output pairs. In the context of healthcare, particularly in the classification of heart diseases, supervised learning algorithms have demonstrated significant success. Two notable algorithms in this domain are Support Vector Machines (SVM) and Decision Trees [7].

1. **Support Vector Machines (SVM):**

SVM is a powerful supervised learning algorithm used for classification and regression tasks. In the context of heart disease classification, SVM works by finding the optimal hyperplane that separates different classes based on features extracted from historical patient data.

Historical patient data includes a variety of parameters such as age, blood pressure, cholesterol levels, and other relevant medical indicators. SVM learns to draw boundaries between different classes (e.g., diseased and non-diseased) by maximizing the margin between data points of different classes.

SVM is particularly effective when dealing with complex, high-dimensional datasets, making it suitable for tasks where the relationships between features are not immediately apparent.

2. **Decision Trees:**

Decision trees are another class of supervised learning algorithms that can be employed in the classification of heart diseases. They operate by iteratively dividing the dataset according to various features, forming a tree-like arrangement where each node signifies a decision based on a distinct feature.

In the context of heart disease classification, a decision tree might ask questions such as "Is the patient's age above a certain threshold?" or "Does the patient have a history of high blood pressure?" to determine the likelihood of heart disease.

Decision trees are interpretable, providing insights into the decision-making process. They are especially useful in healthcare settings where transparency in decision-making is crucial for gaining trust from healthcare professionals and patients alike.

3. **Leveraging Historical Patient Data:**

Both SVM and decision trees rely on historical patient data to train and fine-tune their models. This data typically includes a diverse range of information such as demographic details, medical history, diagnostic test results, and lifestyle factors.

By analyzing this historical data, the algorithms learn to recognize patterns associated with different heart conditions. For instance, they may identify combinations of risk factors that are indicative of a higher probability of heart disease.

The use of large and diverse datasets is essential to ensure that the algorithms generalize well to different populations and account for variations in individual patient characteristics.

In summary, supervised learning algorithms like SVM and decision trees contribute significantly to the accurate classification of heart diseases by leveraging the wealth of information present in historical patient data. These algorithms play a crucial role in assisting healthcare professionals in making informed decisions about diagnosis and treatment strategies, ultimately improving patient outcomes.

2.3.2 Deep Learning Algorithms

Deep learning, a subset of machine learning, has emerged as a transformative force in healthcare, especially in the analysis of medical data. Two key architectures within deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have garnered significant attention for their efficacy in handling specific types of healthcare data [8].

Convolutional Neural Networks (CNNs) for Image-Based Diagnostics:

CNNs are a class of deep neural networks particularly well-suited for tasks involving image analysis. In healthcare, CNNs have revolutionized medical imaging by excelling in tasks such as image classification, segmentation, and detection.

In the context of heart disease and other medical conditions, CNNs can be applied to various imaging modalities such as X-rays, MRIs, and CT scans. These networks automatically learn hierarchical representations of features, enabling them to discern patterns and abnormalities that may not be immediately apparent to the human eye.

CNNs are capable of identifying subtle details in medical images, aiding in the early detection and accurate diagnosis of cardiovascular issues. For example, they can highlight specific regions of interest in an echocardiogram or identify anomalies in a coronary angiogram.

Recurrent Neural Networks (RNNs) for Sequential Data Processing:

RNNs are designed to handle sequential data and are particularly effective in tasks where the order of information is crucial. In healthcare, RNNs find application in processing time-series data, such as electronic health records (EHRs) that contain a chronological sequence of a patient's medical history.

When applied to patient records, RNNs can capture temporal dependencies and trends, allowing for more accurate predictions or classifications. This is especially valuable in monitoring chronic conditions, tracking disease progression, and predicting potential cardiac events based on historical patient data.

RNNs can be used to model the dynamic nature of physiological signals like electrocardiogram (ECG) data, where changes over time may indicate evolving cardiac conditions. This capability makes RNNs instrumental in predicting and preventing adverse events.

Integration of CNNs and RNNs for Comprehensive Analysis:

In some healthcare applications, a combination of CNNs and RNNs is employed to analyze both imaging and sequential data comprehensively. For example, in cardiology, this integration allows for a holistic approach where imaging data (e.g., from cardiac imaging modalities) and temporal data (e.g., EHRs) can be jointly considered for more accurate diagnostics and personalized treatment planning.

The joint use of CNNs and RNNs enhances the ability to uncover complex relationships between visual patterns in medical images and the temporal evolution of a patient's health status.

In conclusion, deep learning, with its CNNs and RNNs, has revolutionized healthcare by providing advanced tools for image-based diagnostics and sequential data processing. These technologies contribute to more accurate and timely diagnoses, enabling healthcare

professionals to intervene proactively and improve patient outcomes in the realm of cardiovascular health and beyond.

2.4 Applications of AI in Heart Disease Diagnosis:

AI applications extend across various facets of heart disease diagnosis:

Risk Assessment:

AI models integrate clinical data, genetic information, and lifestyle factors to assess an individual's risk of developing heart diseases [9].

Early Detection:

AI-assisted medical imaging systems contribute to early detection through the analysis of electrocardiograms (ECGs), cardiac MRI, and other diagnostic images [10].

Decision Support for Clinicians:

AI systems provide clinicians with decision support tools, aiding in accurate diagnosis and personalized treatment plans [11].

2.5 Datasets and Data Preprocessing:

To overcome challenges related to data quality and diversity, researchers curate carefully annotated datasets. Data preprocessing techniques, such as normalization and feature extraction, are applied to optimize model performance [12].

2.6 Performance Evaluation Metrics:

Metrics like sensitivity, specificity, precision, and the area under the receiver operating characteristic (ROC) curve are commonly employed to evaluate the performance of AI models in heart disease diagnosis [13].

2.7 Ethical and Regulatory Considerations:

The integration of AI in medical diagnosis brings forth ethical considerations, including patient privacy, transparency of algorithms, and bias mitigation. Adherence to healthcare regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), is imperative [14].

2.8 Challenges and Future Directions:

Despite successes, challenges persist:

Interpretability:

The interpretability of AI models remains a challenge, with efforts focused on developing transparent models that can be understood and trusted by clinicians [15].

Integration into Clinical Workflows:

Successful integration of AI systems into clinical workflows is crucial for widespread adoption. This involves collaboration between AI developers and healthcare professionals [16].

Addressing Biases:

Efforts are ongoing to address biases in AI algorithms to ensure equitable and unbiased diagnoses across diverse patient populations [17].

2.9 Case Studies and Success Stories

Several real-world implementations have demonstrated the effectiveness of AI-assisted diagnosis:

Automated ECG Analysis:

The integration of AI algorithms into automated electrocardiogram (ECG) analysis marks a significant breakthrough in the realm of cardiovascular diagnostics. These algorithms exhibit remarkable accuracy in the detection of arrhythmias and abnormalities, showcasing their potential to revolutionize the way we interpret and act upon ECG data [18].

AI-driven ECG analysis systems excel in swiftly and precisely identifying irregularities in heart rhythm. By processing ECG signals with intricate patterns, these algorithms contribute to early and accurate diagnoses of various cardiac conditions, including atrial fibrillation, ventricular tachycardia, and other rhythm disturbances.

The automation of ECG analysis not only enhances the efficiency of healthcare workflows but also provides a valuable tool for continuous monitoring. This enables timely intervention in cases where immediate attention is crucial, contributing to the proactive management of cardiovascular health.

Cardiac Imaging:

Deep learning models applied to cardiac imaging represent a cutting-edge application of artificial intelligence in cardiovascular healthcare. These models, particularly utilizing techniques like convolutional neural networks (CNNs), have demonstrated superior performance in deciphering intricate details within cardiac images, thereby enhancing diagnostic accuracy [19].

The application of deep learning in cardiac imaging extends beyond mere identification to the prediction of cardiovascular events. These models can analyze diverse imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT) scans, and echocardiograms to identify structural abnormalities indicative of potential future cardiovascular issues.

The ability of deep learning models to unravel complex patterns in cardiac images contributes to early and more accurate diagnoses, enabling healthcare professionals to intervene at the earliest stages of disease progression. This, in turn, facilitates more effective treatment strategies and better outcomes for patients.

Clinical Decision Support Systems:

AI-driven clinical decision support systems play a pivotal role in augmenting the decision-making capabilities of healthcare professionals. These systems harness advanced algorithms to analyze vast amounts of clinical data, providing timely, evidence-based recommendations that enhance diagnostic accuracy and treatment planning [20].

In the field of cardiology, where the volume of medical knowledge is vast and continually evolving, AI-driven decision support systems become valuable tools for clinicians. They assist in synthesizing complex information, considering patient-specific factors, and delivering actionable insights that contribute to more accurate and timely diagnoses.

The seamless integration of AI into clinical decision-making processes supports healthcare providers in formulating personalized treatment plans, reducing errors, and improving overall patient outcomes. By leveraging these systems, clinicians can navigate the intricacies of cardiovascular medicine with greater confidence and efficiency.

In summary, the incorporation of AI technologies in automated ECG analysis, cardiac imaging, and clinical decision support systems represents a transformative era in cardiovascular healthcare. These advancements empower healthcare professionals with tools that enhance diagnostic accuracy, provide early insights, and contribute to more effective and personalized patient care, ultimately leading to improved outcomes and a more robust healthcare ecosystem.

2.10 Challenges in Real-world Implementation:

It's essential to acknowledge the limitations and challenges associated with AI in medical diagnosis. Studies explore issues such as over-reliance on algorithms, potential misinterpretations, and the need for continuous validation in real-world settings [21].

Despite the promising developments and potential benefits that AI brings to cardiology, the practical implementation of these technologies in real-world healthcare settings encounters several challenges. Addressing these challenges is crucial to ensure the successful deployment and integration of AI systems into cardiology practices [22][23][24][25]. Some of the key challenges include:

Integration into Electronic Health Records (EHR):

One significant challenge is seamlessly integrating AI systems into existing Electronic Health Records (EHR) infrastructure. Many healthcare institutions use diverse EHR systems, and AI applications need to harmonize with these platforms to facilitate smooth data exchange and interoperability.

Ensuring that AI-generated insights are accessible within the clinician's workflow and are well-integrated with patient records is essential for enhancing the efficiency and effectiveness of cardiovascular care.

Data Security and Privacy Concerns:

Healthcare data, particularly in cardiology, is highly sensitive, containing detailed information about patients' cardiovascular health. Maintaining robust data security and privacy is paramount to gaining trust from both healthcare providers and patients.

AI systems need to adhere to stringent data protection standards, encrypting and securing patient information throughout the entire data lifecycle. Compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) is crucial for safeguarding patient confidentiality.

Ensuring Regulatory Compliance:

The deployment of AI in cardiology must align with regulatory frameworks and standards governing the use of medical technologies. Obtaining approvals from health regulatory bodies ensures that AI applications meet the necessary safety and efficacy criteria.

Compliance with regulations, such as FDA approval for medical devices, is essential to guarantee that AI systems adhere to established quality and safety standards. This step is critical in gaining acceptance from healthcare professionals and regulatory authorities.

Interpretable and Explainable AI Models:

AI models often operate as complex, "black-box" systems, making it challenging for clinicians to interpret the reasoning behind specific decisions. In cardiology, where precise and transparent decision-making is crucial, developing interpretable and explainable AI models is essential.

Ensuring that clinicians can understand and trust the AI-generated insights fosters collaboration between human practitioners and AI systems. This transparency is particularly important in critical situations, where the ability to comprehend and validate AI recommendations is paramount.

Data Quality and Bias Mitigation:

AI models heavily depend on the quality and representativeness of the data they are trained on. In cardiology, addressing issues such as data bias and ensuring the inclusion of diverse patient populations is crucial for creating fair and unbiased AI algorithms.

Rigorous data quality assurance processes, along with continuous monitoring for biases, are essential to prevent AI systems from perpetuating disparities in healthcare outcomes.

User Training and Acceptance:

Integrating AI into cardiology practices requires healthcare professionals to adapt to new tools and workflows. Adequate training and education programs are essential to ensure that clinicians understand how to effectively use AI-generated insights in their decision-making processes.

Gaining acceptance and trust from healthcare providers is vital for the successful adoption of AI technologies in cardiology. Collaboration between AI developers and healthcare professionals can bridge the gap and facilitate a smoother transition to AI-assisted healthcare.

Overcoming these challenges demands a collaborative effort involving healthcare providers, AI developers, regulatory bodies, and other stakeholders. By addressing issues related to integration, security, regulation, interpretability, data quality, and user acceptance, the healthcare industry can unlock the full potential of AI in cardiology, leading to improved patient outcomes and more efficient healthcare delivery.

2.11 Related Works

Mirzajani and Siamak [3] undertook a comprehensive study involving the diagnosis and prediction of heart disease, utilizing the potent WEKA data mining tool to apply advanced algorithms for data mining. In this exploratory endeavour, various classification algorithms, namely the SMO, j48 decision tree, KNN, and Naive Bayes (NB), were implemented on a dataset for heart disease. The dataset, comprising 209 records and encompassing 8 distinct features, was meticulously gathered from a hospital in Iran under the supervision of the Ministry of Health. A detailed breakdown of the features in the dataset is provided in Figure 1.

The study aimed to unravel and compare the efficacy of the selected classification algorithms in predicting and diagnosing heart disease. WEKA, renowned for its versatility and robust capabilities in handling diverse data mining tasks, served as the instrumental platform for executing these algorithms. By leveraging this sophisticated tool, Mirzajani and Siamak were able to analyse the experimental results derived from the application of the SMO, j48 decision tree, KNN, and Naive Bayes algorithms on the heart disease dataset.

The heart disease dataset, sourced from a reputable hospital in Iran, is a crucial component of this research, representing a real-world collection of health records under the meticulous control of the health ministry. The dataset comprises 8 features, each providing distinct insights into the factors influencing heart disease. The meticulous description of these features, as illustrated in Table 1, adds a layer of transparency to the research, enabling readers and fellow researchers to comprehend the variables and parameters under scrutiny.

Through this research endeavour, Mirzajani and Siamak not only contribute valuable insights into the application of data mining algorithms for heart disease prediction but also furnish the scientific community with a benchmark dataset sourced from a credible healthcare setting. This dataset, coupled with the outcomes of the analysis and comparison of classification algorithms, becomes a valuable resource for advancing the field of predictive medicine and refining diagnostic tools for heart disease.

Table 1: Dataset Description

#	NAME	POSSIBLE VALUES
1	Age	Numeric
2	Chest Pain Type	Asympt, Atyp_Angina, Non_Angina, Typ_Angina
3	Rest_Bpressure	Numeric
4	Blood Sugar	True, False
5	Rest_Electro	Numeric
6	Max_Heart_Rate	Normal, Left_vent_hyper, st_t_wave_abnormaility
7	Exercise_Angina	Yes, No

Given that the data for this study originated from a single source, the need for integration operations was obviated, streamlining the analytical process. The dataset, comprising 209 samples, demonstrated uniformity as all features across these samples contained values,

mitigating any concerns related to missing data. This meticulous attention to data integrity sets a robust foundation for the ensuing comparative analysis.

Upon conducting a thorough comparison of the selected classification algorithms, distinct performance metrics emerged to gauge their efficacy. The j48 decision tree algorithm demonstrated the highest classification accuracy, registering an impressive 83.73%. Following closely were KNN and SMO, both achieving a commendable accuracy rate of 82.78%, thereby securing the second position. Naive Bayes (NB) trailed slightly behind, yet still demonstrated a respectable accuracy of 81.82%.

While accuracy stands as the most commonly cited metric in classification performance, this study adopts a holistic approach by considering a spectrum of other essential performance measures. Precision, Sensitivity, ROC, F-Measure, and specificity indicators were meticulously factored into the evaluation process. These metrics provide a nuanced understanding of the algorithms' efficiency in not only correctly classifying instances but also in distinguishing between true positives, true negatives, false positives, and false negatives.

By delving into these diverse performance measures, this research aims to present a comprehensive assessment of the classification efficiency of the four selected algorithms. Each metric contributes unique insights into the algorithms' strengths and limitations, fostering a more nuanced interpretation of their predictive capabilities. In doing so, this study not only advances our understanding of heart disease prediction but also provides a methodological benchmark for future research endeavours in the realm of data-driven healthcare analytics.

In a parallel investigation, Enriko [5] ventured into heart disease prediction utilizing KNN, Naive Bayes, and Decision Tree algorithms, albeit employing a distinct dataset sourced from the venerable California University, Irvine (UCI). The dataset, meticulously selected, comprised 76 parameters, from which a subset of 10 parameters was employed for the comprehensive analysis. Notably, the study introduced a database software component, MongoDB, to facilitate the storage and retrieval of pertinent data.

Unlike the previous research, the analysis in this study hinged on a diverse array of parameters, with only 10 chosen out of the 76 available. This strategic selection aimed to distil the most pertinent features for accurate heart disease prediction. The implementation of KNN, Naive Bayes, and Decision Tree algorithms on this curated dataset unfolded a comparative analysis of their predictive capacities.

Remarkably, the accuracy results obtained from this study did not exhibit substantial disparities among the algorithms, suggesting a degree of robustness in their predictive capabilities. However, within the context of the research's specific parameters, KNN emerged as the frontrunner, yielding the highest accuracy at an impressive 81.85%. This outcome underscores the effectiveness of KNN in discerning patterns within the selected subset of parameters for heart disease prediction in the given dataset.

The use of MongoDB as the database software in this research signifies a strategic approach to data management, emphasizing scalability and flexibility in handling diverse parameters. This methodological choice not only adds a layer of sophistication to the study but also contributes to the broader discourse on leveraging advanced database technologies in healthcare analytics.

Enriko's research, therefore, not only augments the understanding of heart disease prediction through machine learning but also highlights the importance of dataset selection and algorithmic choice in influencing the accuracy of predictive models. By drawing comparisons with previous studies and introducing novel elements like MongoDB, this research enriches the evolving landscape of machine learning applications in healthcare analytics, propelling the field towards enhanced diagnostic precision and improved patient outcomes.

As delineated by Kumar, Anand et al. [4], their study delves into the application of Genetic Algorithm (GA) techniques to create computational intellect techniques for heart disease diagnosis. This innovative approach signifies a departure from traditional methodologies, offering a nuanced perspective on enhancing the accuracy and efficiency of diagnostic models. The researchers employed a 3-fold cross-validation approach to rigorously validate the performance of their model, a methodology designed to ensure robustness and generalizability.

The study sourced its input heart disease dataset from the UCIML repository, a repository renowned for its diverse and comprehensive datasets that serve as benchmarks for various machine learning applications. This deliberate choice of dataset reflects the researchers' commitment to utilizing real-world and well-curated data for their investigation.

Crucially, the research assessed the accuracy of the model under different configurations, utilizing various initial rule counts. The outcomes were obtained through the 3-fold cross-validation process. With an initial rule count of 25, the average accuracy stood at an admirable 81.83%. Notably, as the initial rule count increased to 50, there was a marked improvement in accuracy, reaching an impressive 86.83%. This finding underscores the sensitivity of the

model's performance to the configuration of initial rules, showcasing the potential impact of parameter tuning on diagnostic accuracy.

This study, therefore, not only contributes to the burgeoning field of heart disease diagnosis through computational intelligence but also sheds light on the significance of parameter optimization in fine-tuning model performance. The utilization of a 3-fold cross-validation approach adds a layer of robustness to the findings, ensuring that the model's predictive capabilities are rigorously evaluated across different subsets of the dataset.

In essence, Kumar, Anand et al.'s [4] research represents a valuable addition to the landscape of heart disease diagnosis methodologies, emphasizing the potential of Genetic Algorithm techniques and the importance of thoughtful dataset selection and parameter configuration for enhancing the accuracy of computational intelligence models in healthcare applications.

Zagorecki, Orzechowski, and Hołownia [26] embarked on the contraction of a pioneering web-based software system designed to cater to medical diagnosis needs. At the heart of this ground breaking system is an advanced distributed, parallel architecture comprising multiple Bayesian Networks (BN) engines. These engines operate autonomously, conducting queries to individual BNs, and are grounded in the SMILE, a versatile Bayesian Network software accessible at http://genie.sis.pitt.edu. A noteworthy aspect of their design is the stateless nature of the BN engines, a deliberate choice made to optimize scalability and enhance system reliability.

During the software implementation phase, the researchers conducted a meticulous analysis based on an extensive dataset, encompassing over 97,000 diagnoses generated in the initial weeks following the system's deployment. A predominant demographic within this dataset was users aged 25-39, comprising 52.1% females and 47.9% males. The prevalent symptoms reported by this age group included depression, tiredness, tension headaches, and anxiety disorders—common manifestations associated with the demands of modern lifestyles.

Interestingly, the chosen symptoms for analysis were often linked to specific anatomical locations, with a noteworthy 6.8% pertaining to the head, 4.4% to the genitals, and 3.9% to the lower abdomen. Intriguingly, indications related to the anus were reported in 2.2% of cases, a frequency comparable to more conventional symptoms such as chest pain, sleepiness, or tiredness. While lacking concrete evidence, this observation suggests that the system may frequently be employed for self-diagnosing issues associated with sexual health and additional ailments that individuals might find uncomfortable to discuss with a healthcare professional.

Furthermore, the researchers delved into the most common diagnoses for individuals aged 55-70 and 70+, revealing a consistent pattern of age-related health concerns. Diagnoses in these older age groups predominantly centered on issues such as gallstones, joint or bone trauma, osteoarthritis, and ischemic heart disease.

Zagorecki, Orzechowski, and Hołownia [26] ground breaking study not only introduces a cutting-edge web-based medical diagnosis system but also unveils insightful patterns and tendencies within the user demographic and diagnostic outcomes. The utilization of Bayesian Networks and the emphasis on scalability and reliability positions their system as a promising avenue for the intersection of technology and healthcare, paving the way for enhanced self-diagnosis capabilities and informed healthcare decisions.

In a comparative analysis of classification techniques, Repaka, Anjan Nikhil Ravikanti et al. [27] conducted a study that pitted the proposed method against established techniques such as Bayes Net, MLP (Multi-Layer Perception), and SMO (Sequential Minimal Optimization). This research sought to evaluate the performance of these classification methods and determine their efficacy in handling diverse datasets.

The proposed technique, labelled Navies Bayesian, emerged as a standout performer, showcasing an impressive accuracy rate of 89.77%. This marked superiority over the established techniques underscores the effectiveness and potential innovation embedded in the Navies Bayesian classification approach. The comparison not only highlights the prowess of the proposed method but also positions it as a promising alternative for achieving high accuracy in classification tasks.

The utilization of prevailing techniques like SMO, Bayes Net, and MLP as benchmarks in the evaluation process lends credibility to the findings. By contextualizing the proposed method within the landscape of established classification techniques, the research provides a valuable point of reference for practitioners and researchers seeking optimal solutions for their classification needs.

The noteworthy outcome of 89.77% accuracy achieved by the Navies Bayesian method is indicative of its robustness and adaptability across diverse datasets. This superior performance has implications for a wide range of applications, from medical diagnostics to financial forecasting, where accurate classification is paramount.

Repaka, Anjan Nikhil Ravikanti et al.'s [27] research, therefore, not only contributes to the ongoing discourse on classification techniques but also introduces a potentially transformative method in the form of Navies Bayesian. This method, with its demonstrated high performance, holds promise for enhancing the accuracy and reliability of classification tasks, thereby opening avenues for improved decision-making across various domains.

Wang et al. [28] have made a substantial contribution to the field of heart disease diagnosis by advocating for a deep learning-based approach. Their methodology revolves around the application of Convolutional Neural Networks (CNNs) on medical imaging data, specifically cardiac magnetic resonance images and computed tomography scans. The results of their study demonstrated exceptional accuracy in identifying structural abnormalities and cardiac anomalies. This utilization of artificial intelligence (AI) in medical imaging not only signifies a breakthrough in diagnostic accuracy but also holds the promise of early and precise detection of heart diseases.

Li et al. [29] have undertaken another significant study, focusing on the integration of wearable devices and continuous monitoring for heart disease diagnosis. By employing machine learning algorithms to analyse data collected from wearable sensors, encompassing heart rate variability, electrocardiogram signals, and physical activity patterns, their research showcased the potential of continuous monitoring through wearable technology. This approach, emphasizing proactive healthcare, provides a promising avenue for predicting and diagnosing heart conditions, thus potentially revolutionizing how cardiovascular health is managed.

Addressing the critical need for interpretable AI models in medical diagnosis, Ribeiro et al. [30] and Rahim et al. [31] have introduced frameworks that combine machine learning predictions with rule-based models. The integration of decision trees and rule-based systems aims to enhance the transparency and interpretability of AI-assisted diagnosis processes. This innovative approach not only contributes to the technical aspect but also addresses the crucial factor of building trust between healthcare professionals and AI systems, thereby fostering collaborative decision-making in the realm of cardiovascular healthcare.

In the study conducted by Ioannidis in 2016, the application of Random Forests in predicting the risk of heart disease stands out as a notable achievement [32]. This ensemble learning algorithm effectively harnesses electronic health record (EHR) data to generate accurate risk assessments. By considering a myriad of variables and their interactions, Random Forests

exhibit robustness and generalizability, making them valuable tools for identifying individuals at an elevated risk of cardiovascular conditions.

The study by Ioannidis emphasizes the importance of leveraging both genomic and clinical measures to enhance the predictive accuracy of the model, showcasing the potential synergy between biological and clinical data in the era of precision medicine [32].

Gudadhe et al.'s research in 2010 introduces Support Vector Machines (SVM) as a powerful tool for heart disease diagnosis [33]. SVMs excel in classifying patients with high accuracy, leveraging deep learning techniques for effective pattern recognition in diagnostic tasks. The application of SVMs in heart disease diagnosis highlights their ability to discern complex relationships within patient data. This approach contributes to the development of robust diagnostic tools that can aid healthcare professionals in accurately classifying individuals based on their cardiac health status.

Attia et al.'s work in 2019 introduces a neural network-based approach for risk stratification in heart disease, demonstrating the potential of deep learning techniques in this domain [9]. The study focuses on the identification of patients with atrial fibrillation during sinus rhythm, showcasing the specificity of the neural network in outcome prediction.

Neural networks, with their capacity to learn intricate patterns from large datasets, provide a sophisticated means of risk stratification. The study underscores the importance of leveraging artificial intelligence to enhance the accuracy of identifying specific cardiac conditions, paving the way for more targeted interventions and personalized patient care.

Furthermore, the endeavours of Chen et al. [34] and Byrd et al. [35] have explored the integration of natural language processing (NLP) techniques in the analysis of electronic health records (EHRs) for heart disease diagnosis. By extracting valuable information from unstructured clinical narratives, their systems have demonstrated improved accuracy in diagnostic predictions. This research emphasizes the significance of tapping into diverse data sources and leveraging advanced technologies, such as NLP, to enhance the comprehensiveness of AI-assisted diagnosis systems.

Bardhwaj et al. [36], Shailaja et al. [37], Sun et al. [38], and Lee & Yoon [39] extensively explored various machine learning techniques applied in healthcare across a range of diseases. Their investigation delved into the considerable potential of medical big data, demonstrating

its value in clinical decision support, diagnostics, treatment decisions, fraud detection, and prevention. They succinctly outlined a nine-step data mining process, underlining the critical importance of effective decision support in healthcare. Their findings underscored the capability of machine learning models to facilitate early disease diagnosis. While their work aligns with this project's context, it covers a broader spectrum and is less focused on diagnosing heart diseases. Thus, a shift towards literature more closely aligned with the project's objective, specifically investigating how machine learning algorithms can aid in diagnosing heart disease, is necessary.

One such broad analysis by Tripoliti et al. [40] focused on machine learning techniques customized for assessing heart failure, this study specifically investigated the severity estimation of heart failure and the forecasting of critical outcomes such as re-hospitalization, mortality, and destabilizations. Tripoliti et al. thoroughly analysed relevant literature on heart failure, offering valuable perspectives on the utilization of machine learning in addressing cardiac ailments.

In a distinct study by J. & S. [41], two supervised classifiers, namely Naïve Bayes Classifier and Decision Tree Classifiers, were employed to predict heart diseases using a dataset. The Decision Tree model demonstrated an impressive accuracy of 91 percent in predicting heart disease patients, while the Naïve Bayes Classifier achieved an accuracy of 87 percent.

Another notable research by Kamal kant et al. [42] suggested a model employing the Naïve Bayes algorithm for heart disease prediction. The research concluded that, in terms of accuracy, the Naïve Bayes algorithm demonstrated the highest effectiveness for predicting heart diseases, with Neural Networks and Decision Trees following closely behind.

Adding to the body of research, Nidhi Bhatla et al. [43] utilized diverse data mining methods for heart disease prediction. Their results indicated that the Neural Networks algorithm outperformed Decision Trees in terms of accuracy. Importantly, their study incorporated additional factors such as obesity and smoking, broadening the scope beyond conventional attributes.

Rishi Dubey et al. [44] Performed an examination investigating various machine learning algorithms for predicting heart disease. The review concluded that Neural Networks emerged as an effective method for heart disease prediction, indicating its potential to assist in treatment decisions.

Asha Rajkumar et al. [45] utilized a supervised machine learning classification approach to diagnose heart disease. The study meticulously partitioned the dataset into testing and training subsets, employing Naïve Bayes, Decision list, and K-NN algorithms. The findings revealed that Naïve Bayes exhibited a lower error ratio, affirming its effectiveness in heart disease diagnosis.

Sairabi H. Mujawar et al. [46] employed modified versions of K-means and Naïve Bayes algorithms for heart disease prediction. The Naïve Bayes model demonstrated impressive accuracy, achieving a 93 percent prediction rate for heart disease and an 89 percent accuracy rate for cases where patients did not have heart disease.

In 2017, Samuel et al. [47] forecasted the likelihood of heart failure utilizing Artificial Neural Network (ANN). Their methodology incorporated fuzzy analytic hierarchy (AHP) for determining the overall weights of features, facilitating personalized risk assessments. These feature contributions were subsequently utilized to train the ANN classifier to predict the patient's risk of heart failure.

Yekkala et al. [48] investigated bagging ensemble techniques like Random Forest and Adaboost in conjunction with Particle Swarm Optimization (PSO) for heart disease prediction. Their research attained notable accuracy levels using bagging, particularly with PSO.

Dolatabaddi et al. [49] employed an optimized Support Vector Machine for their classification model, extracting HRV signals from ECG in both time and frequency domains for the automated diagnosis of coronary artery disease. The comprehensive accuracy of their study underscored the resilience of the classification approach.

K. Sudhakar et al. [50] Utilized data mining methodologies for heart disease prediction, incorporating classification machine learning techniques like Decision Trees and Neural Network Naïve Bayes. Their study aimed to analyse and compare the effectiveness of classification algorithms on heart disease databases.

K Cinetha et al. [51] presented a decision support system using fuzzy logic for coronary heart disease. The model aimed to predict the possibility of being diagnosed with heart disease in the next ten years, achieving a remarkable accuracy of 97.67% on their dataset consisting of 1230 instances.

Reddy et al. [52] aimed to create the AGAFL (Adaptive Genetic Algorithm using Fuzzy Logic) model for early-stage heart disease prediction. The process involves initial feature extraction using rough set theory, followed by heart disease detection using the AGAFL classifier. The obtained results were compared with publicly presented datasets, demonstrating superior efficiency compared to existing methods. The authors suggest the potential extension of the algorithm by incorporating meta-heuristic algorithms for further enhanced results.

Babu et al. [53] introduced hybrid methods utilizing the Grey Wolf Optimization (GWO) combined with an auto-encoder based on Recurrent Neural Networks (RNN) for identifying various diseases. GWO extracts features, and disease identification is performed by RNN, yielding improved performance compared to existing methods. The models were evaluated using datasets such as Cleveland, mammographic, and Hungarian, showing a 16.82 percent accuracy improvement over other methods. The authors propose further improvements through the implementation of various techniques to enhance accuracy.

Santhi et al. [54] employed Genetic Algorithms to study various models for heart disease prediction and feature selection. The optimized use of genetic algorithms resulted in better performance compared to traditional methods. The models were retested with different heart disease datasets and evaluated in real-time with classifiers like Random Forest, Decision Tree, Naive Bayes, and Support Vector Machine. The findings indicated Naive Bayes as the superior classifier concerning the dataset.

Gokulnath et al. [55] focused on optimizing the function by combining Support Vector Machine (SVM) with Genetic Algorithm (GA) for significant feature selection. The proposed model achieved 88.34 percent accuracy in predicting heart disease, surpassing other techniques like CFS, chi-squared, Filtered subset, Info gain, and consistency subset. Additionally, the study highlighted the effectiveness of ROC analysis in SVM classifiers.

Nayak et al. [56] explored various classifiers for predicting cardiovascular disease at its initial stage. Employing classification methods like Naive Bayes, Support Vector Machine, Decision Tree Classifier, and k-NN classifiers, they aimed to identify early signs of the disease for preventive measures. Through a thorough analysis of detection techniques, Naive Bayes exhibited superior accuracy compared to other classification methods. Implementation in the R data analytical tool was performed for cardiovascular disease identification after thorough

filtration, and further enhancement through ensemble machine learning was suggested for improved accuracy.

Karadeniz et al. [57] introduced two distinct classifiers, namely Reference Vector Classifier (RVC) and Shrunk Covariance Classifier (SCC), for analysing randomness in the distance using medical Spectf and statlog datasets, respectively. The computed accuracies for Spectf and statlog datasets were 88.7 and 88.8 percent, demonstrating the efficiency of the Shrunk Covariance Classifier in measuring distance. The incorporation of different features as a reduction scheme in the architecture contributed to accuracy rate improvement.

Maji et al. [58] discussed the utilization of the decision tree algorithm as a classification method for predicting cardiovascular diseases. They proposed a hybridization method that combines the decision tree and artificial neural network, implemented as a single technique using WEKA. Evaluation through a tenfold test with a heart disease patient dataset from the UCI dataset demonstrated the effectiveness of the hybrid model, showcasing enhanced performance compared to individual classifiers. The output from the hybridization techniques indicated superior predictive capabilities for heart disease. The authors recommended the exploration of various data mining techniques for predicting different diseases at an earlier stage.

Patro et al. [59] employed a combination of methods to enhance classification and feature selection for improved heart disease prediction. The classification algorithm facilitated the analysis of the identification model, employing mathematical tools or computerized techniques to screen data with similar classes. Utilizing Bayesian Optimized Support Vector Machine (BO-SVM), Naive Bayes (NB), Salp Swarm Optimized Neural Network (SSA-NN), and K-Nearest Neighbours (KNN) classifiers, the proposed model was designed to predict heart disease based on the UCI dataset. BO-SVM exhibited superior performance with an accuracy of 93.3 percent and a sensitivity of 80 percent. The authors suggested the potential replacement of deep learning algorithms for further enhancement and improved performance.

In study [60][61][62][63], the researchers proposed various methods for detecting cardiovascular diseases using the UCI dataset. Specifically, they tested Tree Classifier, Random Forest, and K-Nearest Neighbours.

In work [64], a novel algorithm was introduced, analysing different approaches such as Decision Tree (DT), Random Forest (RF), and a hybrid model combining DT and RF. The hybrid model demonstrated improved accuracy compared to the individual models. The

research also involved designing an application for heart disease prediction, using health examination results as input and developing a graphical user interface with Python's Tkinter library.

In [65][66][67][68][69][70][71], Logistic Regression and Stochastic Gradient Descent were applied to predict heart disease, with Logistic Regression yielding the best results in terms of accuracy, precision, and recall. The authors suggested that a larger dataset could enhance results, highlighting the limitation of the small dataset in their models' performance. The proposed algorithms were considered as potential non-invasive diagnostic tools for detecting heart disease.

In research [72], [73][74][75][76], a system to prevent heart disease misdiagnosis was developed using Neural Network (NN) and Support Vector Machine (SVM), with SVM achieving superior results. The authors recommended SVM for the best diagnosis accuracy and proposed feature normalization to enhance classifier performance, a procedure also adopted in this thesis.

In research paper [77], [78], the authors implemented Neural Networks for efficient heart disease classification, emphasizing the accuracy and robustness of the diagnostic system. The study outlined fundamental steps for heart disease diagnosis, including data collection, pre-processing, outliers' elimination, and model development, and patient recruitment, consultation with the developed model, diagnosis, and evaluation by medical professionals.

In [79][80][81], the authors introduced an ensemble of Neural Networks for heart disease prediction. They suggested features removal based on correlation coefficient and employed entropy to select the best components of the Ensemble Neural Network (ENN), emphasizing computational efficiency.

In study [82][83], a Cascaded Neural Network classifier was proposed for heart attack prediction, achieving good results in terms of accuracy, sensitivity, specificity, and short prediction time. The authors described the structure of the Neural Network, which dynamically adds hidden units to minimize error, resulting in a computationally efficient tool for heart attack prediction.

Emerging trends also indicate a shift towards multi-modal approaches, where AI systems combine information from various sources, such as genetic data, imaging, and clinical records, to provide a more holistic understanding of an individual's cardiac health.

In summary, the literature on assisted AI medical diagnosis systems for heart disease reflects a multidimensional landscape. From advanced imaging techniques and wearable devices to interpretable AI models and NLP in EHR analysis, these studies collectively contribute to the evolution of precision medicine and the integration of AI in cardiovascular healthcare. The interdisciplinary nature of this research underscores the potential for synergistic advancements that could significantly impact the future of heart disease diagnosis and patient care.

2.12 A Summary of the Related Works.

Table 2: Literature Review and Gaps

Author	**Topic**	**Findings**	**Gaps**
(Mirzajani & salimi, 2018)	Utilizing Data Mining Techniques for Predicting and Diagnosing Diabetes	- Used WEKA data mining tool on applied classification algorithms to determine the algorithm with the best accuracy. - The classification algorithms used are KNN, j48 decision tree, SMO, and Naive Bayes (NB), and j48 decision tree had the best accuracy with 83.73%.	- Provision of a dashboard for data visualization.
(Enriko et al., 2016)	Creation of a Heart Disease Prediction	- Different classification model are used Decision Tree algorithms, Naive	- This research will focus on a combined

	System Using the k-Nearest Neighbor Algorithm and Streamlined Patient Health Data	Bayes and KNN with a dataset composed from California University, Irvine (UCI). - KNN classification algorithm produced the best accuracy with 81.85%.	dataset with different type of heart disease to create, test and train the model.
(Siva Kumar et al., 2016)	Employing Genetic Algorithms for Efficient Heart Disease Diagnosis	- 3-fold cross validation approach is used to validate the performance of the model which uses Genetic Algorithm (GA) technique. - 81.83% accuracy is achieved with 25 initial rule and 86.83% accuracy with 50 initial rule.	- To validate the performance of a model using a Bayesian classification.
(Zagorecki et al., 2013)	An Automated General Medical Diagnosis System Utilizing Bayesian Networks	- Developed a web-based medical diagnosis system using Bayesian Networks (BN) engines. - The majority of diagnoses were rendered for individuals aged 25-39 (52.1% female and 47.9% male), correlating with symptoms such as depression, anxiety	- Identify the actual likelihood of a patient having heart disease or not on all age groups.

		disorders, tension headaches, and fatigue. - Additionally, they explored the prevalent diagnoses among individuals aged 55-70 and 70+. The predominant diagnoses for patients in these older age brackets consistently revolved around age-related issues, including osteoarthritis, ischemic heart disease, bone injuries or joint, and gallstones.	
(Repaka et al., 2019)	Creating and employing prediction of heart disease system using Naive Bayesian methods	- Compared main techniques of BN and MLP, SMO (Sequential Minimal Optimization), and the result was presented by the recommended Navies Bayesian exhibiting superior performance at 89.77% compared to other methodologies.	- Develop an intelligent computerised system that will produce the likelihood of heart disease in a patient.

2.13 Chapter Summary

In this chapter, an extensive exploration of existing literature has been conducted, scrutinizing the works of various scholars and researchers in the domain of heart disease prediction. A recurrent theme in the literature review is the predominant focus on predicting a specific type of heart disease in isolation. However, a noticeable research gap emerges as many studies have not addressed the comprehensive prediction of the probability associated with all types of heart diseases.

The prevailing trend in the scholarly discourse has leaned towards targeted predictions, often honing in on distinct facets or categories within the spectrum of heart diseases. While these studies have undoubtedly contributed valuable insights into the predictive modelling of individual heart conditions, a notable void remains in the holistic assessment of the probabilities encompassing the diverse array of heart diseases.

This research endeavour, therefore, endeavours to bridge this critical research gap by adopting a more encompassing approach. Rather than concentrating solely on the prediction of a singular type of heart ailment, this study aspires to pioneer a methodology that comprehensively assesses the probabilities associated with the myriad forms of heart diseases. By doing so, it seeks to provide a more nuanced and comprehensive understanding of the predictive dynamics governing the entire spectrum of cardiac health.

In essence, the significance of this study lies not only in its novel contribution to the field of heart disease prediction but also in its endeavour to offer a more inclusive and comprehensive predictive model. By addressing this research gap, the research aims to enhance the efficacy and applicability of predictive models in the realm of cardiovascular health, ultimately contributing to more informed and proactive healthcare interventions.

3 RESEARCH METHODOLOGY

3.1 Introduction

This chapter provides a thorough examination of the research methodologies utilized in the study, covering aspects such as the proposed model, research framework, target population, sampling methodology, data collection methods, data analysis approaches, and the creation and execution of the web application. Each of these components plays a crucial role in shaping the research framework and contributes to the robustness of the study outcomes.

Proposed Model:

The chapter initiates by introducing the proposed model, a conceptual framework that serves as the foundation for the entire study. This model delineates the key variables, relationships, and mechanisms that the research seeks to investigate. It provides a theoretical underpinning for the subsequent stages of the study, guiding the research design and shaping the overall methodology.

Research Design:

The chosen research design is elucidated to articulate the overall plan and structure of the study. Whether it is experimental, observational, case study, or a combination of these, the design provides a blueprint for the systematic exploration of the research questions. The rationale behind selecting a specific design is expounded upon, justifying its appropriateness for the nature and scope of the study.

Study Population and Sampling Design:

A detailed discussion on the study population and the methodology employed for sampling is included. The characteristics of the target population, such as demographics and relevant parameters, are outlined. The sampling design, whether it is random, stratified, or purposive, is justified based on the research objectives and constraints. The chapter delves into considerations regarding sample size, ensuring it is representative of the broader population under scrutiny.

Data Collection Techniques:

This section provides insights into the techniques and tools employed for data collection. Whether it involves surveys, interviews, observations, or a combination of these methods, the chapter elaborates on the reasoning behind the chosen approaches. The development of any survey instruments or interview protocols is detailed, ensuring clarity and transparency in data collection.

Data Analysis Techniques:

The methods used for analysing the collected data are elucidated, underscoring the statistical or qualitative approaches applied. This may involve the use of software tools for statistical analysis or the application of thematic coding for qualitative data. The rationale for selecting specific analysis techniques is discussed, ensuring the robustness and reliability of the results.

Development and Implementation of the Web Application:

As an integral part of the study, the chapter provides a comprehensive overview of the development and implementation of the web application. This involves detailing the programming languages, frameworks, and technologies utilized. The user interface design, functionality, and integration with the proposed model are discussed, offering a holistic understanding of the technological aspects of the study.

In essence, this chapter serves as a methodological compass, guiding readers through the intricacies of how the research was conceptualized, designed, and executed. It lays the groundwork for the subsequent chapters, ensuring a transparent and well-justified methodology that contributes to the credibility and validity of the research findings.

3.2 Proposed Model

The envisioned research endeavours to forecast heart disease utilizing the Bayesian classification algorithm which is probabilistic approach known for its efficacy in handling uncertainties. The main aim of this research is to introduce an advanced medical diagnosis system specifically designed for predicting heart diseases. This system integrates carefully

recorded features by medical professionals to accurately assess the likelihood of a patient having heart disease. The process begins with healthcare professionals entering pertinent values extracted from the health report of a patient, which serve as essential data for the subsequent predictive analysis.

The temperament of this proposed system falls in a sophisticated model that receives the input data, undertaking a comprehensive assessment to prognosticate the probability of the patient being afflicted with heart disease. This predictive model is instrumental in providing valuable insights into the potential risk levels, empowering healthcare practitioners with timely and informed decision-making capabilities.

The intricacies of the heart disease prediction process are elucidated in Figure 2, offering a visual representation of the sequential steps involved in this prognostic endeavour. The interplay between input variables, data processing, and the Bayesian classification algorithm is graphically depicted, providing clarity on the intricate analytical journey from raw medical data to predictive outcomes.

Complementing this, Figure 3 delineates the proposed system architecture specifically designed for heart disease prediction. This architectural blueprint encapsulates the intricate network of components and functionalities that collectively contribute to the seamless operation of the predictive model. The system architecture provides a visual roadmap, offering insights into how the various elements interact to facilitate accurate and timely heart disease predictions.

In essence, this study not only leverages the prowess of the Bayesian classification algorithm for heart disease prognostication but also envisions and concretely develops an assisted medical diagnosis system that aligns with contemporary healthcare needs. By harnessing the insights derived from medical practitioners' inputs and encapsulating them within a robust predictive model, this research strives to contribute to the ongoing evolution of precision medicine and diagnostic efficacy in the realm of cardiovascular health.

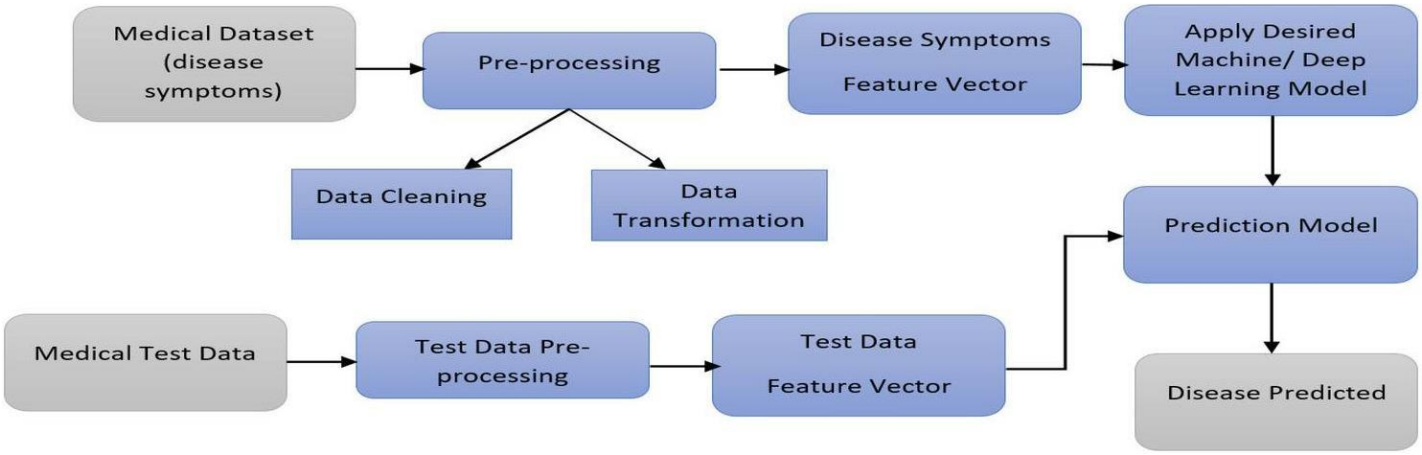


Figure 1: Framework for Heart Disease Prediction Model

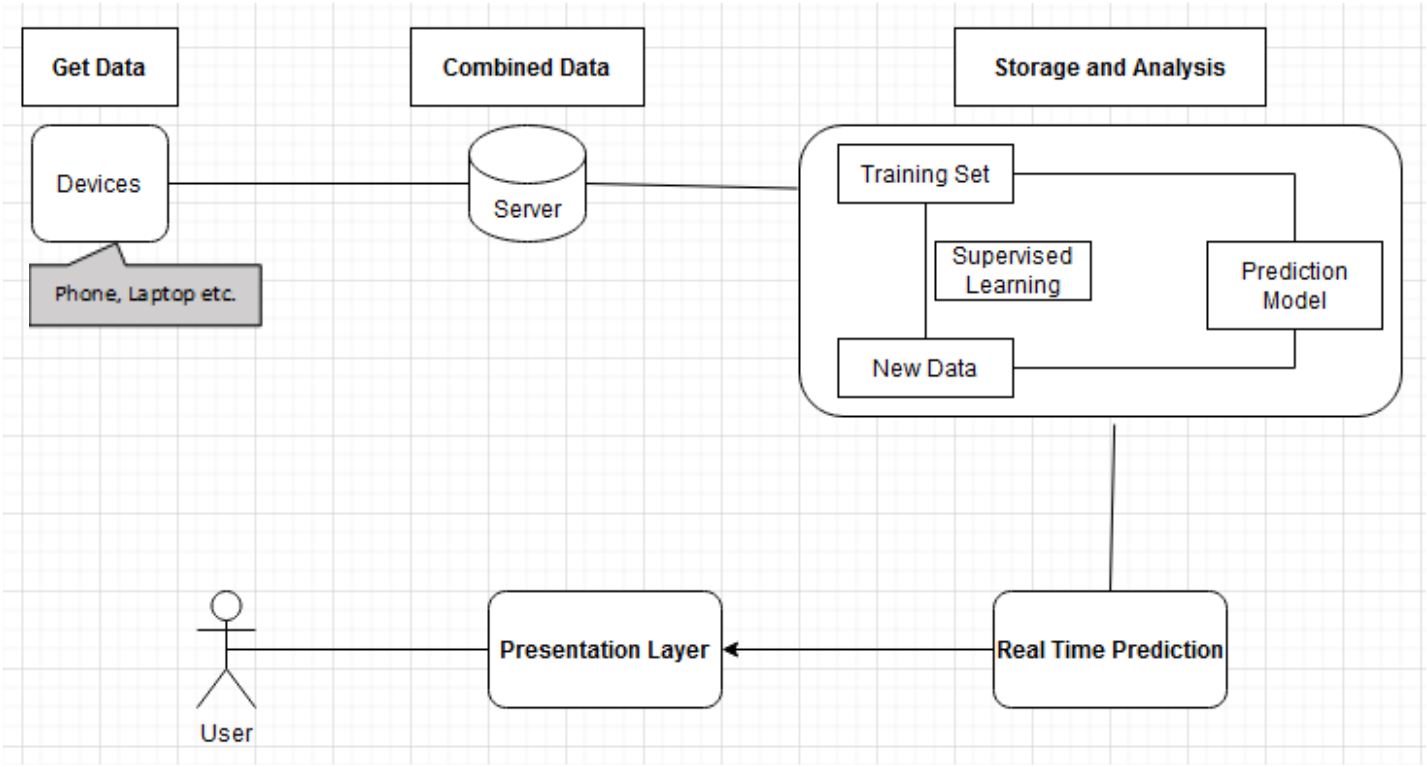


Figure 2: Proposed architecture for detection and prediction of heart disease

3.3 Research Design

The research approach adopted for this study was characterized by a mixed-methods design, seamlessly integrating both quantitative and qualitative data collection methodologies. This hybrid approach was deliberately chosen to facilitate a comprehensive and nuanced exploration of the research objectives, ensuring a well-rounded understanding that can be easily observed and analysed.

3.3.1 Quantitative Data Collection:

Quantitative data was collected to capture numerical and measurable aspects of the phenomena under investigation. This involved the systematic gathering of structured data through surveys, questionnaires, or other quantitative instruments. The emphasis on quantitative data allowed for the quantification of variables related to heart disease prediction, enabling statistical analysis and the derivation of objective insights. For instance, numerical indicators like heart rate, blood pressure, and cholesterol levels may have been systematically collected to quantify the risk factors associated with heart disease.

3.3.2 Qualitative Data Collection:

In parallel, qualitative data was also gathered to delve into the more nuanced and subjective dimensions of the research questions. Qualitative methods, such as interviews, focus group discussions, or open-ended survey questions, were employed to capture the richness and depth of participants' experiences, perceptions, and attitudes towards the assisted medical diagnosis system. This qualitative strand allowed for the exploration of the human aspects, contextual factors, and potential challenges associated with the utilization of the predictive model.

3.3.3 Combination of Qualitative and Quantitative Data:

The deliberate integration of both quantitative and qualitative data collection methods was strategic. It enabled a triangulation of findings, where the strengths of one method could complement the limitations of the other. For instance, quantitative data might provide statistical significance and generalizability, while qualitative data could offer depth and context to the statistical trends observed. This triangulation enhanced the overall reliability and validity of the study's findings.

3.3.4 Ease of Observation and Analysis:

The mixed-methods approach was particularly advantageous in facilitating the observation and analysis of data. The structured nature of quantitative data allowed for straightforward statistical analyses, facilitating the identification of patterns, correlations, and trends. On the other hand, the qualitative data, rich in descriptive content, provided a more in-depth understanding of the intricacies involved. The complementary nature of these data types contributed to a holistic interpretation of the research outcomes, making it easier for researchers to draw meaningful conclusions.

In summary, the mixed-methods research approach employed in this study, by combining quantitative and qualitative data collection techniques, aimed to provide a comprehensive and well-rounded exploration of the assisted medical diagnosis system for heart disease prediction. This approach facilitated a more nuanced understanding of the subject matter, allowing for both statistical rigor and contextual insights in the analysis of the collected data.

3.4 Baseline Study

3.4.1 Population of the Study

The study relied on sample sizes derived from heart disease patients, which were generously provided by medical practitioners associated with the National Heart Disease Hospital in Lusaka, Zambia. This collaborative effort with healthcare professionals specializing in cardiovascular care was pivotal in ensuring the availability of relevant and authentic data essential for the study's objectives.

To facilitate the collection of this valuable information, a meticulously designed questionnaire was developed. This questionnaire served as a structured tool tailored to extract specific data points relevant to the study's focus on heart disease prediction. The questions encompassed a range of variables such as age, gender, blood pressure, cholesterol levels, and other pertinent medical parameters. This instrument was crafted with precision to elicit comprehensive responses that could contribute to a thorough analysis of the factors influencing heart disease prediction.

The distribution of the questionnaire was conducted in a streamlined manner, leveraging the convenience and efficiency of modern communication methods. Specifically, the questionnaire was disseminated to medical practitioners via email. This electronic distribution not only expedited the data collection process but also facilitated ease of access for the busy healthcare

professionals involved in the study. The use of email as a communication channel allowed for a seamless exchange of information, enabling medical practitioners to provide responses at their convenience while maintaining the integrity and security of the data.

The engagement with medical practitioners through email correspondence was characterized by a collaborative and communicative approach. Clear instructions were provided alongside the questionnaire to ensure that the respondents understood the study's objectives and the type of information sought. Moreover, this mode of data collection offered the flexibility for practitioners to contribute their insights and expertise, enriching the dataset with valuable qualitative nuances that complemented the quantitative data.

By establishing this collaborative relationship with medical practitioners and utilizing email as a means of data collection, the study not only benefited from the expertise of professionals directly involved in heart disease care but also ensured a systematic and ethical approach to acquiring patient-related information. This collaborative effort between researchers and medical practitioners at the National Heart Disease Hospital stands as a testament to the synergy between the scientific community and healthcare practitioners in advancing knowledge and understanding in the field of cardiovascular health.

3.4.2 Sampling Technique and Sample Size

The foundation of this study rested on the utilization of sample sizes extracted from a comprehensive dataset comprising 1190 patients diagnosed with heart disease. This dataset, a reservoir of diverse patient profiles, became an invaluable resource for the investigation, enabling a robust analysis of patterns, trends, and factors influencing heart disease prediction. The sheer breadth of the dataset, encompassing a considerable number of patients, fortified the study's statistical power and capacity for generating meaningful insights.

In addition to leveraging the dataset, the study forged a collaborative partnership with medical practitioners to acquire supplemental and contextually rich data. Recognizing the significance of real-world clinical expertise, the study sought the input of healthcare professionals who actively engage with heart disease patients. This collaborative approach ensured that the study's findings were grounded in the practical realities of cardiovascular care, enhancing the external validity and applicability of the research outcomes.

To facilitate the collection of this dual-source data — both from the dataset and the insights provided by medical practitioners — a meticulously crafted questionnaire was developed. This instrument was meticulously designed to extract a spectrum of information vital for the study's objectives. The questionnaire encompassed an array of variables, including patient demographics, medical history, lifestyle factors, and specific diagnostic parameters, ensuring a holistic exploration of factors influencing heart disease prediction.

To engage medical practitioners effectively, the questionnaire was distributed using a multifaceted approach. A digital dissemination method was employed, where the questionnaire was sent to practitioners via email. This electronic mode not only expedited the data collection process but also offered a convenient platform for practitioners to provide detailed responses. Simultaneously, recognizing the diversity in communication preferences, hard copies of the questionnaire were also distributed to practitioners, ensuring inclusivity and flexibility in the data collection process.

The collaborative endeavour with medical practitioners went beyond a mere exchange of data; it fostered a meaningful dialogue between the research team and frontline healthcare providers. This two-way interaction not only enriched the quantitative dataset with nuanced qualitative insights but also ensured that the study's objectives were aligned with the practical challenges and considerations faced by practitioners in the field of cardiovascular medicine.

In summary, this study embraced a dual-source approach, drawing on both a substantial dataset and the expertise of medical practitioners. The amalgamation of these distinct but complementary data sources empowered the study to explore heart disease prediction comprehensively, blending statistical rigor with real-world clinical insights. This multidimensional approach aligns with contemporary research paradigms that recognize the multifaceted nature of healthcare phenomena and the need for collaborative efforts to advance knowledge in the field.

3.4.3 Methods for Data Collection

This research employed a comprehensive approach by harnessing data from both primary and secondary sources. The dataset under scrutiny was meticulously crafted by amalgamating disparate datasets, each independently available, yet never before combined. This innovative approach resulted in the creation of a unique dataset that stands as the most extensive and robust compilation for heart disease research to date. This amalgamated dataset integrates information

from five distinct heart datasets, unifying them over 11 common features, thereby establishing its distinction as the largest dataset for heart disease tailored explicitly for the purpose of research.

The datasets contributing to this curated compilation include:

1. Cleveland: Encompassing 303.
2. Hungarian: Comprising 294.
3. Switzerland: Containing 123.
4. Beach VA: Incorporating 200.
5. Stalog (Heart) Data Set: Contributing 270.

In aggregate, this amalgamation results in a dataset of substantial magnitude, totalling 1190 observations. The utilization of these diverse datasets adds a layer of complexity and richness to the study, ensuring that the findings draw upon a varied and representative sample of heart-related cases.

The dataset encompasses a total of 11 features, forming the cornerstone of the analytical framework. The amalgamated dataset, readily accessible on the Kaggle website [84], was employed for the analysis, allowing for a meticulous examination of the interconnected features and their impact on heart disease prediction.

Table 1, provided below, furnishes a detailed description of the 11 features encapsulated within the dataset, offering a comprehensive understanding of the variables considered in the study. This amalgamated dataset, born from the harmonization of disparate sources, stands as a testament to the research's commitment to leveraging diverse and extensive data for a thorough exploration of heart disease prediction dynamics.

[Table 1: Description of 11 Features in the Dataset]

This innovative approach to dataset curation not only broadens the scope of the study but also ensures that the analysis is conducted on a dataset with unparalleled breadth and depth, setting the stage for nuanced and informed insights into heart disease prediction.

Table 3: Description Features

No	Features Description	Distinct Values of Features
1	Age: patient's age	Years
2	Sex: patient's sex [M: Male, F: Female]	M, F
3	ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	TA, ATA, NAP, ASY
4	RestingBP: resting blood pressure [mm Hg]	Values in mm/hg
5	Cholesterol: serum cholesterol [mm/dl]	Values in mm/dl
6	FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	1, 0
7	RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]	Normal, ST, LVH
8	MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]	Numerical values between 60 & 202
9	ExerciseAngina: exercise-induced angina [Y: Yes, N: No]	Y, N
10	Oldpeak: oldpeak = ST [Numeric value measured in depression]	Numeric value measured in depression
11	ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Up, Flat, Down
12	HeartDisease: output class [1: heart disease, 0: Normal]	1, 0

3.4.4 Data Collection Instruments

The cornerstone of data collection in this study was a carefully designed and multifaceted approach, employing a combination of instruments to ensure a comprehensive and robust gathering of information essential for addressing the research questions. The primary data collection instruments employed were questionnaires, in-depth interviews, and a pre-existing dataset sourced from Kaggle.

In essence, this multifaceted data collection strategy underscored the commitment of the study to capture the complexity of heart disease prediction comprehensively. The synergy between

structured instruments, qualitative exploration, and the utilization of an extensive dataset reflects a methodological rigor aimed at providing nuanced and informed answers to the research questions.

3.4.5 Questionnaire

Roopa and Rani [85] defines a questionnaire as ‘a list of mimeographed or printed questions that is completed by or for a respondent to give his opinion’. The questionnaire was designed in four parts: the first part being demographic information, the second part targeted computer knowledge and experience, the third part was adoption factors for the system and the fourth part focused on the usage of the system.

A structured questionnaire served as a pivotal tool for systematically collecting quantitative data. This instrument was meticulously crafted to capture specific variables relevant to the study's objectives. The questionnaire, administered to a targeted sample, facilitated the acquisition of standardized responses, allowing for quantitative analysis and statistical interpretation. The structured nature of the questionnaire ensured consistency in data collection across respondents, contributing to the reliability of the gathered information.

3.4.6 In-depth Interviewing

In-depth interviewing is a qualitative research technique that comprises of intensive individual interviews with a small number of respondents to explore their viewpoints on a particular idea, program, or situation [86]. The interviews were conducted on the health personnel’s at the National Heart Disease Hospital in Zambia.

In addition to structured questionnaires, qualitative insights were gleaned through in-depth interviews. This method provided a platform for open-ended discussions with selected participants, including medical practitioners, researchers, or individuals with expertise in the field of heart disease. In-depth interviews allowed for the exploration of nuanced perspectives, experiences, and contextual factors that might not be fully captured through quantitative measures alone. The qualitative data derived from interviews enriched the study with a deeper understanding of the intricacies surrounding heart disease prediction.

3.4.7 Kaggle Dataset

To complement the primary data collection methods, a pre-existing dataset sourced from Kaggle was employed. Kaggle, a platform known for hosting diverse datasets, provided a valuable resource for this study. The dataset, pre-compiled and accessible, originated from a combination of independent datasets related to heart disease. Leveraging Kaggle's repository added a layer of efficiency to the research process, allowing for the utilization of a substantial dataset already curated and processed. This approach not only saved time but also ensured access to a large and diverse dataset, enhancing the study's capacity for comprehensive analysis.

The amalgamation of these diverse data collection instruments — questionnaires, in-depth interviews, and a Kaggle dataset — created a methodological synergy. The combination of quantitative and qualitative data allowed for a triangulated understanding of heart disease prediction, offering a holistic view that goes beyond numerical trends to encompass the human and contextual dimensions of the subject matter.

3.4.8 Data Analysis

Following the meticulous collection of data, the researcher embarked on a crucial phase of data processing and editing to enhance the overall quality and coherence of the dataset. The objective was to ensure consistency among responses, eliminate any discrepancies, and prepare the data for subsequent analyses. To achieve this, advanced statistical tools, with a specific focus on the Jupyter Notebook, were employed to streamline and simplify the dataset, making it both interpretable and understandable.

Data Editing and Consistency Check:

The initial step in this post-collection phase involved a comprehensive review of the gathered data. The researcher meticulously examined each dataset, scrutinizing responses for completeness, accuracy, and uniformity. Any missing or anomalous entries were identified and addressed to uphold the integrity of the dataset. This process aimed at rectifying discrepancies and ensuring that the data accurately reflected the intended information.

Utilization of Statistical Tools - Jupyter Notebook:

Jupyter Notebook, a powerful and versatile tool in the realm of data science, played a pivotal role in the subsequent stages of data processing. Leveraging the capabilities of Jupyter Notebook, the researcher initiated a series of statistical analyses, employing Python

programming language and associated libraries. This interactive computing environment facilitated the execution of code snippets, allowing for real-time exploration and manipulation of the dataset.

Data Simplification and Interpretability:

Jupyter Notebook was particularly instrumental in simplifying the dataset and enhancing its interpretability. Through the application of Python scripts and statistical functions, the researcher transformed raw data into a more structured and readable format. This not only facilitated a clearer understanding of the dataset but also laid the groundwork for more sophisticated analyses.

Visual Representation and Exploration:

Visualizations, generated within the Jupyter Notebook environment, further aided in comprehending the patterns and trends embedded in the data. Graphical representations, such as charts and graphs, were created to present key insights visually. This not only simplified complex information but also allowed for a more intuitive grasp of the dataset's characteristics.

Iterative Process and Quality Assurance:

The utilization of Jupyter Notebook in the data processing phase was an iterative process. The researcher engaged in multiple rounds of analyses, refining the dataset and ensuring that it met the predefined criteria for quality and consistency. This iterative approach, coupled with the interactivity of Jupyter Notebook, enabled real-time adjustments and quality assurance checks.

In summary, the post-data collection phase was marked by a meticulous editing process aimed at enhancing data consistency. The integration of statistical tools, notably Jupyter Notebook, not only simplified the dataset but also empowered the researcher to delve into complex analyses. This systematic approach laid the groundwork for subsequent stages of the study, ensuring that the dataset was well-prepared for meaningful interpretations and insights.

3.5 System Design and Implementation

3.5.1 Requirements Specification

The requirement specification phase in system architectural analysis is a pivotal step that involves the delineation of two integral components: hardware and software. Each of these components plays a crucial role in shaping the overall system architecture [87].

In the context of system architecture, hardware components represent the tangible and physical elements that constitute the system's infrastructure. These include the client's computers, servers, and the underlying database that form the interconnected backbone of the system [87]. The client's computer serves as the end-user interface, providing the point of interaction with the system. Servers act as the central processing units, handling requests, and managing data flow, while the database serves as the repository for storing and retrieving information. The harmonious integration of these hardware components is paramount for the seamless functioning of the system.

Contrasting with hardware, software components define the dynamic behavior of the system and elucidate how its various elements interact. These components cater to different stakeholders, including architects, programmers, and customers, offering diverse architectural views tailored to their specific needs [87]. Within the realm of software components, detailed descriptions emerge, encompassing the system's functionality, the interconnection of its components, the collaborative workflow, the developmental methodologies, and the application of software on the underlying hardware infrastructure.

Software components are intricately linked to the essential requirements of the system, which can be broadly categorized into functional and non-functional requirements [87]. Functional requirements delineate the specific functionalities and features that the software must deliver. These include user interactions, system responses, and the overall behavior of the software in response to user input. On the other hand, non-functional requirements encompass aspects like performance, security, scalability, and reliability, addressing the quality attributes that characterize the software's overall effectiveness and efficiency.

In summary, the requirement specification phase not only entails the identification of hardware components that form the foundational structure of the system but also delves into the intricate details of software components. This includes the behavioral aspects, architectural views catering to diverse stakeholders, and the alignment with both functional and non-functional requirements. By elucidating the hardware and software components in the requirement specification, this phase lays the groundwork for the subsequent design and development stages of the system.

3.5.1.1 Functional Requirements

These are requirements that specify what the system must do according to the functions of the system or its components, when the function is described as defining the behaviour between outputs and inputs. The following are the requirements that makeup the heart disease medical diagnosis system:

- System should allow health personnel's to register.
- System should allow health personnel's registered to login.
- System should allow health personnel's input required data for prediction.
- System should display the probability output results.

3.5.1.2 Non-Function Requirements

These are requirements that define the characteristics that a system is expected to have and can be used to evaluate its performance rather than its behaviour. All components of the proposed heart disease medical diagnosis system require the following non-functional requirements:

- System must have acceptable performance.
- System services should be durable. This means the services should be able to recover from failures or interruption by automatically resuming or restarting the service.
- System must be scalable.

3.5.1.3 Hardware Requirements

- Processor : Intel(R) Core(TM) i7-5500U CPU @ 2.40GHz 2.40 GHz
- RAM : 10.0 GB
- System type : 64-bit operating system, x64-based processor
- Version : 21H2
- Edition : Windows 10 Pro
- Device Name : Lenovo

3.5.1.4 Software Requirements

- Operating System : Windows
- Technology : Python (Flask)
- Web Technologies : Html, JavaScript, CSS
- IDE : Visual Studio Code
- Management System : SQLite

- ❖ Model Creation : Jupyter Notebook

3.5.2 Design Specification

3.5.2.1 Use Case Diagrams

Use case diagrams serve as crucial behavioral diagrams in the realm of system design, offering a visual representation that encapsulates the interactions between users and the system. These diagrams provide a simplified and comprehensible view of the system's functionalities, categorizing them into distinct use cases and illustrating how these functionalities interact with both internal and external users [88][89].

3.5.2.1.1 Key Elements of Use Case Diagrams:

Actors:

In a use case diagram, actors are the entities engaging with the system, which may include external users, other systems, or even time-triggered events. Actors are typically depicted by stick figures or other symbols, and they interact with the system through various use cases.

Use Cases:

Use cases depict the particular functionalities or tasks that the system executes to achieve a user's objective. They are illustrated as ovals and are linked to actors via lines, denoting the interaction between the actor and the system.

Relationships:

The lines connecting actors and use cases denote relationships. For instance, a solid line signifies a primary relationship, while additional notations may indicate associations, dependencies, or include arrows to illustrate the direction of the interaction.

System Boundary:

The system boundary, typically depicted by a box, encompasses all the actors and use cases within the system. This boundary defines the extent of the system being analyzed.

3.5.2.1.2 Benefits of Use Case Diagrams:

Simplified Communication:
Use case diagrams provide a simplified means of communication by presenting a high-level overview of system functionalities. This clarity aids in better understanding, especially for stakeholders who may not have technical expertise.

Requirements Clarification:
Use case diagrams play a crucial role in requirements analysis and clarification. They help in identifying and defining the various use cases, ensuring that all potential functionalities are considered during system development.

User-Centric Design:
By focusing on user interactions, use case diagrams facilitate a user-centric design approach. They ensure that the system is designed to meet the needs and goals of the users effectively.

Bridging the Gap:
Use case diagrams serve as a bridge between technical and non-technical stakeholders. They offer a common language that both developers and business stakeholders can understand, fostering collaboration and alignment.

3.5.2.1.3 Application in System Development:

In the context of system development, use case diagrams are often employed during the early stages of requirements gathering and analysis. They act as a foundational blueprint, guiding the development team in understanding user interactions and system functionalities. As the project progresses, use case diagrams continue to be valuable for communication, testing, and ensuring that the final system aligns with the initially envisioned functionalities.

In essence, use case diagrams are a powerful tool for capturing, communicating, and understanding the dynamic aspects of a system, contributing significantly to the success of system development projects.

In Figure 4 is the use case diagram represented for the system.

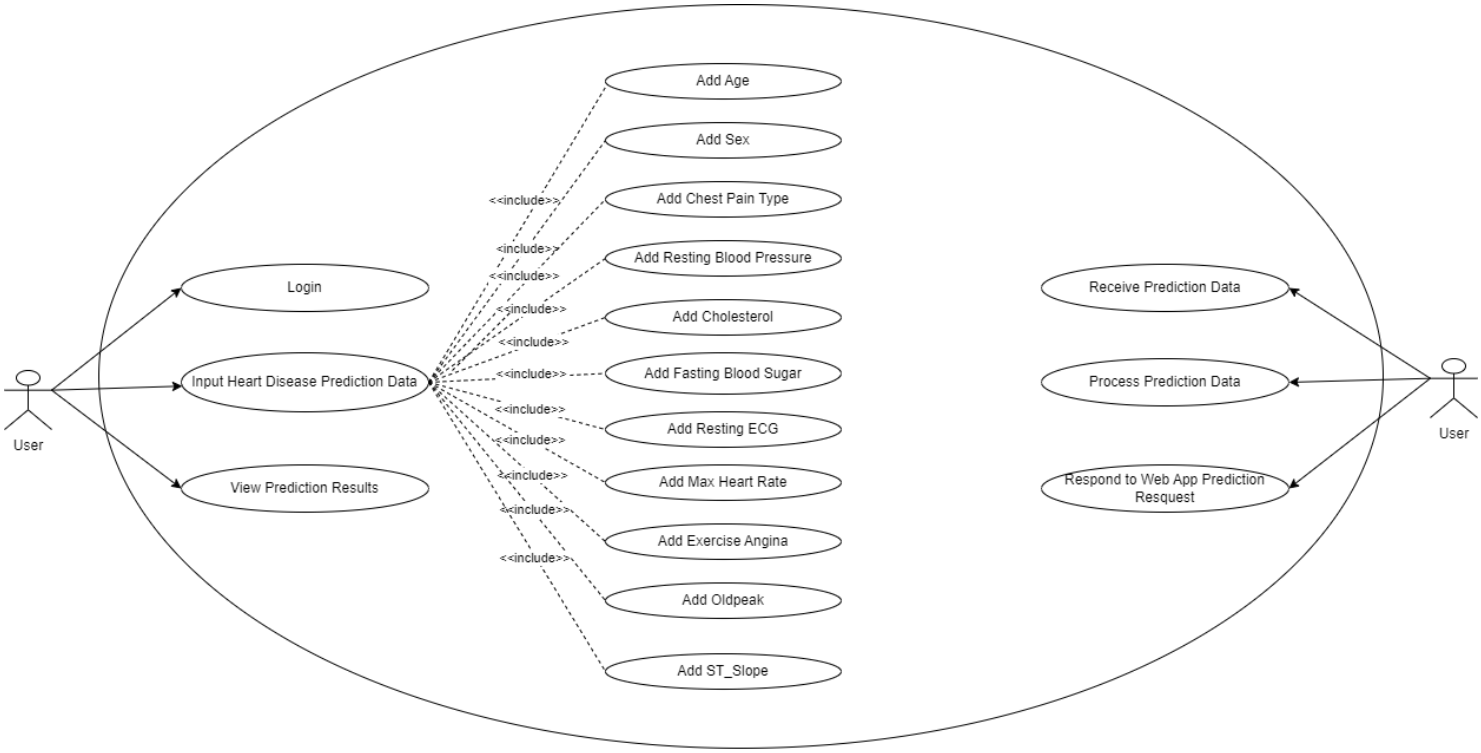


Figure 3: Web App Use Case Diagram

3.5.2.2 Flow Chart Diagram

Flowchart diagrams play a pivotal role in system design, providing a visual representation that illustrates the sequential flow of data within the system [90][91]. Figure 5, in this context, serves as a comprehensive flowchart detailing the intricacies of the entire Prediction System, elucidating how data should navigate through the system and facilitating a clear understanding of communication pathways.

3.5.2.2.1 Key Elements of the Flowchart Diagram:

Start and End Points:

Flowcharts typically commence with a start point and conclude with an end point, symbolizing the initiation and termination of the process. These symbols are often represented by circles, denoting the beginning and conclusion of the data flow.

Process Boxes:

Rectangular boxes within the flowchart denote specific processes or actions. Each box represents a discrete step or operation in the data flow. In the context of the Prediction System, these processes might encompass data pre-processing, feature extraction, model training, and result generation.

Arrows:

Arrows connect the various elements of the flowchart, indicating the directional flow of data. The arrows delineate the sequence of steps, guiding the reader through the logical progression of the system.

Decision Points:

Diamond-shaped symbols represent decision points where the flow of data can take different paths based on certain conditions. These decision points introduce conditional logic, allowing for dynamic and context-dependent data routing.

Input/Output:

Input and output symbols, often represented by parallelograms, signify the points where data enters or exits the system. These points denote interactions with external entities or the initiation and conclusion of specific processes.

3.5.2.2.2 Benefits of Flowchart Diagrams:

Visual Clarity:

Flowchart diagrams provide a visual and intuitive representation of complex processes, enhancing clarity and understanding. This visual clarity is particularly valuable for stakeholders who may have varying levels of technical expertise.

Sequential Logic:

The sequential nature of flowcharts helps in detailing the step-by-step logic of the system. This sequential logic aids in identifying dependencies, ensuring that each step is executed in a logical order.

Process Optimization:

Flowcharts are instrumental in identifying opportunities for process optimization. By visually mapping the data flow, inefficiencies or redundancies can be identified and addressed to enhance overall system efficiency.

Communication Tool:

Flowchart diagrams serve as a powerful communication tool among diverse stakeholders. They provide a common language for discussing and understanding the intricacies of the Prediction System, fostering collaboration and alignment.

3.5.2.2.3 Application in System Design:

In the context of the Prediction System, the flowchart diagram presented in Figure 5 serves as a blueprint for the entire data processing journey. It delineates the systematic flow of data from its initiation through various processes, decision points, and ultimately to the generation of predictions. This visualization is instrumental in guiding system architects, developers, and other stakeholders through the intricacies of the Prediction System, ensuring a shared understanding of the data flow logic.

In conclusion, flowchart diagrams are indispensable in system design for their ability to provide a structured, sequential representation of data flow. The presented flowchart for the Prediction System serves as a valuable tool for understanding, communicating, and refining the processes involved in making accurate predictions based on the system's architecture.

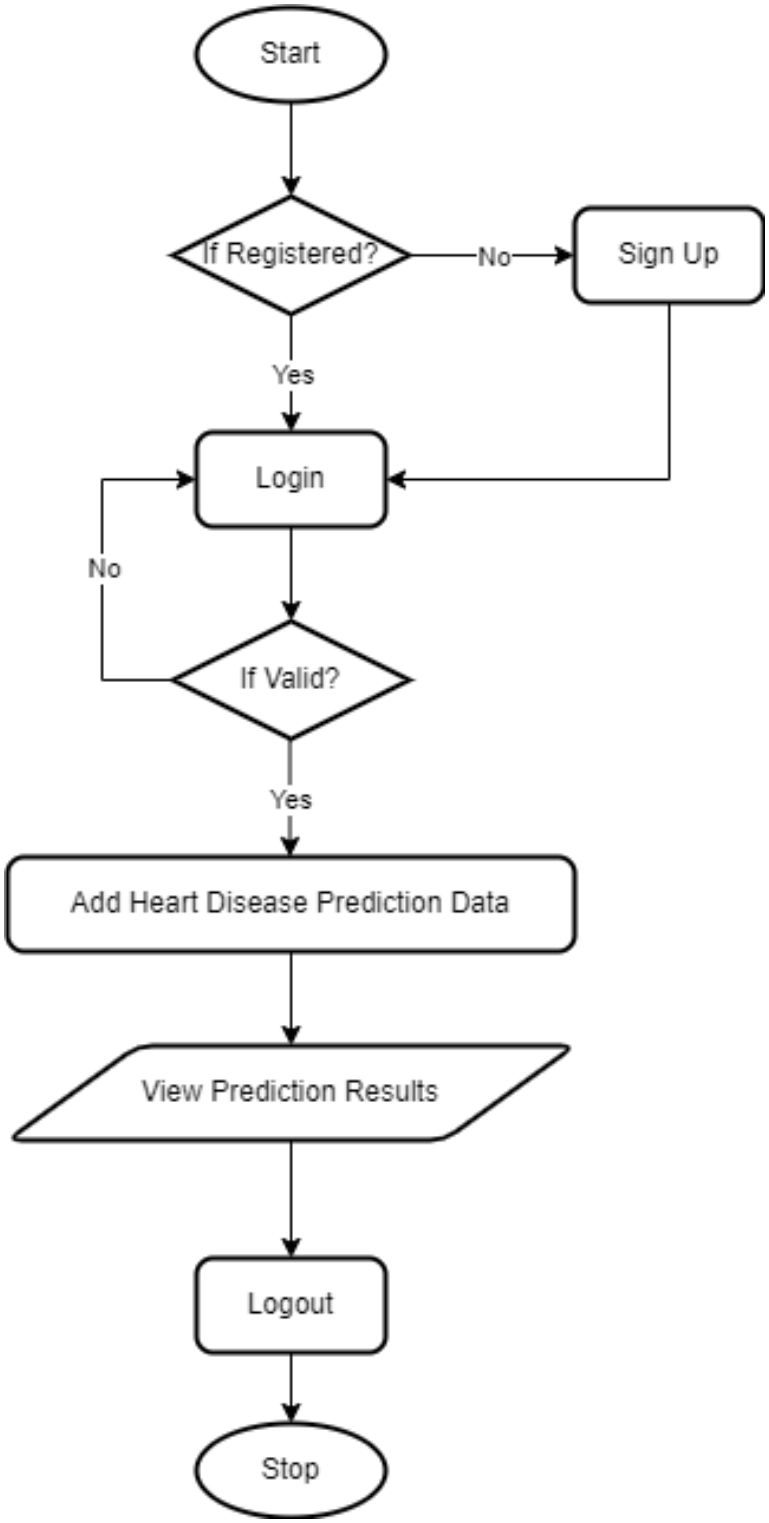


Figure 4: System Analysis Flow Chart Diagram

3.5.2.3 Data Flow Diagram

A data flow diagram (DFD) serves as a powerful visualization tool in systems analysis and design, illustrating how information traverses through processes and subsystems within a larger system [92][93][94]. Figure 6, in this context, provides a comprehensive representation of how data flows within the Prediction System, capturing the intricate movements and transformations that occur throughout the entire process.

Key Components of the Data Flow Diagram:

Processes:
Processes in a DFD represent specific functions or operations performed within the system. In the context of the Prediction System, these processes might include data preprocessing, feature extraction, model training, and result generation. Each process is depicted as a labeled circle or ellipse.

Data Flows:
Data flows are represented by arrows connecting various components in the diagram, indicating the direction in which data moves. These arrows showcase the path of data from its sources to its destinations, revealing the relationships and dependencies between different processes and data entities.

Data Stores:
Data stores are represented by rectangles and symbolize repositories where data is stored. In the Prediction System, data stores could include databases, files, or any other storage mechanisms holding relevant information.

External Entities:
External entities are entities outside the system boundary that interact with the system. These entities are often represented by squares or rectangles. In the context of the Prediction System, external entities could be sources of input data or recipients of prediction results.

Data Movement in the Prediction System:

Input Data Flow:

The diagram captures the entry point of data into the system, typically originating from external entities. This input data flow symbolizes the initial information that triggers the predictive processes.

Internal Data Flows:

As data progresses through the system's processes, the diagram delineates internal data flows, showcasing how information undergoes various transformations and analyses. These internal data flows highlight the dynamic nature of data as it moves through different stages.

Output Data Flow:

The ultimate result of the Prediction System is depicted through the output data flow. This represents the predictions or insights generated by the system, which may be delivered to external entities or stored for further analysis.

Data Stores:

The presence of data stores in the diagram emphasizes the storage and retrieval of data at different stages of the process. These data stores act as temporary repositories or long-term storage for information critical to the system.

Benefits of Data Flow Diagrams:

Visualization of Processes:

DFDs provide a visual representation of the processes within a system, aiding in the understanding of how data is processed and transformed.

Identification of Dependencies:

By showcasing the flow of data, DFDs assist in identifying dependencies between different processes and data entities, contributing to a holistic understanding of system dynamics.

Communication Tool:

DFDs serve as effective communication tools, enabling stakeholders, including developers and non-technical personnel, to grasp the data movement and processing logic in a system.

Application in System Analysis:

In the case of the Prediction System, Figure 6 serves as a valuable analytical tool for understanding the flow of data. It assists system analysts and designers in dissecting the intricacies of information movement, guiding the development process, and ensuring that data is processed in a logical and efficient manner.

In conclusion, the data flow diagram for the Prediction System encapsulates the essence of how information traverses through the system, offering a structured visualization that aids in system analysis, design, and communication among diverse stakeholders.

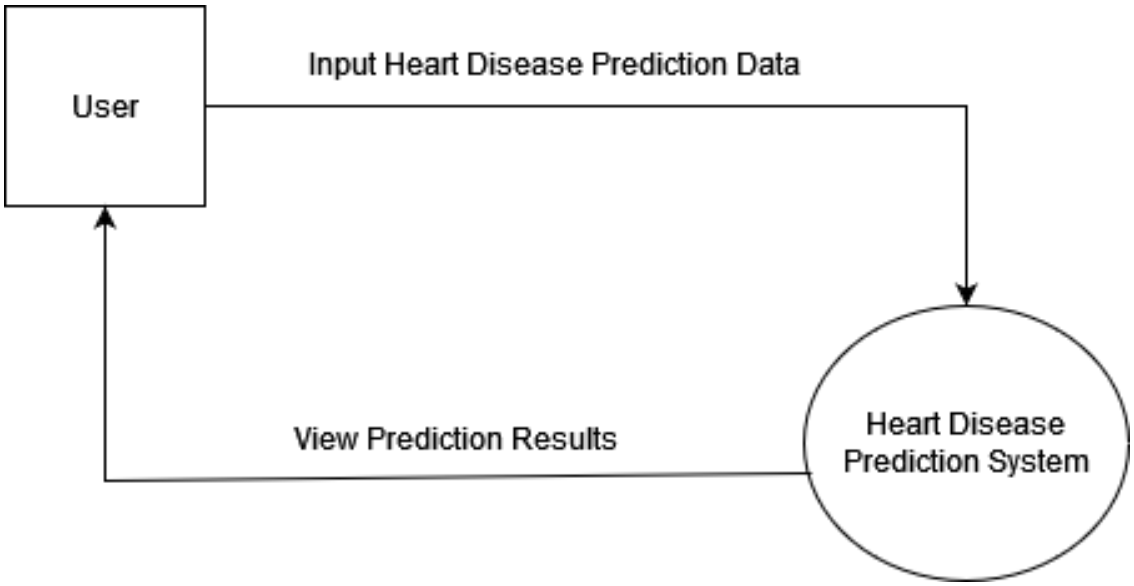


Figure 5: Web App Data Flow Diagram

3.5.2.4 Sequence Diagrams

Sequence diagrams stand as a valuable tool in systems modelling, specifically designed to depict the dynamic interactions between actors and objects within a system. They provide a sequential representation of the processes, showcasing how different components collaborate to fulfil specific functionalities[95][88][96]. In the context of the Prediction System, Figures 7 through 9 present sequence diagrams, each capturing the dynamic aspect of crucial functions performed by system components.

Key Aspects of Sequence Diagrams:

Actors:

In sequence diagrams, actors are entities external to the system that initiate interactions. These can be users, external systems, or any other components that trigger processes within the system.

Objects:

Objects represent instances of classes or entities within the system. They are central to sequence diagrams as they actively participate in interactions, responding to messages and executing specific actions.

Lifelines:

Lifelines represent the existence of an object over time during the sequence. They are depicted as vertical dashed lines, providing a visual representation of the temporal aspect of the object's involvement.

Messages:

Messages are arrows indicating communication between objects. They illustrate the flow of information, signalling method calls, responses, or other types of interactions.

Dynamic Modelling in Sequence Diagrams:

Initiation of Processes:

Sequence diagrams illustrate how processes are initiated, often triggered by external actors. These diagrams effectively capture the chronological order of events, providing a dynamic portrayal of system functionality.

Object Interactions:

The interactions between objects are meticulously detailed, showcasing how messages are passed between them. This dynamic representation is essential for understanding the sequence of actions that take place during the execution of a particular function.

Temporal Relationships:

The lifelines in sequence diagrams offer insight into the temporal relationships between objects. By visualizing the duration of an object's involvement in a process, stakeholders can discern the order and duration of actions.

Benefits of Sequence Diagrams:

Dynamic System Understanding:

Sequence diagrams offer a dynamic perspective, aiding in the understanding of how processes unfold and objects interact over time.

Communication and Collaboration:

These diagrams serve as effective communication tools, fostering collaboration among development teams, stakeholders, and other involved parties.

Identification of System Flows:

Sequence diagrams assist in identifying the flow of information and interactions within the system, helping to pinpoint potential bottlenecks or areas for optimization.

Application in System Design:

In the context of the Prediction System, sequence diagrams play a crucial role in system design. They provide a blueprint for developers, guiding them in implementing functions by illustrating the step-by-step interactions and collaborations between different components. Moreover, sequence diagrams facilitate alignment among stakeholders, ensuring a shared understanding of the dynamic aspects of the system.

In conclusion, Figures 7 through 9, depicting sequence diagrams, offer a dynamic visualization of important functions within the Prediction System. These diagrams serve as essential tools for system analysts, developers, and other stakeholders, contributing to a comprehensive understanding of the system's dynamic behaviour and interactions.

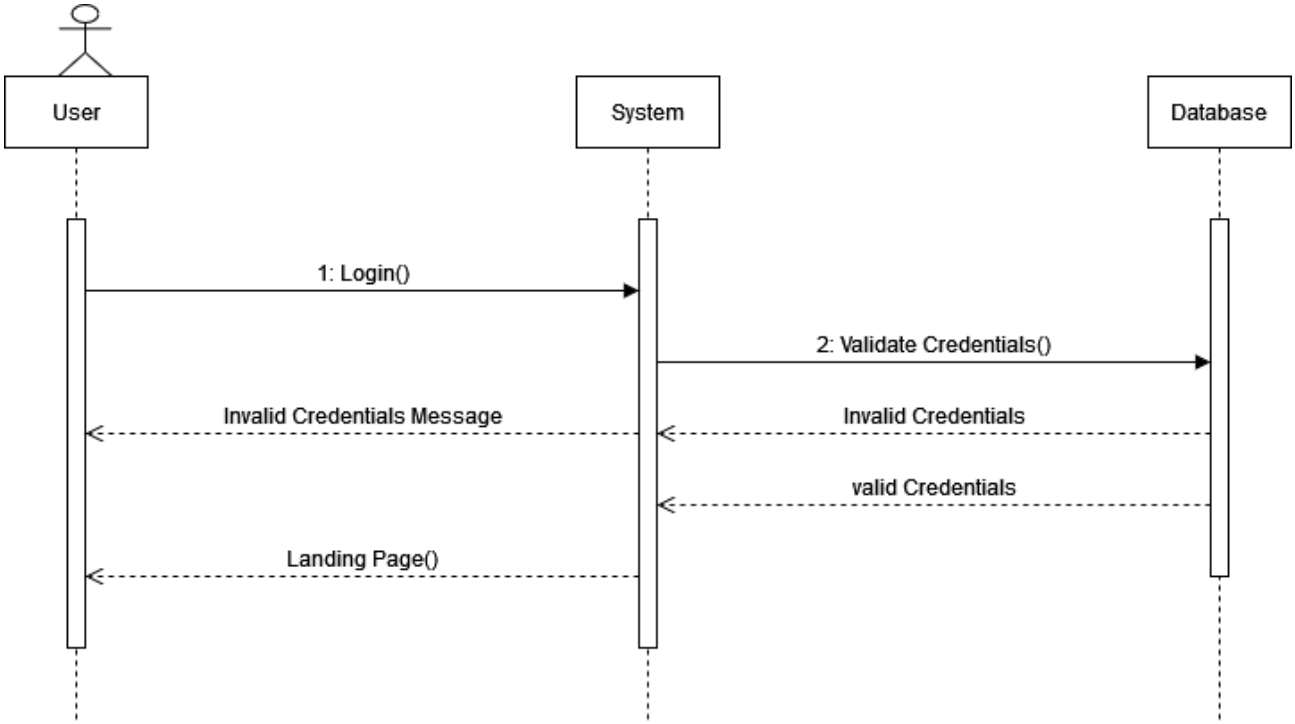


Figure 6: Login Sequence Diagram

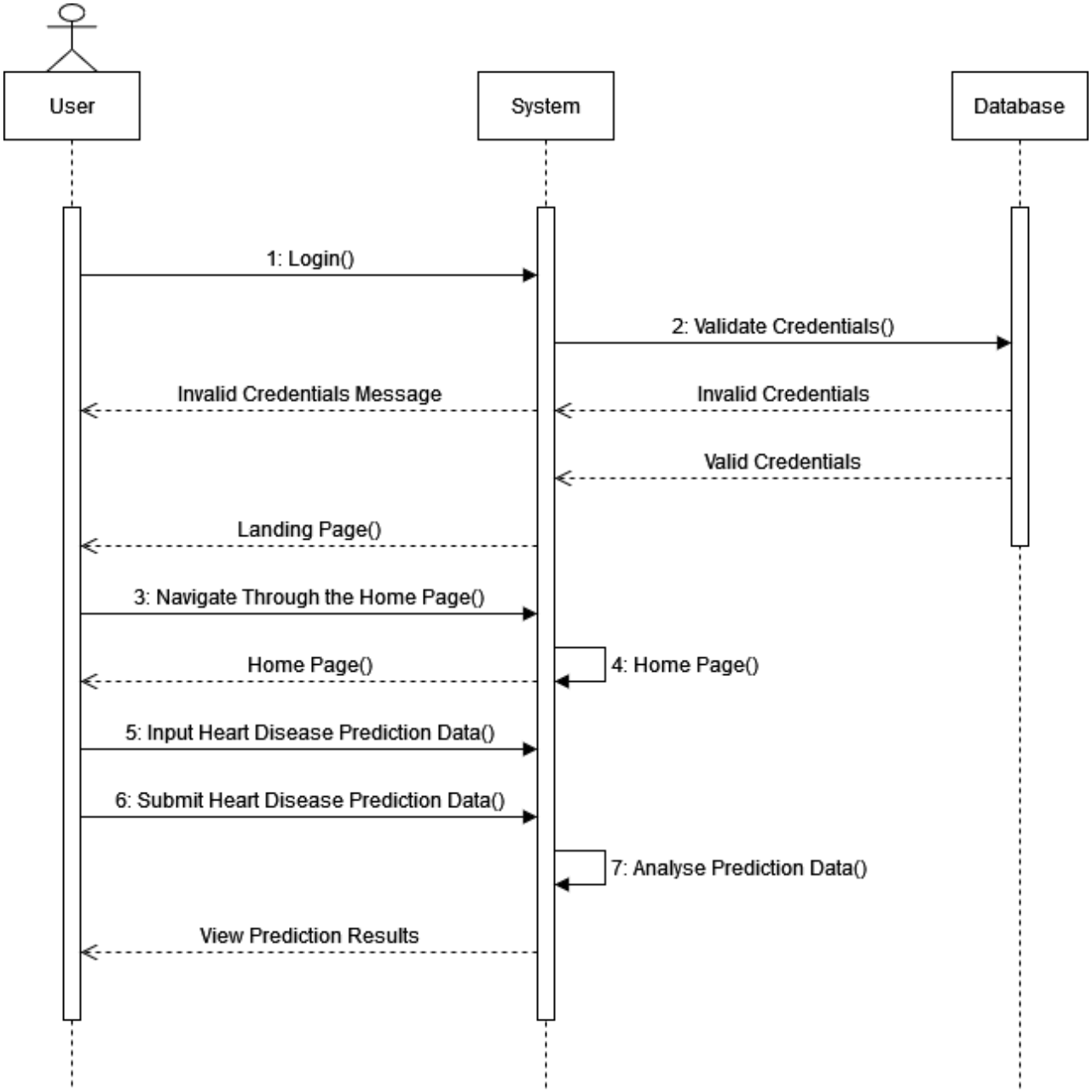


Figure 7: User Management Sequence Diagram

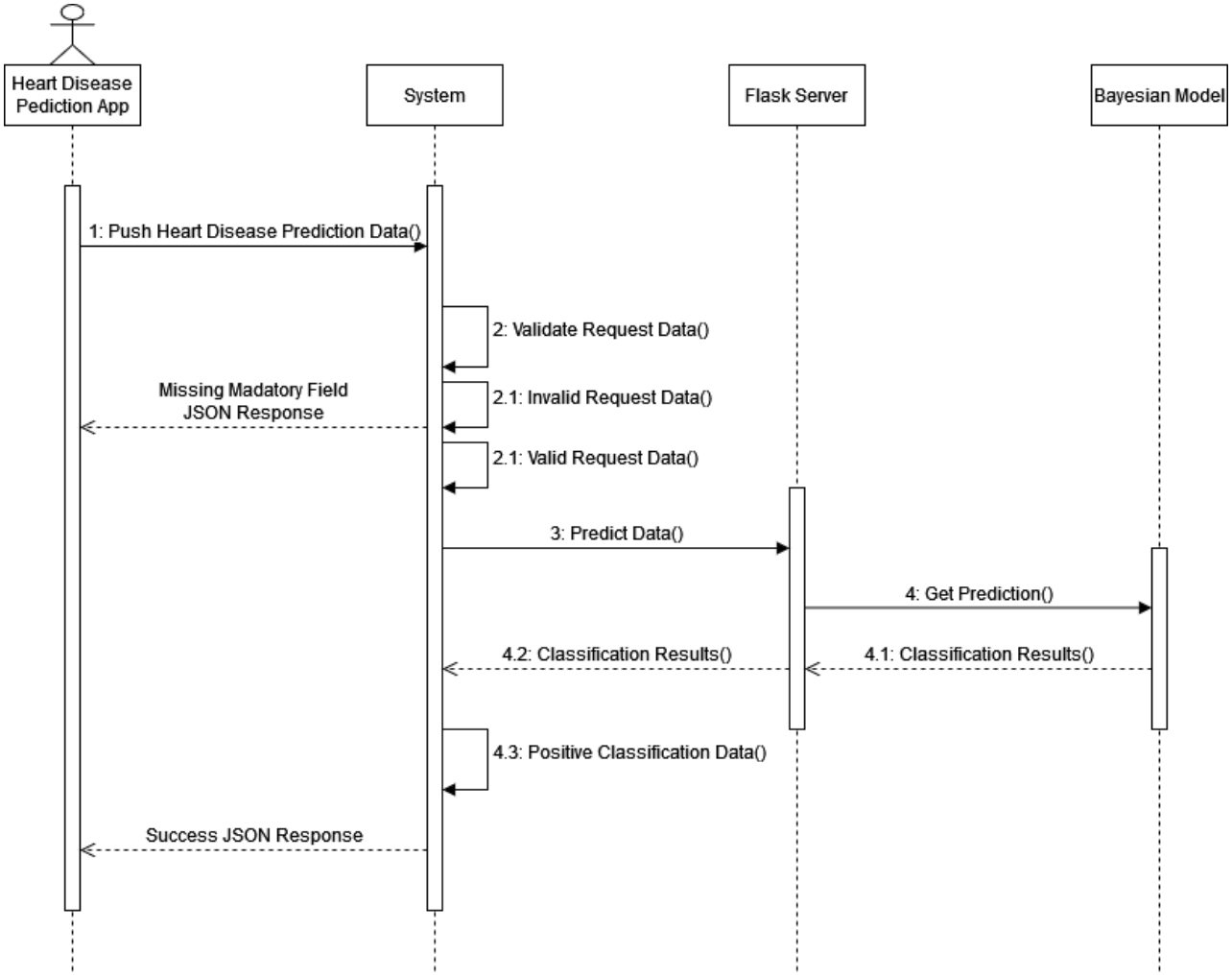


Figure 8: Prediction Sequence Diagram

3.5.2.5 Database Design

Database design is a comprehensive and iterative process that plays a pivotal role in the creation, implementation, and maintenance of an organization's information management systems. It involves a systematic series of steps aimed at developing both physical and logical models for the proposed database system. The ultimate objective is to design a robust, efficient, and scalable database structure that effectively manages and organizes data to meet the organization's requirements [97][98][99].

Key Steps in Database Design:

Requirements Analysis:

The first step in database design involves thorough requirements analysis. This phase involves engaging with stakeholders, understanding the organizational needs, and identifying the data

requirements of various departments or functions. Clear comprehension of user expectations and system objectives lays the foundation for a successful database design.

Conceptual Design:

Following requirements analysis, the conceptual design phase focuses on creating an abstract representation of the data model. This involves defining entities, their relationships, and the constraints that govern them. Entity-Relationship Diagrams (ERDs) are commonly used during this phase to visually represent the high-level structure of the database.

Normalization:

Normalization is a critical process that aims to eliminate data redundancy and dependency. It involves organizing the data to ensure that relationships between tables are well-structured and that the database adheres to a set of rules. Normal forms, such as First Normal Form (1NF), Second Normal Form (2NF), and so on, are applied to achieve a more efficient and maintainable database schema.

Logical Design:

During the logical design phase, the conceptual model is translated into a logical model that can be implemented in a relational database management system (RDBMS). This involves defining tables, attributes, primary keys, foreign keys, and specifying data types. The resulting logical model provides the groundwork for the actual database schema.

Physical Design:

The physical design phase focuses on the implementation details of the logical model. It involves considerations such as indexing, partitioning, storage optimization, and the definition of data structures that will be used by the database management system. The goal is to enhance performance, ensure data integrity, and accommodate the anticipated data volume.

Implementation:

Once the physical design is complete, the database is implemented based on the finalized design. This involves creating tables, relationships, indexes, and other database objects in the chosen database management system. The implementation phase brings the theoretical design into a practical and functioning database.

Testing and Refinement:

Rigorous testing is conducted to ensure that the database meets the specified requirements. This includes functionality testing, performance testing, and validation of data integrity. Based on test results, refinements and optimizations are made to address any issues or discrepancies identified during testing.

Documentation:

Comprehensive documentation is an integral part of the database design process. This includes documentation of the data model, schema definitions, relationships, constraints, and any specific design decisions made. Well-documented databases facilitate easier maintenance, updates, and troubleshooting.

Benefits of Effective Database Design:

Data Integrity:

Well-designed database ensures data accuracy, consistency, and reliability. Constraints and relationships are defined to enforce data integrity rules, reducing the risk of errors and inconsistencies.

Efficient Retrieval and Storage:

Properly designed databases optimize data storage and retrieval. Indexing, partitioning, and other techniques are employed to enhance performance, enabling faster access to information.

Scalability:

A carefully designed database can easily adapt to growing data volumes and evolving business requirements. Scalability is a key consideration to accommodate future expansion without compromising performance.

Maintainability:

Clear documentation and adherence to design best practices enhance the maintainability of the database. This is crucial for ongoing updates, modifications, and troubleshooting.

Security:

Security considerations are integrated into the design process to safeguard sensitive data. Access controls, encryption, and other security measures are implemented to protect the database from unauthorized access.

In essence, database design is a dynamic and collaborative process that requires a deep understanding of organizational needs, effective communication with stakeholders, and adherence to best practices in information management. A well-designed database serves as the backbone of an organization's data infrastructure, supporting efficient data handling, decision-making, and overall business success.

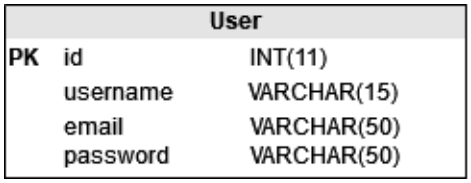


Figure 9: Database Schema

3.6 System Implementation

3.6.1 System Development

The proposed system was developed using an agile system development life cycle. Agile software development, in essence, encompasses a set of system development approaches centered around iterative progress, wherein both requirements and solutions evolve through collaborative efforts among self-organizing cross-functional teams [87][100]. The following diagram in Figure 11 depicts the agile process used in the system development and a rundown of each stage is provided thereafter.

Figure 10: Agile Development Process

Expanding on the Agile Development Life Cycle:

The Agile Development Life Cycle consists of six distinct phases, each playing a crucial role in the successful delivery of a software project. Let's delve deeper into each phase to understand its significance and contributions to the overall development process:

i. **Requirements Gathering:**

In the initial phase, the focus is on understanding the client's needs and expectations. This involves engaging in extensive communication with the client to gather detailed insights into how they envision the system's functionality, appearance, and performance. The goal is to create a comprehensive and accurate set of requirements that will serve as the foundation for the entire development process.

ii. **Design:**

With the requirements in hand, the design phase kicks in, where the architectural blueprint of the software is crafted. This involves creating a robust and scalable framework that aligns with the specified requirements. By establishing a well-defined structure, the design phase aims to prevent potential issues and streamline the development process. It sets the standards and guidelines that will guide the development team throughout the project.

iii. **Development:**

The development phase is where the actual coding takes place. Developers start translating the design into functional code, and data recording mechanisms are implemented in the background. The focus is on building the software incrementally, with regular checks to ensure that each unit is functioning as intended. This iterative approach allows for flexibility and adaptability as the project evolves.

iv. **System Testing:**

The system testing phase is dedicated to thoroughly assessing the software for errors, bugs, and inconsistencies. Various testing methodologies, including unit testing and integration testing, are employed to verify that the system meets the specified requirements. Identifying and rectifying issues at this stage is crucial to delivering a high-quality, reliable product.

v. **Deployment:**

Once the system has successfully passed testing and received approval from both developers and end-users, it is ready for deployment. Deployment involves making the software accessible and operational for its intended users. This phase ensures a seamless transition from development to practical use, marking a crucial milestone in the Agile Development Life Cycle.

vi. **Review:**

The review phase is a reflective stage where the project team evaluates the entire development process. Developers compare the estimated time for tasks with the actual time taken, identifying any discrepancies and reasons for potential delays. This retrospective analysis is

instrumental in improving estimation accuracy for future projects. Additionally, maintenance activities are initiated to ensure that the system remains adaptable and up-to-date with evolving requirements.

In summary, the Agile Development Life Cycle promotes collaboration, adaptability, and continuous improvement. By breaking down the software development process into these six phases, teams can efficiently navigate complex projects, respond to changing requirements, and deliver high-quality software solutions.

3.7 System Testing

Testing a heart disease application in practice involves several key steps to ensure its accuracy, reliability, and effectiveness. Here's a highlight of how the application might be tested:

Unit Testing:

Individual components of the application, such as algorithms for data preprocessing, feature extraction, and prediction models, are tested in isolation to ensure they function correctly. This involves testing various input scenarios and verifying that the expected outputs are produced.

Integration Testing:

Once individual components are tested, they are integrated to form the complete application. Integration testing ensures that different modules work together seamlessly. This involves testing the flow of data and interactions between components to identify and fix any integration issues.

Validation Testing:

This step involves validating the performance of the heart disease prediction model using separate datasets. The model is tested on data that it hasn't seen during training to assess its generalization ability. Performance metrics such as accuracy, sensitivity, specificity, and area under the ROC curve are used to evaluate the model's predictive power.

User Acceptance Testing (UAT):

In UAT, the application is tested by end-users to ensure that it meets their requirements and expectations. This involves real users interacting with the application to perform tasks relevant to their roles. Feedback from users is collected and used to improve the application's usability and functionality.

Cross-Validation:

Cross-validation is a method employed to evaluate the reliability and resilience of the prediction model. It involves partitioning the dataset into several subsets, and then iteratively training and testing the model on various combinations of these subsets. This approach aids in identifying overfitting and ensures that the model performs consistently across diverse data subsets.

Performance Testing:

Performance testing evaluates the application's responsiveness, scalability, and resource usage under different conditions. This involves testing the application with varying levels of workload, data volume, and concurrent users to identify any performance bottlenecks or issues that may affect its reliability in real-world usage scenarios.

Security Testing:

Security testing is conducted to identify and mitigate potential vulnerabilities in the application that could compromise the confidentiality, integrity, or availability of sensitive data. This involves testing for common security threats such as injection attacks, cross-site scripting, and authentication bypass vulnerabilities.

Regulatory Compliance Testing:

Depending on the jurisdiction and intended use of the application, regulatory compliance testing may be required to ensure that the application meets applicable regulations and standards related to healthcare data privacy and security.

By following these testing steps, developers can ensure that the heart disease application is thoroughly evaluated and validated before being deployed for real-world use, minimizing the risk of errors and ensuring the safety and effectiveness of the application for end-users.

3.8 Chapter Summary

The chapter dedicated to methods and materials serves as a critical component in understanding the intricacies of developing the website and implementing the model within the context of the study. This section delves into the systematic processes, tools, and resources employed to create the website, conduct training sessions, and execute the implementation of the model. Furthermore, the chapter sheds light on the meticulously crafted plans that guided the development of essential components crucial to the study's objectives.

4 RESULTS

4.1 Introduction

The chapter discusses the most important research results for the development of an AI-powered medical diagnostic assistance system for heart disease. The main objective of the research was to develop a predictive model leveraging machine learning to forecast heart disease. The chapter presents the results of the Bayesian classification model training and web application development.

4.2 Bayesian Classification Model

The model's training process, as outlined in Chapter 3, laid the foundation for the subsequent presentation and discussion of results in this section. The utilization of the data collection and training methods established a robust framework for training the Bayesian machine learning classifier. The ensuing discussion delves into the features incorporated in the model, the dataset partitioning strategy, and the performance evaluation metrics employed for a comprehensive analysis.

4.2.1 Features and Dataset Partitioning:

The features enlisted in Table 2 serve as the inputs for the Bayesian machine learning classifier. These features, encompassing crucial medical parameters such as blood pressure, sex, cholesterol levels, blood pressure, age, blood sugar, heart rate, and various others, contribute to the model's ability to make predictions regarding the likelihood of heart disease.

The dataset derived from the data collection methods was strategically divided into two subsets: an 80% training dataset and a 20% testing dataset. The training dataset played a pivotal role in imparting knowledge to the model, allowing it to learn and discern patterns from the input features. On the other hand, the testing dataset served as a critical benchmark to assess the model's generalization and predictive capabilities. This partitioning strategy ensures a robust evaluation of the model's performance on unseen data.

4.2.2 Performance Evaluation Metrics:

Performance evaluation hinged on a comprehensive set of metrics to gauge the model's effectiveness. Precision, F1 scores, recall, and precision were among the key measurement systems leveraged to analyse the model's predictive capabilities. Precision, denoting the ratio of true positive predictions to the total predicted positives, assesses the model's accuracy in

identifying positive instances. F1 score, a harmonic mean of precision and recall, provides a balanced measure of a model's overall performance.

Recall, often referred to as sensitivity, and gauges the model's ability to correctly identify positive instances among all actual positives. Precision, conversely, emphasizes the correctness of positive predictions among all predicted positives. These metrics collectively offer a nuanced understanding of the model's strengths and areas that may necessitate refinement.

4.2.3 Bayesian Classification and Probabilistic Learning:

The Bayesian classification approach adopted in this study signifies a probabilistic perspective on learning and inference. Unlike deterministic models, Bayesian classification incorporates probability to express the uncertainty inherent in the relationships learned from the data [101]. This probabilistic framework allows the model to not only make predictions but also quantify the degree of uncertainty associated with each prediction.

Bayesian classification inherently embraces the probabilistic nature of real-world scenarios, acknowledging that data relationships are subject to uncertainty and variability. The model leverages this probabilistic viewpoint to make informed predictions, providing a more nuanced and flexible approach to learning from the dataset.

A limitation of the Bayesian classification model is its assumption of independence among features, which may not always hold true in real-world datasets. Additionally, Bayesian classifiers may struggle with high-dimensional data or datasets with complex relationships between features.

In summary, the training of the model, fuelled by the meticulously collected dataset and robust training methods, culminated in a comprehensive evaluation of its predictive prowess. The presentation and discussion of results illuminated the model's performance through various metrics, emphasizing its capacity to make probabilistic predictions in the realm of heart disease diagnosis. The Bayesian classification approach introduced a layer of sophistication by embracing uncertainty, contributing to a more nuanced understanding of the model's learning and inference capabilities.

$$P(m|k) = P(k|m) * P(m)/P(k) \qquad (1)$$

Equation (1) articulates a fundamental probabilistic relationship that plays a crucial role in the context of test theory, where the focus is on assessing the probability of an event denoted as 'm' given the occurrence of another event represented as 'k.' This equation is expressed as:

$$P(m|k) = P(k \cap m) / P(k)$$

This formulation entails that the likelihood of event 'm' given the occurrence of event 'k' is equal to the joint likelihood of both 'k' and 'm' divided by the probability of 'k.' The interpretation of this equation is central to understanding the conditional probability associated with test theories, where 'm' is the test theory, and 'k' constitutes the evidence or proof linked to 'm.'

Breaking Down the Equation:

i. Probability of k∩m:

This represents the joint probability of both events 'k' and 'm' occurring simultaneously. It signifies the likelihood of the evidence 'k' and the test theory 'm' happening together.

ii. Probability of K:

Denominator of the equation, it denotes the probability of the evidence 'k' occurring. This serves as the normalization factor, ensuring that the conditional probability is appropriately scaled.

iii. Conditional Probability of m|k:

The result of the equation provides the conditional probability of the test theory 'm' given the occurrence of the evidence 'k.' In other words, it quantifies the probability of 'm' being true or valid given the presence of 'k.'

Interpretation in Test Theory:

In the realm of test theory, 'm' typically represents a hypothesis or a testable statement, while 'k' serves as the evidence or proof associated with the hypothesis. The equation captures the essence of how the probability of the hypothesis being true, given the observed evidence, is influenced by the joint probability of both the hypothesis and the evidence.

Application in Bayesian Inference:

Equation (1) aligns with the principles of Bayesian inference, where prior beliefs (probability of 'm') are updated based on new evidence ('k'). The numerator reflects the joint probability of prior beliefs and new evidence, and the denominator ensures the scaling factor for the updated probability.

In conclusion, Equation (1) encapsulates a foundational concept in probability theory and Bayesian inference, particularly relevant in test theory where hypotheses are evaluated based on observed evidence. The equation provides a mathematical framework for understanding and quantifying the conditional probability of a hypothesis given the presence of specific evidence.

4.2.4 Data Analysis

Figure 12 serves as a visual representation of the sample observations gleaned from the dataset, offering a snapshot that encapsulates the essence of the data under consideration. This graphical depiction is instrumental in providing a quick and accessible overview of the dataset, allowing stakeholders, researchers, or practitioners to glean insights into the nature and characteristics of the collected observations.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	TA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	TA	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	TA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	TA	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	TA	150	195	0	Normal	122	N	0.0	Up	0

Figure 11: Collected Observations Sample

Figure 13 provides a visual representation of the distribution of categorical variables, specifically delineating the patients categorized by their gender. This insightful depiction offers a clear overview of the gender distribution within the dataset, unravelling valuable insights into the composition of the patient population.

Key Aspects of Figure 13:

i. **Sex-based Categorization:**

The figure distinctly segregates the dataset into two categories based on sex—male and female. Each category is represented graphically, offering a comparative view of the proportions of male and female patients.

ii. Percentage Representation:

The accompanying information in the caption, "76.38% male and 23.61% female," provides a quantitative breakdown of the distribution. This adds a layer of precision to the visual representation, facilitating a more detailed understanding of the gender composition.

Interpretation and Analysis:

- **Gender Disparity:**

Figure 13 readily highlights the gender disparity within the dataset, showcasing that a predominant percentage of patients are male. The visual contrast aids in swiftly recognizing any imbalances in the representation of male and female individuals.

- **Demographic Overview:**

The visualization serves as a demographic snapshot, enabling a quick assessment of the dataset's composition. This information is vital for contextualizing subsequent analyses and interpretations, particularly in scenarios where gender may be a critical factor.

- **Potential Gender-Related Insights:**

Depending on the nature of the study or analysis, the gender distribution may offer insights into gender-specific patterns or tendencies related to heart disease. Exploring such nuances becomes more feasible with a clear understanding of the dataset's gender composition.

Complementary Analysis:

- **Statistical Testing:**

While Figure 13 provides a visual overview, statistical tests, such as chi-square tests, can be employed to ascertain whether the observed gender distribution is statistically significant. This aids in determining whether any imbalances are indicative of a broader trend or if they could be attributed to random variability.

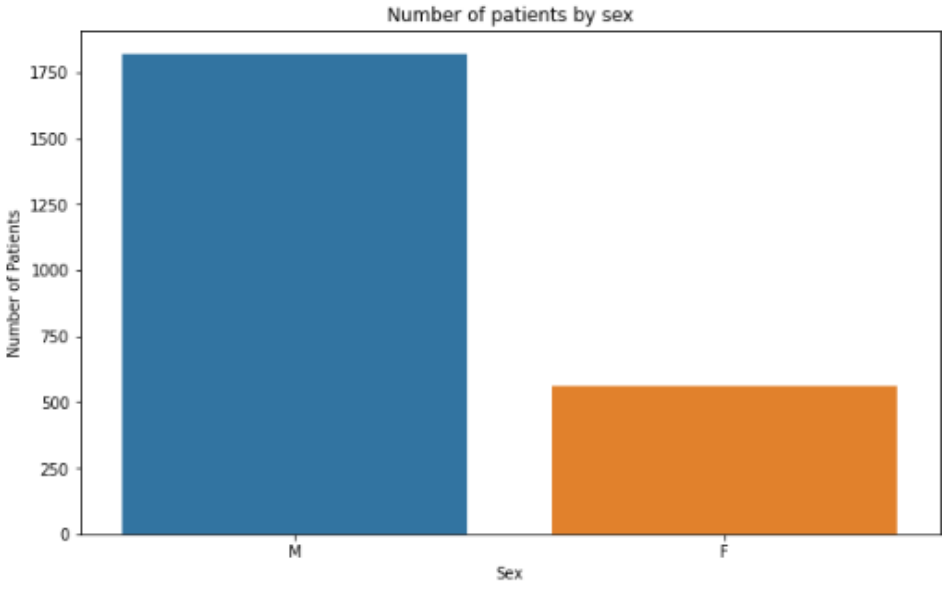
- **Correlation with Other Variables:**

Exploring potential correlations between gender and other relevant variables within the dataset enriches the analysis. For instance, examining whether certain medical parameters vary significantly between male and female patients can unveil nuanced patterns.

Incorporating Context:

Understanding the broader context of the dataset, such as the demographic characteristics of the population under study, is essential for a holistic interpretation of Figure 13. Demographic considerations, such as age distribution and geographical location, may influence the observed gender distribution.

In summary, Figure 13 serves as a valuable visual tool for grasping the gender distribution within the dataset. Its straightforward representation aids in quickly discerning the proportions of male and female patients, paving the way for further analyses and contextual interpretations in the broader context of heart disease research.



Key: M – Male, F - Female

Figure 12: Patients by gender

Figure 14 intricately details the prevalence of heart disease within the dataset, specifically delving into the gender-based distribution of heart disease among male and female patients. This nuanced representation not only elucidates the overall prevalence but also sheds light on potential gender-specific patterns that may be pivotal in the context of heart disease research.

Key Aspects of Figure 14:

i. Gender-Specific Heart Disease Rates:

The figure distinctly categorizes patients based on gender and further delineates the proportion of individuals within each gender category who are diagnosed with heart disease. This provides a clear and targeted understanding of heart disease prevalence.

ii. Percentage Breakdown:

The accompanying information in the caption, "Heart disease is present in 88.87% of male patients, while it affects 11.12% of female patients," quantifies the prevalence rates within each gender. This breakdown enhances the granularity of the visual representation, offering precise insights into gender-specific heart disease occurrences.

Interpretation and Analysis:

- **Gender Disparities in Heart Disease:**

Figure 14 unveils notable gender disparities in heart disease prevalence. The stark contrast between the percentages of male and female patients with heart disease underscores potential gender-specific factors or vulnerabilities related to heart health.

- **Gender-Specific Health Considerations:**

The visualization prompts consideration of gender-specific health factors that may contribute to the observed prevalence rates. This could encompass physiological differences, lifestyle factors, or other variables that may influence the likelihood of heart disease diagnosis.

- **Clinical Implications:**

The insight into the gender-specific distribution of heart disease has clinical implications, guiding healthcare practitioners in tailoring diagnostic and preventive measures. Understanding how heart disease manifests differently among males and females is crucial for personalized and effective healthcare strategies.

Complementary Analysis:

- **Statistical Significance Testing:**

Conducting statistical tests, such as chi-square tests, could ascertain whether the observed differences in heart disease prevalence between male and female patients are statistically significant. This is pivotal for validating the robustness of the observed patterns.

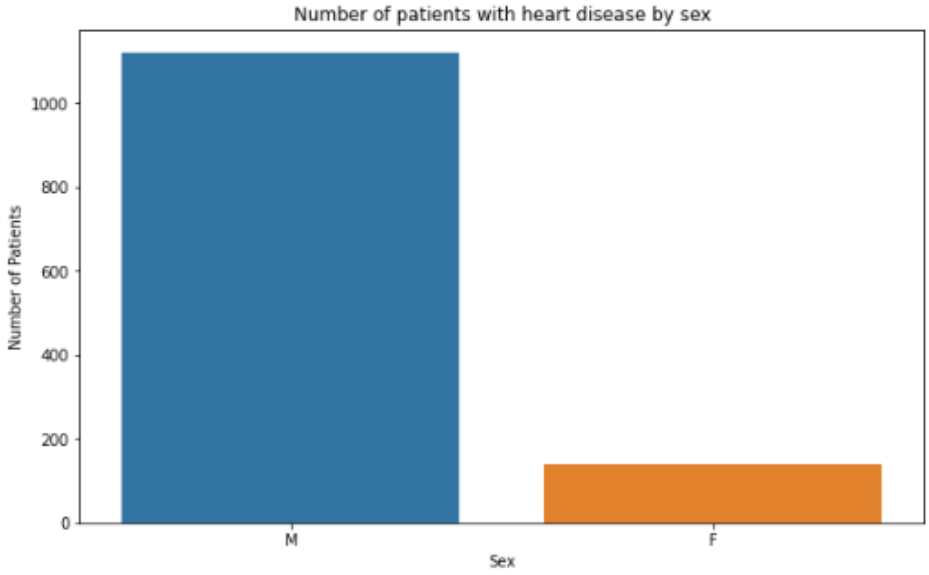
- **Exploration of Contributing Factors:**

Beyond the prevalence rates, delving into potential contributing factors, such as lifestyle choices, genetic predispositions, or socio-economic factors, enriches the analysis. This exploration aids in understanding the multifaceted nature of heart disease occurrence.

Incorporating Context:

Understanding the broader context of the dataset, including demographic factors, regional variations, and other pertinent details, is vital for a comprehensive interpretation of Figure 14. The interplay of these factors may contribute to the observed gender-specific patterns.

In summary, Figure 14 serves as a focal point for comprehending the nuanced relationship between gender and heart disease prevalence within the dataset. Its detailed breakdown facilitates targeted analyses and underscores the importance of gender-specific considerations in the broader landscape of heart disease research and healthcare.



Key: M – Male, F - Female

Figure 13: Patients with heart disease by gender

Figure 15 encapsulates a comprehensive overview of heart disease prevalence within the complete dataset, offering a balanced perspective that considers both the presence and absence of heart disease among the patients under study. This visual representation serves as a pivotal snapshot, shedding light on the overall distribution of heart disease within the broader context of the dataset.

Key Aspects of Figure 15:

i. **Global Heart Disease Prevalence:**

The figure holistically captures the prevalence of heart disease across the entire dataset, presenting a dual categorization of patients based on whether they have heart disease or not. This binary classification lays the foundation for a nuanced understanding of the dataset's overall cardiac health landscape.

ii. **Percentage Breakdown:**

The accompanying information in the caption, "52.85% of patients have heart disease and 47.14% do not have heart disease," quantifies the prevalence rates, providing a clear and concise breakdown of the distribution. This percentage breakdown enhances the interpretability of the visual representation.

Interpretation and Analysis:

- **Global Heart Disease Trends:**

Figure 15 serves as a vital tool for discerning overarching trends in heart disease prevalence across the entire dataset. The balanced representation of both positive and negative cases offers a holistic perspective.

- **Dataset Imbalance:**

The approximate equality in the percentage distribution between patients with and without heart disease suggests a dataset that is relatively balanced in terms of heart disease occurrences. This balance is critical for robust model training and analysis.

- **Baseline Understanding:**

The visualization establishes a baseline understanding of heart disease prevalence, providing a starting point for more nuanced analyses. Understanding the baseline is pivotal for contextualizing subsequent investigations and drawing meaningful comparisons.

Complementary Analysis:

- **Temporal or Demographic Breakdown:**

While Figure 15 offers a global perspective, additional analyses that break down heart disease prevalence over time or across different demographic segments can provide deeper insights. Understanding how these rates vary can uncover dynamic patterns.

- **Risk Factor Exploration:**

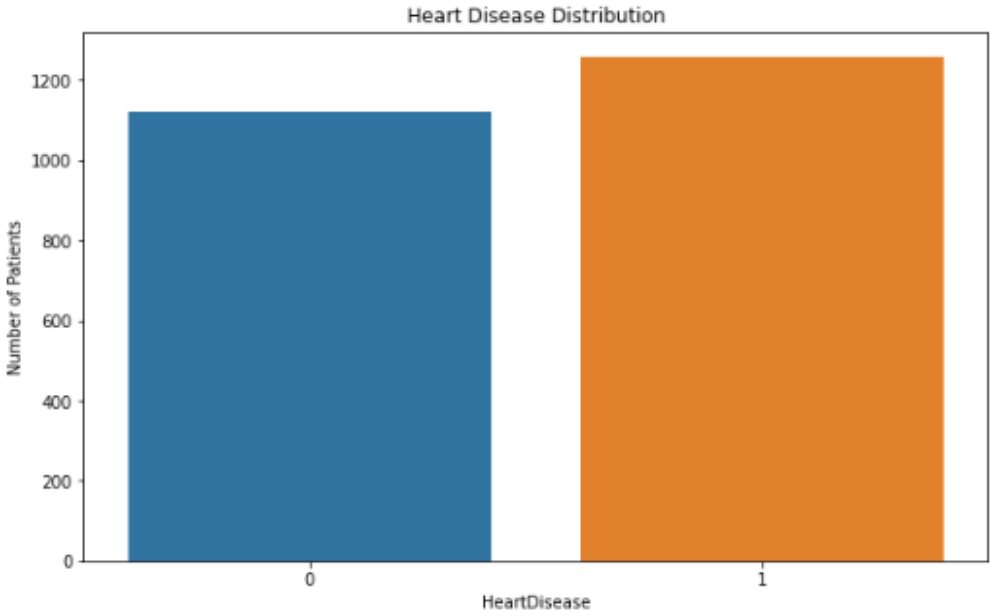
Exploring potential risk factors associated with heart disease within the dataset contributes to a more comprehensive analysis. Identifying correlations between specific features and heart disease prevalence enhances the understanding of contributing factors.

Incorporating Context:

Understanding the broader context of the dataset, including the characteristics of the patient population, geographical considerations, and any temporal variations, enriches the

interpretation of Figure 15. The interplay of these contextual factors can influence the observed prevalence rates.

In summary, Figure 15 serves as a foundational visual representation, offering a balanced portrayal of heart disease prevalence within the complete dataset. Its clarity and straightforward categorization set the stage for more in-depth analyses, allowing researchers and practitioners to delve into the complexities of heart health within the studied population.



Key: 0 – No Heart Disease, 1 – Heart Disease

Figure 14: Heart Disease Distribution

4.2.5 Pre-processing

The pre-processing stage of the data plays a crucial role in ensuring the robustness and effectiveness of machine learning models. In the context of the study, a pivotal step involved the division of the dataset into two distinct categories: the training dataset, encompassing 80% of the data, and the testing dataset, comprising the remaining 20%. This division is integral to the model development process and contributes to the reliability of subsequent predictions and evaluations.

Key Aspects of Data Division:

i. **Training Dataset (80%):**

The training dataset, comprising 80% of the original data, serves as the foundational subset used to train the machine learning model. During this phase, the model learns patterns, relationships, and features inherent in the data, enabling it to make informed predictions when confronted with new, unseen data. The larger proportion allocated to training ensures that the model gains a comprehensive understanding of the dataset's intricacies.

ii. **Testing Dataset (20%):**

The testing dataset, constituting 20% of the original data, operates as an independent subset that the model has not encountered during the training phase. This subset is reserved to assess the model's generalization capability—its ability to accurately predict outcomes on new, unseen data. The testing dataset is crucial for evaluating the model's performance and gauging its effectiveness in real-world scenarios.

Rationale for Data Division:

- **Preventing Overfitting:**

Allocating a substantial portion of the data to the training dataset guards against overfitting, where a model becomes excessively attuned to the training data but struggles to generalize to new data. A well-balanced division mitigates the risk of the model memorizing specific instances rather than learning underlying patterns.

- **Performance Evaluation:**

The testing dataset acts as a benchmark for evaluating the model's performance on unseen data. This external evaluation is essential for assessing the model's efficacy and determining whether it can effectively extrapolate its learning to novel instances.

Implementation of Data Division:

i. **Randomized Split:**

The division of the dataset into training and testing subsets typically involves a randomized process to ensure a representative distribution of data in both categories. Randomization minimizes biases that may arise from specific patterns or sequences present in the original dataset.

ii. **Stratified Sampling:**

In scenarios where maintaining the distribution of classes or categories is critical, stratified sampling may be employed. This technique ensures that both training and testing datasets retain a proportional representation of each class, preserving the overall class distribution.

Ensuring Data Quality:

Prior to division, pre-processing may involve steps such as handling missing values, normalizing features, or addressing outliers. These measures contribute to the overall quality of the data used for model training and testing, enhancing the model's performance and interpretability.

Cross-Validation Consideration:

While the described division is common, in-depth model evaluation may involve techniques such as cross-validation. Cross-validation iteratively partitions the dataset into training and testing subsets, providing a more comprehensive assessment of the model's stability and generalization.

In summary, the division of the dataset into training and testing subsets is a pivotal step in the pre-processing stage, contributing to the integrity and effectiveness of the subsequent machine learning model. The careful allocation of data ensures that the model is adequately trained and rigorously evaluated, laying the foundation for reliable predictions in real-world applications.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
1851	59	M	ASY	126	218	1	Normal	134	N	2.2	Flat
388	53	M	NAP	130	0	0	LVH	135	Y	1.0	Flat
2037	61	F	ASY	145	307	0	LVH	146	Y	1.0	Flat
230	37	F	ATA	130	173	0	ST	184	N	0.0	Up
1892	42	M	ASY	148	244	0	LVH	178	N	0.8	Up

Figure 15: Sample of the training dataset

4.2.6 Initial model training

The implementation and evaluation of the CatBoostClassifier machine learning model for heart disease prediction represent critical steps in gauging the model's performance and reliability. The following expands on the training and testing process, emphasizing the key metrics derived from the evaluation, and elaborates on the insights provided by the confusion matrix depicted in Figure 17.

Training and Testing Process:

i. **CatBoostClassifier Model:**

The choice of the CatBoostClassifier as the machine learning model underscores its suitability for handling categorical features and robust performance in classification tasks. The model was configured with 11 features relevant to heart disease prediction.

ii. **Learning Rate and Iterations:**

The learning rate, set at 0.018447, determines the step size during optimization, influencing how quickly the model adapts to the training data. The model was trained over multiple iterations, with the best test performance achieved at 91.28% accuracy after 3760 iterations. This iterative process ensures that the model refines its understanding of the data with each cycle.

Model Evaluation:

i. **Accuracy:**

The evaluation of the model on the test data yielded an accuracy of 98%. Accuracy represents the overall correctness of the model's predictions, indicating the proportion of correctly classified instances among all instances.

ii. **Precision, F1-Score, Recall, and Support:**

Figure 17 illustrates the confusion matrix, it offers an elaborate overview of the model's effectiveness. Precision assesses the precision of positive predictions, recall evaluates the model's capacity to identify all positive instances, and the F1-score strikes a balance between precision and recall. The support metric signifies the quantity of instances in each class.

Confusion Matrix (Figure 17):

The confusion matrix is a powerful visual representation that encapsulates the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model. These elements contribute to the computation of precision, recall, and the F1-score.

Key Aspects of Figure 17:

i. **True Positives (TP):** Instances where the model correctly predicted the positive class (heart disease) are represented in the top-left quadrant.
ii. **True Negatives (TN):** Instances correctly predicted as the negative class (no heart disease) are found in the bottom-right quadrant.
iii. **False Positives (FP):** Incorrectly predicted positive instances are located in the top-right quadrant.
iv. **False Negatives (FN):** Instances wrongly classified as negative instances (missed heart disease predictions) are in the bottom-left quadrant.

Interpretation of Confusion Matrix Metrics:

- **Precision (Positive Predictive Value):**

Precision is the ratio of true positives to the total number of predicted positives. A high precision score indicates a low rate of false positives.

- **Recall (Sensitivity or True Positive Rate):**

Recall represents the ratio of true positives to the total number of actual positives. High recall signifies effective capturing of positive instances.

- **F1-Score:**

The F1-score offers a combined measure of precision and recall, taking into account both false positives and false negatives.

- **Support:**

Support indicates the number of instances in each class, offering context to the precision, recall, and F1-score metrics.

Implications and Further Analysis:

- **Model Robustness:**

The high accuracy and performance metrics in the confusion matrix suggest that the CatBoostClassifier model demonstrates robustness in predicting heart disease.

- **Fine-Tuning Opportunities:**

The confusion matrix metrics provide insights into areas where the model may benefit from fine-tuning, such as minimizing false positives or false negatives, depending on the specific application and consequences of prediction errors.

- **Clinical Relevance:**

Translating these findings into a clinical context involves considering the implications of false positives and false negatives, as well as exploring additional domain-specific metrics.

In summary, the training and evaluation of the CatBoostClassifier model showcase its efficacy in heart disease prediction. The rich information provided by the confusion matrix aids in understanding the nuances of the model's performance, facilitating informed decisions for further refinement or deployment in real-world healthcare scenarios.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	226
1	0.98	0.98	0.98	250
accuracy			0.98	476
macro avg	0.98	0.98	0.98	476
weighted avg	0.98	0.98	0.98	476

Figure 16: Confusion matrix

4.2.7 Hyperparameter tuning

Hyperparameters play a pivotal role in machine learning models, as they govern the training process and influence the behavior of the model. The training process involves configuring the model based on the training data, adjusting its parameters, and optimizing hyperparameters to enhance overall performance.

Components of the Training Process:

i. **Training Data:**

The training data forms the foundation for configuring the model. It consists of instances used to teach the model patterns, relationships, and features, enabling it to make accurate predictions on new, unseen data.

Notably, the values in the training data do not become direct components of the model but serve as the basis for model learning.

ii. **Model Parameters:**

Model parameters are the internal variables that the chosen machine learning technique adapts during training. These parameters are adjusted to align the model with the patterns present in the training data.

iii. **Hyperparameters:**

Hyperparameters govern the training process itself. They remain constant during a training job and influence how the model learns from the training data.

Examples of hyperparameters include learning rates, regularization factors, and the number of iterations.

Hyperparameter Tuning:

Hyperparameter tuning involves finding the optimal configuration for these governing variables to enhance the model's performance. This process often employs techniques like grid search or random search to explore the hyperparameter space and identify the most effective combination.

Results of Hyperparameter Tuning:

i. **Number of Trials:**

The training process involved conducting 11 trials, each representing a different configuration of hyperparameters.

ii. **Best Trail (Trail 5):**

Among the 11 trials, Trail 5 emerged as the most effective, achieving a test accuracy value of 99%. This indicates that the hyperparameter configuration in Trail 5 resulted in superior model performance on unseen data.

iii. **Best Test Value and Iteration:**

The best test value, denoted as 0.0849, and the best iteration, set at 999, are key indicators of the model's efficiency. These values represent the optimal balance achieved during training.

Final Model Evaluation:

i. **Test Data Prediction Accuracy:**

After the final model was trained with the refined hyperparameters, the accuracy prediction of the test data reached an impressive 99.07%. This reflects the model's ability to generalize well to new, unseen data.

ii. **Training Data Prediction Accuracy:**

The accuracy prediction of the training data 99.97% underscores the model's proficiency in capturing patterns from the training dataset.

Implications and Insights:

i. **Optimized Training Process:**

The hyperparameter tuning process resulted in a well-optimized training process, leading to a highly accurate model.

ii. **Model Generalization:**

The high test accuracy suggests that the model generalizes effectively to new instances, demonstrating its robustness in real-world applications.

iii. **Performance Consistency:**

The consistency between training and test accuracy indicates that the model's performance is reliable and not overfitting to the training data.

In summary, hyperparameters play a critical role in shaping the training process, and the optimization achieved through hyperparameter tuning has resulted in a highly accurate and

reliable machine learning model for heart disease prediction. The findings underscore the importance of meticulous tuning to enhance model efficacy and generalization.

4.3 Web Application Stimulation Results

1.3.1. Addressing Heart Disease Challenges in Zambia:

Heart diseases continue to pose a significant threat to public health in Zambia, emerging as one of the leading causes of mortality. This pressing issue can be largely attributed to the absence of advanced software tools that can complement healthcare professionals efforts in predicting and diagnosing heart diseases promptly.

1.3.2. Research Initiatives in AI-Based Predictive Mechanisms:

Recognizing the urgent need for improved diagnostic techniques, various scholars and researchers, including Kaur et al. [1], Goma et al. [2], Mirzajani et al. [3], Kumar et al. [4], Enriko et al. [5], Wiharto et al. [6], Zagorecki et al. [26], and Repaka et al. [27], have undertaken commendable efforts to develop AI-based predictive mechanisms. These initiatives leverage deep learning and machine learning techniques to enhance the accuracy and efficiency of heart disease diagnosis.

However, a critical evaluation of existing research reveals certain limitations. Many of these studies either rely on a limited dataset type or lack a seamless integration into a clinical setting. These shortcomings underscore the necessity for a more robust and clinically applicable solution to address the unique challenges posed by heart diseases in Zambia.

1.3.3. Study Objectives and Purpose:

The primary aim of this study is to Create and implement a computer-assisted diagnostic system employing artificial intelligence (AI) methodologies. This system is envisioned to play a pivotal role in mitigating the death rate associated with heart diseases by providing crucial decision support to medical practitioners. The emphasis is placed on facilitating early diagnosis and prompt initiation of treatment protocols.

1.3.4. Proposed Design and Implementation Analysis:

In the subsequent sections, the study delves into a comprehensive analysis of the design and implementation phases of the prediction system for heart disease. Figures 18 to 21 illustrate the intricacies of the proposed system, emphasizing its ability to monitor and analyze features logged during the diagnostic process. The ultimate output of the system is the probability of heart disease based on the extracted features.

1.3.5. Key Components of the Proposed System:

Data Monitoring and Feature Extraction:

The system actively monitors relevant features, capturing crucial data points during the diagnostic process.

Features logged include critical parameters such as age, exercise-induced angina, sex, blood sugar levels, cholesterol levels, blood pressure, heart rate, resting electrocardiogram results, chest pain type, ST-slope, and oldpeak.

Artificial Intelligence Techniques:

Leveraging advanced AI techniques, the system employs machine learning algorithms for predictive analysis.

These algorithms are designed to process the logged features and generate a probability assessment for the presence of heart disease.

Clinical Applicability:

A pivotal aspect of the system's design is its seamless integration into clinical settings, ensuring practical utility for healthcare professionals.

The system aims to enhance the diagnostic capabilities of medical practitioners, enabling them to make informed decisions in a timely manner.

1.3.6. Conclusion:

The envisioned computer-assisted diagnosis system represents a significant step forward in the quest to combat heart diseases in Zambia. By addressing the limitations of existing approaches,

this study strives to contribute a clinically relevant and impactful solution that aligns with the unique healthcare landscape of the region. The subsequent sections will delve into a detailed exploration of the proposed system's design and implementation, shedding light on its potential to revolutionize heart disease diagnosis and treatment strategies.

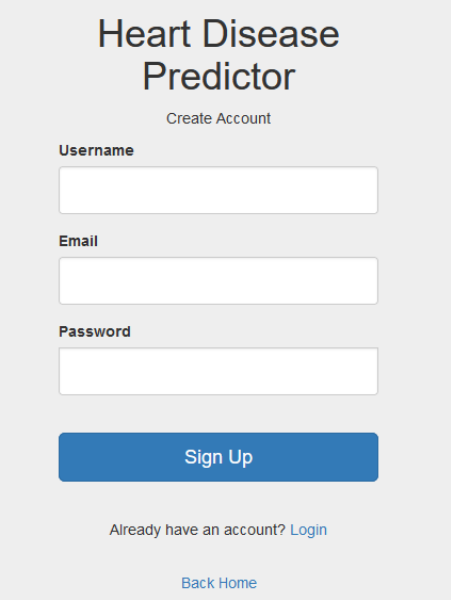


Figure 17: Medical professionals are required to input information to establish an account.

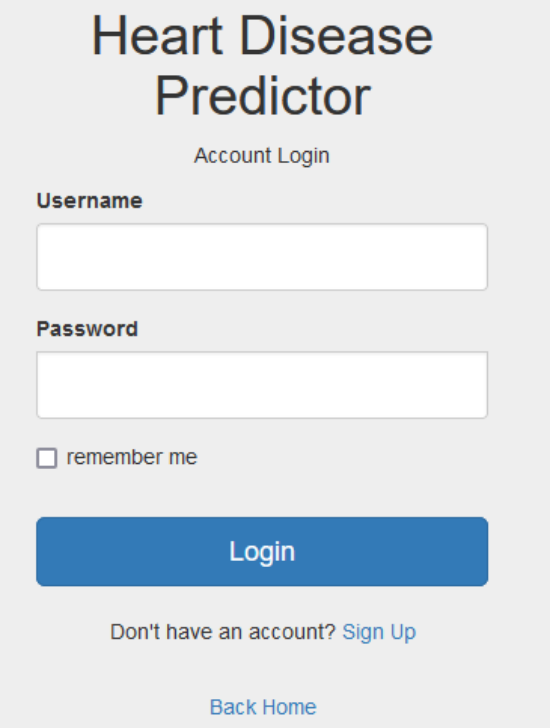


Figure 18: Medical professionals need to input their password and username to access the login system

Figure 19: Medical professionals are required to input health information such as blood pressure, age, heart rate etc., and then initiate the prediction process by clicking on the 'run model' button

Figure 20: Medical professionals have the ability to view the likelihood of a patient having heart disease or not

4.4 Inference Statistics

Scenario: An evaluation is conducted to assess the performance of an AI-powered medical diagnostic assistance system for Heart Disease. The system processes patient data and provides predictions regarding the likelihood of heart disease.

Inference Statistics:

i. **Descriptive Statistics:**

Table 4: Descriptive Statistics

Statistics	Values
Number of Cases Analyzed	n=1190
Cases with Heart Disease	Cases=52.85%
Cases without Heart Disease	Cases=47.14%
Accuracy of the AI Diagnosis System (Test Data)	Accuracy=99.07%
Accuracy of the AI Diagnosis System (Training Data)	Accuracy=99.97%
Precision	Precision=98%
F-measure	F-measure=0.98

4.5 Chapter Summary

This chapter presents the results of the study. Bayesian Classification training results were presented, and the application of the generated Bayesian classification model was presented. The chapter also covered the implementation of a web application.

5 DISCUSSION AND CONCLUSIONS

5.1 Introduction

The study questions from the first chapter are addressed in this chapter. The findings analyses the conclusions and responses to the study's questions. The chapter also offers suggestions for on heart disease predictions using Zambian data.

5.2 Discussion

This section discusses the finding to answer the research questions developed in the first chapter.

5.2.1 Objective 1 Discussion

Developing a Machine Learning Model for Heart Disease Prediction:

The primary research question aimed at developing a model using machine learning to assist in heart disease prediction. In response to this imperative query, the researcher undertook a multifaceted approach, employing advanced computational tools and innovative techniques to craft a robust predictive model. The pivotal steps involved in addressing this research question are elucidated below:

i. Web-Based Interactive Computing Platform:

To embark on the journey of model development, the researcher harnessed the power of a web-based interactive computing platform. Specifically, the Python programming language coupled with Jupyter Notebook provided the ideal ecosystem for designing, implementing, and refining the heart disease prediction model.

Jupyter Notebook Utilization:

Jupyter Notebook, known for its versatility and interactivity, served as the primary coding environment. Its integration with Python facilitated the seamless execution of code snippets, allowing for a dynamic and iterative model development process.

Python Programming Language:

Python's prominence in the field of machine learning and data science made it the language of choice. Its extensive libraries, including but not limited to NumPy, Pandas, and Scikit-Learn, provided a comprehensive toolkit for data manipulation, analysis, and model implementation.

ii. Training, Testing, and Model Creation:

The development lifecycle of the machine learning model unfolded through distinct phases, each contributing to the model's refinement and predictive efficacy.

Dataset Preparation:

The foundation of any machine learning model lies in the quality and diversity of the dataset. The researcher curated a dataset encompassing relevant features associated with heart health, ensuring a representative and comprehensive sample.

Data Preprocessing:

Before model training commenced, the dataset underwent meticulous preprocessing. This involved handling missing values, normalizing data, and addressing any anomalies that could impact the model's learning process.

Training and Testing Split:

A fundamental step involved partitioning the dataset into two subsets: a training dataset (80%) and a testing dataset (20%). This segregation enabled the model to learn patterns from the training data and assess its performance on unseen test data.

Machine Learning Algorithm Selection:

The choice of a machine learning algorithm is pivotal in determining the model's predictive prowess. The researcher opted for the Bayesian Classification algorithm, a probabilistic approach known for its efficacy in handling uncertainty.

Model Iteration and Optimization:

Model development is an iterative process. The researcher fine-tuned parameters, adjusted hyperparameters, and iterated on the model architecture to enhance its predictive accuracy.

iii. Model Evaluation and Validation:

The efficacy of the developed model was rigorously evaluated using various metrics to ensure its reliability and generalization capabilities.

Performance Metrics:

Metrics such as precision, recall, F1 score, and accuracy were calculated to gauge the model's performance across different dimensions. These metrics provided a nuanced understanding of the model's strengths and areas for improvement.

Validation on Testing Data:

The ultimate litmus test for the model was its performance on the testing dataset. This phase validated the model's ability to make accurate predictions on new and unseen data, ensuring its applicability beyond the training set.

In essence, the researcher's approach to addressing the first research question involved a meticulous blend of computational tools, algorithmic selection, and iterative refinement. The utilization of Jupyter Notebook, coupled with Python, provided a dynamic environment for crafting an advanced machine learning model poised to contribute significantly to heart disease prediction.

5.2.2 Objective 2 Discussion

Utilizing Bayesian Classification for Heart Disease Prediction:

In addressing the research question concerning heart disease prediction, the researcher employed a Bayesian-based classification model. The choice of Bayesian classification stems from its probabilistic approach to learning and inference, offering a distinctive perspective on uncertainty in the context of data-driven learning [101]. The comprehensive methodology employed to develop and assess the Bayesian classification model is elucidated below:

i. Bayesian Classification Model:

The Bayesian classification model forms the cornerstone of the predictive framework. Bayesian methods utilize probability to encapsulate the uncertainty inherent in the relationships learned from data. This nuanced approach aligns well with the intricate nature of heart disease prediction, where uncertainties and complexities abound.

ii. Features as Input to the Classifier:

The success of the Bayesian classification model hinges on the quality and relevance of the features incorporated into the predictive system. The researcher curated a dataset comprising pertinent features associated with heart health. These features, ranging from demographic information to physiological parameters, served as the input variables for the Bayesian machine learning classifier.

iii. Dataset Division for Training and Testing:

To facilitate effective model development and evaluation, the dataset was bifurcated into two subsets: the training dataset and the testing dataset. The training dataset, constituting 80% of the data, was instrumental in imparting knowledge to the model, while the testing dataset (20%) gauged the model's performance on unseen data.

iv. Performance Measurement:

Performance evaluation was conducted through a robust set of metrics, including precision, recall, F1-scores, and accuracy. These metrics collectively provided a comprehensive view of the model's effectiveness in making accurate predictions. Precision reflected the model's ability to minimize false positives, recall gauged its capability to capture true positives, and F1-score provided a balance between precision and recall.

v. Iterative Training Process:

The development of the Bayesian classification model involved an iterative training process. Eleven trials were conducted to fine-tune the model, with the fifth trial emerging as the optimal configuration, achieving a remarkable accuracy of 99%. The iterative nature of the training

process underscores the commitment to refining the model for optimal predictive performance.

vi. Final Model Parameters:

The culmination of the training process led to the identification of optimal parameters for the Bayesian classification model. The best test value, recorded at 0.0849, and the corresponding best iteration, reaching 999, epitomize the precision and granularity achieved through the iterative trials.

vii. Prediction Accuracy:

Post-training, the final model, armed with the refined parameters, underwent testing on new and unseen data. The accuracy of the model in predicting heart disease on the test dataset reached an impressive 99.07%. Moreover, the model demonstrated exceptional accuracy on the training data, recording an accuracy rate of 99.97%.

In summary, the researcher's utilization of the Bayesian classification model for heart disease prediction reflects a meticulous and thorough approach. The emphasis on iterative refinement, performance metrics, and optimal parameter selection underscores the commitment to developing a highly accurate and reliable predictive system for heart disease diagnosis.

5.2.3 Objective 3 Discussion

Evaluating Model Accuracy on Zambian Patients:

The final research question sought to gauge the accuracy of the developed model in predicting heart disease specifically for Zambian patients. This objective was accomplished through the implementation of a web-based application, leveraging Python programming, to assess the model's performance with real-world data obtained from medical practitioners at the National Heart Hospital in Lusaka, Zambia. The meticulous process and notable outcomes are detailed below:

i. Web-Based Application Development:

To execute Objective 3 effectively, a web-based application was meticulously crafted using Python. This application served as the operational interface for testing the Bayesian

classification model on Zambian patients' data. The application was designed to seamlessly integrate the model and facilitate accurate predictions based on the unique characteristics of the Zambian patient dataset.

ii. Data Collection from National Heart Hospital:

Real-world data for Zambian patients were procured from medical practitioners at the National Heart Hospital in Lusaka, Zambia. This dataset, reflective of the local context and patient demographics, was instrumental in evaluating the model's adaptability and accuracy within the Zambian healthcare landscape.

iii. Difference between the Kaggle and Zambian dataset:

The main difference between the Kaggle heart disease dataset and a Zambian heart disease dataset primarily lies in their source and possibly their specific attributes. Let's compare them:

Source:

Kaggle Heart Disease Dataset: This dataset is sourced from Kaggle, a platform for data science competitions and collaboration. It might have been compiled from various sources, potentially including medical institutions, research studies, or public health databases, and made available for analysis and modeling.

Zambian Heart Disease Dataset: A Zambian heart disease dataset, on the other hand, is sourced from healthcare institutions, research conducted within Zambia, or public health records specific to Zambia. It focuses specifically on heart disease cases within the Zambian population.

Scope and Attributes:

Kaggle Heart Disease Dataset: This dataset typically contains a standardized set of attributes commonly used in heart disease prediction models. These attributes often include demographic information (such as age and sex), physiological measurements (such as blood pressure and cholesterol levels), and possibly diagnostic test results (such as electrocardiogram findings). The dataset might be curated to represent a diverse population but may not specifically

represent Zambian demographics or health profiles. Also measurements for some attributes like blood sugar and cholesterol differ in how they are measured in Zambia.

Zambian Heart Disease Dataset: A dataset focusing on heart disease within Zambia include similar attributes as the Kaggle dataset, but with a specific emphasis on the Zambian population. It includes additional factors relevant to Zambia, such as socioeconomic status, dietary habits, prevalence of specific risk factors (like hypertension or diabetes), access to healthcare services, and possibly genetic predispositions that are more prevalent within the Zambian population.

Availability and Context:

Kaggle Heart Disease Dataset: Being available on Kaggle, this dataset is accessible to a wide audience of data scientists, researchers, and enthusiasts for analysis, modeling, and potentially for participation in data science competitions. Its context is typically broader, not specific to any particular region or population.

Zambian Heart Disease Dataset: A dataset focusing on heart disease within Zambia has a narrower context, specifically targeting the healthcare challenges and realities within Zambia. It may be used for research within Zambia or for international comparisons with other datasets, contributing to the understanding of heart disease within the Zambian population.

In summary, while both datasets might share similarities in terms of their focus on heart disease, their source, attributes, and context would differ based on whether it's a dataset sourced from Kaggle or one specifically focusing on heart disease cases within Zambia.

iv. Sample Data and Prediction Accuracy using the Zambian dataset:

A subset of the Zambian patient dataset, consisting of 102 patients, was utilized for testing the Bayesian classification model. The model's predictive accuracy was assessed by comparing its predictions against the actual outcomes for these patients. The results revealed a commendable prediction accuracy of 89%.

v. Model Generalizability and Adaptability:

The success of achieving an accuracy rate of 89% underscores the model's generalizability and adaptability to the unique characteristics of Zambian patients. The incorporation of real-world data from the National Heart Hospital facilitated a robust evaluation, affirming the model's efficacy in a specific geographical and demographic context.

vi. Implications for Clinical Decision Support:

The high prediction accuracy achieved in the Zambian patient context holds significant implications for clinical decision support. The model's reliability in accurately predicting heart disease in a local setting enhances its utility as a valuable tool for medical practitioners, contributing to timely diagnosis and intervention.

vii. Continuous Monitoring and Refinement:

The development of the web-based application and the subsequent evaluation of the model on Zambian patients mark a pivotal stage in the continuous monitoring and refinement of the predictive system. Ongoing efforts to collect additional data and monitor the model's performance in real-world scenarios contribute to its continual improvement and adaptability.

In conclusion, the meticulous integration of the Bayesian classification model into a web-based application, coupled with its successful evaluation on Zambian patient data, signifies a pivotal milestone in the study. The achieved prediction accuracy of 89% reflects the model's efficacy and potential for practical application in the clinical realm, offering valuable insights for heart disease diagnosis in the Zambian healthcare landscape.

viii. Integration and System Test Case

Integration and system test cases are specific scenarios or sets of conditions that are designed to validate the proper functioning of software applications, particularly during the development and testing phases. Both integration testing and system testing are crucial components of the software testing life cycle and help ensure the reliability and correctness of a software system.

Integration Test Case:

Integration testing focuses on verifying that individual software modules or components work together as intended when integrated into larger subsystems or the entire system. Integration test cases aim to identify issues related to the interactions between different parts of the software. Here's an overview of what an integration test case might involve:

i. **Objective:** Ensure the seamless collaboration of integrated components.
ii. **Scope:** Test interactions between two or more software modules or subsystems.
iii. **Test Steps:**

- Input data into one module and check the output in another.
- Verify the correct flow of data between interconnected components.
- Test error handling and exception scenarios during data exchange.
- Confirm that interfaces between modules adhere to specifications.
- Assess the impact of changes in one module on the overall system.

System Test Case:

System testing involves evaluating the entire software system as a whole. It is concerned with validating that the system meets the specified requirements and functions as intended in its entirety. System test cases address end-to-end scenarios, ensuring that all components work together harmoniously. Here's an overview of what a system test case might involve:

i. **Objective:** Confirm the overall functionality, performance, and reliability of the complete software system.
ii. **Scope:** Test the system in its entirety, including all integrated components and external interfaces.
iii. **Test Steps:**

- Input data through the user interface or external interfaces.
- Verify that data processing, calculations, and decision-making are accurate.
- Evaluate system response times under normal and peak loads.
- Test the system's ability to handle different inputs and use cases.
- Assess error handling, security features, and recovery mechanisms.
- Confirm compliance with specified requirements and user expectations.

In summary, integration test cases focus on the interactions between integrated components, ensuring they work together correctly. On the other hand, system test cases evaluate the overall functionality of the entire software system, including its interactions with external elements. Both types of test cases play a crucial role in identifying and addressing issues at different levels of the software development process. Table 4 shows the integration and System Test Case for the developed software system.

Table 5: Integration and System Test Case

ID	Test Type	Test Steps	Expected Results	Status
1	Integration Test	Data Input	- Input data is accurately captured and processed.	Pass
2	Integration Test	Model Training	- Bayesian classification algorithm is successfully applied.	Pass
			- Model is trained with the provided dataset.	Pass
3	System Test	Prediction Accuracy	- System generates accurate probability of heart disease.	Pass
			- Predicted probability matches expected outcomes.	Pass
4	System Test	Real-time Monitoring	- System consistently monitors logged features.	Pass
			- Probability of heart disease is dynamically updated based on input.	Pass
5	Integration Test	System Components	- All system components are effectively integrated.	Pass
			- Proper communication and data flow between modules.	Pass
6	System Test	User Interface	- System provides clear and accurate feedback on predicted probability.	Pass
			- User interface is responsive and user-friendly.	Pass
7	Integration Test	Error Handling	- System appropriately handles and reports errors.	Pass
			- Overall system remains robust.	Pass

Cost Benefit Analysis of the Heart Disease Prediction System

The cost-benefit analysis of an AI heart disease prediction system for healthcare providers and patients involves evaluating the financial costs and potential benefits associated with its implementation and utilization.

For healthcare providers, the costs may include initial investment in acquiring the AI system, infrastructure setup, staff training, maintenance, and ongoing support. Additionally, there may be costs related to data storage and security measures to protect patient information. However, the benefits for healthcare providers can be significant. These may include improved accuracy and efficiency in diagnosing heart disease, reduced workload for medical staff, better allocation of resources, and potentially lower costs associated with preventable heart-related complications through early detection and intervention.

For patients, the costs may include expenses related to accessing the AI system, such as co-payments or insurance deductibles, as well as concerns about data privacy and security. However, the benefits for patients can be substantial as well. These may include earlier detection of heart disease, personalized risk assessment and treatment plans, improved health outcomes, and potentially reduced healthcare costs over time through preventive measures and timely interventions.

Overall, the cost-benefit analysis of an AI heart disease prediction system should weigh the initial financial investment against the potential long-term benefits for both healthcare providers and patients, including improved efficiency, accuracy, and outcomes in heart disease management and treatment.

5.3 Conclusions

The overarching goal of this study was to introduce a transformative Machine Learning system designed to assist healthcare professionals, specifically medical practitioners, in the timely and accurate prediction of heart diseases in Zambia. Focusing on the renowned National Heart Disease Hospital situated in Lusaka, Zambia, the study aimed to address the alarming mortality rates associated with heart diseases in the country. To accomplish this, the development of a web application emerged as a crucial step in streamlining the prediction process for enhanced efficiency and accuracy.

Rationale for the Web Application:

The study's rationale stemmed from the imperative need to combat the high mortality rate attributed to heart diseases in Zambia. Recognizing the pivotal role of predictive technologies in mitigating this health challenge, the decision to create a web application became paramount. The application served as a dynamic platform to facilitate precise predictions of heart diseases, offering valuable decision support to medical practitioners and enabling timely interventions.

Designing an Assisted Medical Diagnostic System:

The studies core aim was to craft an assisted medical diagnostic system tailored for predicting heart diseases. Leveraging the Bayesian classification approach, the system utilized a curated set of 11 features sourced from the Kaggle repository dataset. This comprehensive dataset formed the foundation for training the model and fine-tuning its parameters to optimize accuracy.

Training Data Accuracy and Model Performance:

The study's success was underscored by the remarkable accuracy achieved during the training phase. The model, fueled by new parameters, exhibited an exceptional accuracy score of 99.97% when applied to the training data. This outcome showcased the robustness of the machine learning model in capturing patterns and relationships within the data.

Test Data Validation and Continued Improvement:

The validation process, conducted on test data, yielded a notable accuracy of 99.07%, affirming the model's proficiency in generalizing predictions beyond the training set. While these results were impressive, the study acknowledged the importance of continuous improvement and refinement. The pursuit of additional data from the National Heart Hospital aimed to enhance prediction accuracy further, underlining a commitment to ongoing development.

Application to Real-World Scenarios:

The culmination of the study's efforts was the utilization of the developed model on sample data collected from the National Heart Hospital. The achieved prediction accuracy of 89% in this real-world scenario demonstrated the practical applicability of the machine learning

system. It laid the groundwork for future endeavors, emphasizing the potential impact on clinical decision-making in a healthcare setting.

Anticipated Future Enhancements:

The study concluded with an optimistic outlook for the future. Acknowledging the potential for improved accuracy with an expanded dataset from the National Heart Hospital, the study hinted at ongoing efforts to refine the predictive model continually. This forward-looking approach aligns with the dynamic nature of healthcare data and the evolving landscape of predictive technologies.

In essence, the study's purpose was not only met but exceeded expectations by delivering a robust, accurate, and practical machine learning system tailored for heart disease prediction in Zambia. The envisioned impact on healthcare outcomes and the commitment to ongoing refinement underscore the study's significance in contributing to the advancement of predictive healthcare technologies in the Zambian context

5.4 Recommendations

Recommendations for Future Research and Implementation:

i. Exploration of Additional Features:

The study suggests a future avenue for research that involves the incorporation of more features into the Zambian dataset. By expanding the set of features available for analysis, researchers can explore the potential of various machine learning models to predict heart disease with increased precision and reliability. This recommendation aligns with the evolving landscape of healthcare data and the continuous quest for improved predictive capabilities.

ii. Augmentation of Data Samples:

To enhance the predictive accuracy of the model, the study advocates for the collection of more extensive data samples from heart disease patients in Zambia. A larger and more diverse dataset can contribute to a more robust and nuanced understanding of the factors influencing heart diseases in the local context. This recommendation underscores the importance of continually enriching the dataset to bolster the effectiveness of predictive models.

iii. Development of a Comprehensive System Model:

The study proposes the development of a system model that goes beyond predicting the occurrence of heart disease. This future model aims to provide accurate forecasts regarding the specific type and severity of cardiac illnesses. Such a comprehensive system would offer invaluable insights to healthcare professionals, enabling tailored treatment plans and interventions based on a nuanced understanding of the disease spectrum.

iv. Extension to Predict Other Diseases:

Building on the success of the heart disease prediction model, the study encourages the extension of developed models and tools to predict a broader range of diseases prevalent in Zambia. By leveraging the foundational infrastructure created for heart disease prediction, researchers can explore the applicability of the developed system to address and forecast various health conditions. This extension aligns with the holistic approach to healthcare and emphasizes the versatility of predictive technologies.

v. Continuous Monitoring and Validation:

In the realm of predictive healthcare technologies, continuous monitoring and validation are crucial. The study recommends establishing mechanisms for ongoing validation and refinement of the developed models. Regular updates, incorporating new data insights, and adapting to emerging healthcare trends will ensure the sustained relevance and accuracy of the predictive tools.

vi. Collaborative Efforts and Knowledge Sharing:

To foster a collaborative approach in advancing healthcare predictive models, the study proposes initiatives for knowledge sharing and interdisciplinary collaboration. Establishing partnerships between data scientists, healthcare professionals, and technology experts can accelerate the development and implementation of innovative solutions. Open channels for communication and collaboration will contribute to a collective effort in improving healthcare outcomes.

vii. Ethical Considerations and Data Privacy:

As predictive models in healthcare evolve, the study underscores the importance of robust ethical considerations and data privacy measures. Future research should prioritize the development of frameworks that ensure responsible and secure handling of patient data. Ethical guidelines should be established to govern the use of predictive models in healthcare settings, safeguarding patient privacy and fostering trust in these technologies.

By addressing these recommendations, future research endeavors can contribute to the ongoing evolution of predictive healthcare technologies in Zambia, fostering advancements that resonate with the dynamic nature of healthcare data and the evolving needs of the healthcare ecosystem.

5.5 Chapter Summary

This chapter marks the culmination of the study, providing a comprehensive review and conclusion. It encapsulates the research questions along with their corresponding answers, offering a synthesis of the key findings and insights derived throughout the study.

REFERENCES

[1] B. Kaur and W. Singh, "Review on Heart Disease Prediction System using Data Mining Techniques," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 10, pp. 3003–3008, 2014.

[2] F. Goma, W. Scholtz, O. Scarlatescu, G. Nel, and J. M. Fourie Zambia, "Zambia Country Report PASCAR and WHF Cardiovascular Diseases Scorecard project," *Cardiovasc. J. afriCa* •, vol. 31, no. 4, 2020, doi: 10.5830/CVJA-2020-038.

[3] S. S. Mirzajani and siamak salimi, "Prediction and Diagnosis of Diabetes by Using Data Mining Techniques," *Avicenna J. Med. Biochem.*, vol. 6, no. 1, pp. 3–7, 2018, doi: 10.15171/ajmb.2018.02.

[4] P. Siva Kumar, D. Anand, V. Uday Kumar, and D. Bhattacharyya, "A computational intelligence method for effective diagnosis of heart disease using genetic algorithm," *Int. J. Bio-Science Bio-Technology*, vol. 8, no. 2, pp. 363–372, 2016, doi: 10.14257/ijbsbt.2016.8.2.34.

[5] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 12, pp. 59–65, 2016.

[6] W. Wiharto, H. Kusnanto, and H. Herianto, "Intelligence system for diagnosis level of coronary heart disease with K-star algorithm," *Healthc. Inform. Res.*, vol. 22, no. 1, pp. 30–38, 2016, doi: 10.4258/hir.2016.22.1.30.

[7] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat. Med. 2019 251*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[8] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digit. Med. 2018 11*, vol. 1, no. 1, pp. 1–10, May 2018, doi: 10.1038/s41746-018-0029-1.

[9] Z. I. Attia *et al.*, "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *Lancet (London, England)*, vol. 394, no. 10201, pp. 861–867, Sep. 2019, doi: 10.1016/S0140-6736(19)31721-0.

[10] K. W. Johnson *et al.*, "Artificial Intelligence in Cardiology," *J. Am. Coll. Cardiol.*, vol. 71, no. 23, pp. 2668–2679, Jun. 2018, doi: 10.1016/J.JACC.2018.03.521.

[11] Z. I. Attia, D. M. Harmon, E. R. Behr, and P. A. Friedman, "Application of artificial intelligence to the electrocardiogram," *Eur. Heart J.*, vol. 42, no. 46, pp. 4717–4730, Dec. 2021, doi: 10.1093/EURHEARTJ/EHAB649.

[12] D. Dey *et al.*, "Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review," *J. Am. Coll. Cardiol.*, vol. 73, no. 11, pp. 1317–1335, Mar. 2019, doi: 10.1016/J.JACC.2018.12.054.

[13] C. Krittanawong, H. J. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial Intelligence in Precision Cardiovascular Medicine," *J. Am. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, May 2017, doi: 10.1016/J.JACC.2017.03.571.

[14] A. Ghaffari and N. Madani, "Atrial fibrillation identification based on a deep transfer learning approach," *Biomed. Phys. Eng. Express*, vol. 5, no. 3, p. 035015, Mar. 2019,

doi: 10.1088/2057-1976/AB1104.

[15] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med. 2019 251*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[16] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/JAMA.2016.17216.

[17] E. K. Oikonomou, M. Siddique, and C. Antoniades, "Artificial intelligence in medical imaging: A radiomic guide to precision phenotyping of cardiovascular disease," *Cardiovasc. Res.*, vol. 116, no. 13, pp. 2040–2054, Nov. 2020, doi: 10.1093/CVR/CVAA021.

[18] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." PMLR, pp. 301–318, Dec. 10, 2016. Accessed: Dec. 12, 2023. [Online]. Available: https://proceedings.mlr.press/v56/Choi16.html

[19] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Futur. Healthc. J.*, vol. 6, no. 2, p. 94, Jun. 2019, doi: 10.7861/FUTUREHOSP.6-2-94.

[20] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, May 2016, doi: 10.1038/SREP26094.

[21] J. Wiens *et al.*, "Author Correction: Do no harm: a roadmap for responsible machine learning for health care (Nature Medicine, (2019), 25, 9, (1337-1340), 10.1038/s41591-019-0548-6)," *Nat. Med.*, vol. 25, no. 10, p. 1627, Oct. 2019, doi: 10.1038/S41591-019-0609-X.

[22] B. Vandenberk, D. S. Chew, D. Prasana, S. Gupta, and D. V. Exner, "Successes and challenges of artificial intelligence in cardiology," *Front. Digit. Heal.*, vol. 5, 2023, doi: 10.3389/FDGTH.2023.1201392.

[23] T. Nakamura and T. Sasano, "Artificial intelligence and cardiology: Current status and perspective," *J. Cardiol.*, vol. 79, no. 3, pp. 326–333, Mar. 2022, doi: 10.1016/J.JJCC.2021.11.017.

[24] Ł. Ledziński and G. Grześk, "Artificial Intelligence Technologies in Cardiology," *J. Cardiovasc. Dev. Dis. 2023, Vol. 10, Page 202*, vol. 10, no. 5, p. 202, May 2023, doi: 10.3390/JCDD10050202.

[25] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nat. Rev. Cardiol. 2021 187*, vol. 18, no. 7, pp. 465–478, Feb. 2021, doi: 10.1038/s41569-020-00503-2.

[26] A. Zagorecki, P. Orzechowski, and K. Hołownia, "A system for automated general medical diagnosis using bayesian networks," *Stud. Health Technol. Inform.*, vol. 192, no. 1–2, pp. 461–465, 2013, doi: 10.3233/978-1-61499-289-9-461.

[27] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 2019-April, no. April 2019, pp. 292–297, 2019, doi: 10.1109/icoei.2019.8862604.

[28] J. Wang *et al.*, "Detecting Cardiovascular Disease from Mammograms with Deep Learning," *IEEE Trans. Med. Imaging*, vol. 36, no. 5, pp. 1172–1181, May 2017, doi: 10.1109/TMI.2017.2655486.

[29] C. Li, X. Hu, and L. Zhang, "The IoT-based heart disease monitoring system for pervasive healthcare service," *Procedia Comput. Sci.*, vol. 112, pp. 2328–2334, Jan. 2017, doi: 10.1016/J.PROCS.2017.08.265.

[30] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, "MLaaS: Machine learning as a service," *Proc. - 2015 IEEE 14th Int. Conf. Mach. Learn. Appl. ICMLA 2015*, pp. 896–902, Mar. 2016, doi: 10.1109/ICMLA.2015.152.

[31] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[32] J. P. A. Ioannidis, "Prediction of Cardiovascular Disease Outcomes and Established Cardiovascular Risk Factors by Genome-Wide Association Markers," *Circ. Cardiovasc. Genet.*, vol. 2, no. 1, pp. 7–15, Feb. 2009, doi: 10.1161/CIRCGENETICS.108.833392.

[33] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," *2010 Int. Conf. Comput. Commun. Technol. ICCCT-2010*, pp. 741–745, 2010, doi: 10.1109/ICCCT.2010.5640377.

[34] Q. Chen *et al.*, "An automatic system to identify heart disease risk factors in clinical texts over time," *J. Biomed. Inform.*, vol. 58, pp. S158–S163, Dec. 2015, doi: 10.1016/J.JBI.2015.09.002.

[35] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," *Int. J. Med. Inform.*, vol. 83, no. 12, pp. 983–992, Dec. 2014, doi: 10.1016/J.IJMEDINF.2012.12.005.

[36] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," *Annu. Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 236–241, Sep. 2017, doi: 10.1109/COMPSAC.2017.164.

[37] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine Learning in Healthcare: A Review," *Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018*, pp. 910–914, Sep. 2018, doi: 10.1109/ICECA.2018.8474918.

[38] J. Zou, Y. Han, and S. S. So, "Overview of artificial neural networks," *Methods in Molecular Biology*, vol. 458. pp. 15–23, 2008. doi: 10.1007/978-1-60327-101-1_2.

[39] C. H. Lee and H. J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, Mar. 2017, doi: 10.23876/J.KRCP.2017.36.1.3.

[40] E. E. Tripoliti, T. G. Papadopoulos, G. S. Karanasiou, K. K. Naka, and D. I. Fotiadis, "Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 26–47, Jan. 2017, doi: 10.1016/J.CSBJ.2016.11.001.

[41] J. Santhana Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms.," *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. ICIICT*

2019, Apr. 2019, doi: 10.1109/ICIICT1.2019.8741465.

[42] K. Kant, K. Garg, and A. Professor, "Review of Heart Disease Prediction using Data Mining Classifications," *IJSRD-International J. Sci. Res. Dev.*, vol. 2, pp. 2321–0613, 2014, Accessed: Dec. 13, 2023. [Online]. Available: www.ijsrd.com

[43] N. Bhatla and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", Accessed: Dec. 13, 2023. [Online]. Available: www.ijert.org

[44] "Review on Hybrid Data Mining Techniques for the Diagnosis of Heart Diseases in Medical Ground - IJAR - Indian Journal of Applied Research." https://www.worldwidejournals.com/indian-journal-of-applied-research-(IJAR)/fileview/August_2015_1443520109__211.pdf (accessed Dec. 13, 2023).

[45] A. Rajkumar, M. G. Sophia, and Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," 2010.

[46] S. Mujawar and P. Devale, "Heart Disease Prediction Using Modified K Means and Using Naive Baiyes," *Int. J. Comput. Sci. Eng. Int. J. Comput. Sci. Eng.*, 2015, Accessed: Dec. 13, 2023. [Online]. Available: www.ijcseonline.org

[47] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/J.ESWA.2016.10.020.

[48] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," *Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017*, pp. 691–698, May 2018, doi: 10.1109/SMARTTECHCON.2017.8358460.

[49] A. Davari Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," *Comput. Methods Programs Biomed.*, vol. 138, pp. 117–126, Jan. 2017, doi: 10.1016/J.CMPB.2016.10.011.

[50] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," 2014.

[51] K. Cinetha and D. Maheswari, "Decision Support System for Precluding Coronary Heart Disease (CHD)," 2014.

[52] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185–196, Jun. 2020, doi: 10.1007/S12065-019-00327-1.

[53] B. Sankara Babu, A. Suneetha, G. Charles Babu, Y. Jeevan Nagendra Kumar, and G. Karuna, "Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network," *Period. Eng. Nat. Sci.*, vol. 6, no. 1, pp. 229–240, 2018, doi: 10.21533/PEN.V6I1.286.

[54] B. Santhi and K. Renuka, "Study and analysis of prediction model for heart disease data using machine learning techniques," *J. Comput. Sci.*, vol. 16, no. 3, pp. 344–354, 2020, doi: 10.3844/JCSSP.2020.344.354.

[55] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Comput.*, vol. 22, pp. 14777–14787, Nov. 2019, doi: 10.1007/S10586-018-2416-4.

[56] S. Nayak, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "Prediction of heart disease by mining frequent items and classification techniques," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, pp. 607–611, May 2019, doi: 10.1109/ICCS45141.2019.9065805.

[57] T. Karadeniz, G. Tokdemir, and H. H. Maraş, "Ensemble Methods for Heart Disease Prediction," *New Gener. Comput.*, vol. 39, no. 3–4, pp. 569–581, Nov. 2021, doi: 10.1007/S00354-021-00124-4.

[58] S. Maji and S. Arora, "Decision Tree Algorithms for Prediction of Heart Disease," *Inf. Commun. Technol. Compet. Strateg.*, vol. 40, pp. 447–454, 2018, doi: 10.1007/978-981-13-0586-3_45.

[59] S. P. Patro, G. S. Nayak, and N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," *Informatics Med. Unlocked*, vol. 26, p. 100696, Jan. 2021, doi: 10.1016/J.IMU.2021.100696.

[60] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES," *Malaysian J. Comput. Sci.*, vol. 2022, no. Special Issue 1, pp. 132–148, 2022, doi: 10.22452/MJCS.SP2022NO1.10.

[61] S. Pandey, S. Prabhakaran, and S. Reddy, "To cite this article: Chiradeep Gupta et al 2022," *J. Phys. Conf. Ser*, vol. 2161, p. 12013, doi: 10.1088/1742-6596/2161/1/012013.

[62] M. Nawaz Ahmad, "COMPREHENSIVE ANALYSIS OF HEART DISEASE PREDICTION USING SCIKIT-LEARN," *Int. Res. J. Mod. Eng. Technol. Sci.*, pp. 2582–5208, Accessed: Dec. 14, 2023. [Online]. Available: www.irjmets.com

[63] F. Fatima, A. Jaiswal, and N. Sachdeva, "Heart Disease Prediction Using Supervised Classifiers," *SSRN Electron. J.*, May 2022, doi: 10.2139/SSRN.4121817.

[64] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021*, pp. 1329–1333, Jan. 2021, doi: 10.1109/ICICT50816.2021.9358597.

[65] E. Miranda, F. M. Bhatti, M. Aryuni, and C. Bernando, "Intelligent Computational Model for Early Heart Disease Prediction using Logistic Regression and Stochastic Gradient Descent (A Preliminary Study)," *Proc. 2021 1st Int. Conf. Comput. Sci. Artif. Intell. ICCSAI 2021*, pp. 11–16, 2021, doi: 10.1109/ICCSAI53272.2021.9609724.

[66] A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.008.

[67] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3. pp. 31–44, Mar. 1996. doi: 10.1109/2.485891.

[68] A. P. Jawalkar *et al.*, "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," *J. Eng. Appl. Sci.*, vol. 70, no. 1, pp. 1–18, Dec. 2023, doi: 10.1186/S44147-023-00280-Y/FIGURES/6.

[69] "(PDF) Heart Disease Prediction Using Logistic Regression." https://www.researchgate.net/publication/368848738_Heart_Disease_Prediction_Using_Logistic_Regression (accessed Dec. 14, 2023).

[70] K. K. Wong, D. N. Ghista, A. W. Ip, W. Zhang, N. Chandrasekhar, and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Process. 2023, Vol. 11, Page 1210*, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/PR11041210.

[71] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," *Int. J. Adv. Intell. Paradig.*, vol. 10, no. 1–2, pp. 118–132, 2018, doi: 10.1504/IJAIP.2018.089494.

[72] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart Diseases Diagnosis Using Neural Networks Arbitration," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 12, pp. 75–82, Nov. 2015, doi: 10.5815/IJISA.2015.12.08.

[73] E. O. Olaniyi, O. Kayode Oyedotun, and K. Adnan, "Intelligent Systems and Applications," *Intell. Syst. Appl.*, vol. 12, pp. 75–82, 2015, doi: 10.5815/ijisa.2015.12.08.

[74] M. I. Hossain *et al.*, "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," *Iran J. Comput. Sci. 2023 64*, vol. 6, no. 4, pp. 397–417, Jun. 2023, doi: 10.1007/S42044-023-00148-7.

[75] A. Khaleel Faieq and M. M. Mijwil, "Prediction of heart diseases utilising support vector machine and artificial neural network," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 1, pp. 374–380, Apr. 2022, doi: 10.11591/IJEECS.V26.I1.PP374-380.

[76] E. Owusu, P. Boakye-Sekyerehene, J. K. Appati, and J. Y. Ludu, "Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/3152618.

[77] A. Sayad and P. Halkarnikar, "DIAGNOSIS OF HEART DISEASE USING NEURAL NETWORK APPROACH," 2014.

[78] "DIAGNOSIS OF HEART DISEASE USING NEURAL NETWORK APPROACH | Semantic Scholar." https://www.semanticscholar.org/paper/DIAGNOSIS-OF-HEART-DISEASE-USING-NEURAL-NETWORK-Sayad-Halkarnikar/c24f2253ca508a47e90ec442c52a0cd9ee90dea9 (accessed Dec. 14, 2023).

[79] S. Priscila and M. Hemalatha, "Improving the Performance of Entropy Ensembles of Neural Networks (EENNS) on Classification of Heart Disease Prediction," 2017.

[80] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," *Int. Conf. Syst. Eng.*, pp. 406–409, 2011, doi: 10.1109/ICSENG.2011.80.

[81] M. B. Wadhonkar, "Artificial Neural Network Approach for Classification of Heart Disease Dataset," 2014.

[82] R. Chitra and D. Seenivasagam, "Heart Attack Prediction System using Cascaded Neural Network," 2013.

[83] R. Chitra, "Heart Attack Prediction System Using Fuzzy C Means Classifier," *IOSR J. Comput. Eng.*, vol. 14, no. 2, pp. 23–31, 2013, doi: 10.9790/0661-1422331.

[84] "Heart Failure Prediction Dataset | Kaggle." https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (accessed Oct. 10, 2022).

[85] S. Roopa and M. S. Rani, "Questionnaire Designing for a Survey," *J. Indian Orthod. Soc.*, vol. 46, no. 4, pp. 273–277, doi: 10.5005/jp-journals-10021-1104.

[86] B. Carolyn Boyce, “PATHFINDER I NTERNATIONAL T OOL SERIES Monitoring and Evaluation-2 CONDUCTING IN-DEPTH INTERVIEWS: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input,” 2006.

[87] I. Sommerville, *Software Engineering*, 9th ed. United States of America: Pearson Education, Inc., 2011. doi: 10.1016/B978-0-12-396961-3.00009-3.

[88] S. Kumar Swain and D. Prasad Mohapatra, “Test Case Generation Based on Use case and Sequence Diagram,” 2010, Accessed: Dec. 14, 2023. [Online]. Available: https://www.researchgate.net/publication/45363457

[89] “Use Case Diagram ~ Computer Science ~ 2420 ~ kelas-karyawan-bali.kurikulum.org.” http://kelas-karyawan-bali.kurikulum.org/IT/en/2420-2301/use-case-diagrams_4557_kelas-karyawan-bali-kurikulum.html (accessed Dec. 14, 2023).

[90] N. Ensmenger, “The Multiple Meanings of a Flowchart,” *http://utpress.utexas.edu/index.php/genders*, vol. 51, no. 3, pp. 321–351, Jul. 2016, doi: 10.7560/IC51302.

[91] I. Nassi and B. Shneiderman, “Flowchart techniques for structured programming,” *ACM SIGPLAN Not.*, vol. 8, no. 8, pp. 12–26, Jan. 1973, doi: 10.1145/953349.953350.

[92] Q. Li and Y.-L. Chen, “Data Flow Diagram,” *Model. Anal. Enterp. Inf. Syst.*, pp. 85–97, 2009, doi: 10.1007/978-3-540-89556-5_4.

[93] W. Wulandari and A. D. Y. Widiantoro, “Design Data Flow Diagram for Supporting the User Experience in Applications,” 2017, Accessed: Dec. 14, 2023. [Online]. Available: http://www.ijcim.th.org/past_editions/2017V25N2/v25n2.html

[94] K. S. R. Anjaneyulu and J. R. Anderson, “The advantages of data flow diagrams for beginning programming,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 608 LNCS, pp. 585–592, 1992, doi: 10.1007/3-540-55606-0_68/COVER.

[95] S. Bernardi, S. Donatelli, and J. Merseguer, “From UML sequence diagrams and statecharts to analysable petri net models,” pp. 35–45, Jul. 2002, doi: 10.1145/584369.584376.

[96] M. I. Kurke, “Operational Sequence Diagrams in System Design,” *https://doi.org/10.1177/001872086100300107*, vol. 3, no. 1, pp. 66–73, Mar. 1961, doi: 10.1177/001872086100300107.

[97] S. Ceri, B. Pernici, and G. Wiederhold, “Distributed Database Design Methodologies,” *Proc. IEEE*, vol. 75, no. 5, pp. 533–546, 1987, doi: 10.1109/PROC.1987.13771.

[98] G. Wiederhold and H. Broun, “are-Da%aßa DATABASE DESIGN,” 1976.

[99] “Beginning Database Design - Gavin Powell - Google Books.” https://books.google.co.zm/books?hl=en&lr=&id=HbAhv1zAIQ8C&oi=fnd&pg=PR17&dq=database+design+&ots=iYM-_ZwU61&sig=T32_Ha5XJFiSoV8ZHVJn-ui5GCU&redir_esc=y#v=onepage&q=database design&f=false (accessed Dec. 14, 2023).

[100] *software development process*. 2020. Accessed: Aug. 21, 2023. [Online]. Available: https://taazaa.com/software-development-process/

[101] C. J. Du and D. W. Sun, “Object Classification Methods,” *Comput. Vis. Technol. Food Qual. Eval.*, pp. 81–107, 2008, doi: 10.1016/B978-012373642-0.50007-7.

APPENDICES

Appendix 1 - Questionnaire

The University of Zambia

School of Natural and Applied Sciences

Computer Science Department

Assisted Artificial Intelligence Medical Diagnosis System for Heart Disease

---

Mweemba Maambo

Master of Science Computer Science

For more information or any queries, kindly get in touch on:
Cell: **+260976248979**
Email: **maambomweemba4@gmail.com**

Dear Respondent,

I am a student at the University of Zambia in my final stage pursuing a Master of Science Computer Science. As partial fulfilment for the award of a master's degree, I am conducting a baseline study on: "***Assisted Artificial Intelligence Medical Diagnosis System for Heart Disease.***"

You have been purposefully sampled to provide information for the topic indicated above. The information being collected is purely for academic purposes as such, it will be treated with maximum confidentiality. Subsequently, you are not supposed to indicate your name or any personal information that can lead to revealing of your identity.

Your co-operation will be greatly appreciated.

For more information or any queries, kindly get in touch with the following:

Project Supervisor: Prof. Jackson Phiri (Jackson.phiri@cs.unza.zm)

SURVEY QUESTIONNAIRES

PART ONE: DEMOGRAPHIC INFORMATION (PLEASE TICK [√])

1. Gender: Male [ ] Female [ ]

2. Marital Status: Single [ ] Married [ ] Divorced [ ] Other [ ]

3. Age: 20 or under [ ] 21-30 [ ] 31-40 [ ] 41-50 [ ] 51-60 [ ] 61+ [ ]

4. Highest level of education: SHS and below [ ] Diploma [ ] First degree [ ] Masters [ ] Ph.D. [ ]

5. Type of employment: Not working [ ] Salaried worker [ ] Self-employed [ ] Pensioner [ ]

6. Occupation (Please specify, e.g., "University Lecturer in Graduate School of Natural Science")

PART TWO: COMPUTER KNOWLEDGE AND EXPERIENCE (PLEASE TICK [√])

7. How do you describe your general knowledge about computers? Very poor [ ] Poor [ ] Moderate [ ] Good [ ] Very good [ ]

8. How would you describe your Internet knowledge? Very poor [ ] Poor [ ] Moderate [ ] Good [ ] Very good [ ]

9. How long have you been using Internet? Don't use [ ] Less than 1yr [ ] 1- 2 yrs. [ ] More than 2 yrs. [ ]

10. How often do you use the Internet per day? Don't use [ ] Less than 1hr [ ] 1-2 hrs. [ ] 3- 4 hrs. [ ] More than 4 hrs. [ ]

PART THREE: ASSISTED MEDICAL DIAGNOSIS HEART DISEASE SYSTEM ADOPTION FACTORS

Using a rating scale from the lowest point of 1 to the highest point of 5, please circle the number that indicates your level of agreement or disagreement with the following statement.

SD = strongly disagree | D = Disagree | N = Neutral | A = Agree | SA = Strongly Agree | NA= Not Application

No	Statement					
	Performance Expectancy	**SD**	**D**	**N**	**A**	**SA**
1	I think that the assisted medical diagnosis heart disease system will benefit health practitioners	1	2	3	4	5
2	I think that using the assisted medical diagnosis heart disease system would help me predict heart disease more quickly	1	2	3	4	5
3	I think that using the assisted medical diagnosis heart disease system would increase my productivity	1	2	3	4	5
4	I think using the assisted medical diagnosis heart disease system would improve my performance	1	2	3	4	5
	Effort Expectancy	**SD**	**D**	**N**	**A**	**SA**

1	I think that interaction with the assisted medical diagnosis heart disease system is clear and easily understandable	1	2	3	4	5
2	I think it's easy to become skillful at using the assisted medical diagnosis heart disease system	1	2	3	4	5
3	I find the assisted medical diagnosis heart disease system easy to use	1	2	3	4	5
4	I think that learning to operate the assisted medical diagnosis heart disease system is easy for me	1	2	3	4	5
5	I think the user interface for the assisted medical diagnosis heart disease system is well designed for any one no matter one 's computer literacy level	1	2	3	4	5
Social Influence		**SD**	**D**	**N**	**A**	**SA**
1	People who influence my behaviour think that I should use the assisted medical diagnosis heart disease system	1	2	3	4	5
2	People who are important to me think that I should use the assisted medical diagnosis heart disease system	1	2	3	4	5
Facilitating Conditions		**SD**	**D**	**N**	**A**	**SA**
1	I have the resources necessary to use the assisted medical diagnosis heart disease system	1	2	3	4	5
2	I have the knowledge necessary to use the assisted medical diagnosis heart disease system	1	2	3	4	5
3	Help/guidance is available on using the assisted medical diagnosis heart disease system	1	2	3	4	5
4	The assisted medical diagnosis heart disease system has most of the services I need from heart disease prediction	1	2	3	4	5
5	I am aware and understand the services/activities that can be done on the assisted medical diagnosis heart disease system	1	2	3	4	5
Behavioural Intention		**SD**	**D**	**N**	**A**	**SA**
1	I intend to use the system in the next months.	1	2	3	4	5
2	I predict I would use the assisted medical diagnosis heart disease system in the next months.	1	2	3	4	5
3	I plan to use the assisted medical diagnosis heart disease system in the next months.	1	2	3	4	5
4	I intend to predict the type of heart disease on the assisted medical diagnosis heart disease system	1	2	3	4	5
5	I intend to register on the assisted medical diagnosis heart disease system	1	2	3	4	5

PART FOUR: ACTUAL USE OF THE ASSISTED MEDICAL DIAGNOSIS HEART DISEASE SYSTEM (PLEASE TICK [√])

1. How long have you been using the assisted medical diagnosis heart disease system? Under 1 year [ ] 1-2 years [ ] 3- 4 years [ ] more than 4 years [ ]

2. On a weekly basis, how many times do you use the assisted medical diagnosis heart disease system? Not at all [ ] once a week [ ] 2-3 times [ ] more than 3 times [ ]

3. How frequently do you use the assisted medical diagnosis heart disease system for the following services?

Functionality	Never 1	Rarely 2	Sometimes 3	Often 4	Always 5
View stored heart disease data					
Predict heart disease					

Action Control	Never 1	Rarely 2	Sometimes 3	Often 4	Always 5
Input heart disease features					
Filter number of heart disease patients					

Appendix 2: NHRA Certificate of Registration

NATIONAL HEALTH RESEARCH AUTHORITY

The Health Research Act

(Act No. 2 of 2013)

CERTIFICATE OF REGISTRATION

THIS IS TO CERTIFY THAT

Maambo Mweemba

has been registered as a Health Researcher

Dated this 16th Day of November 2022

Registration NHRAR-R-1401/21/11/2022

DIRECTOR

PROF. GODFREY BIEMBA

Appendix 3: UNZA Approval of Study

THE UNIVERSITY OF ZAMBIA
DIRECTORATE OF RESEARCH AND GRADUATE STUDIES

Great East Road Campus | P.O. Box 32379 | Lusaka10101 | Tel: +260-211-290 258/291 777
Fax: (+260)-211-290 258/253 952 | E-mail: director.dres@unza.zm | Website: www.unza.zm

APPROVAL OF STUDY

IORG No. 0005376
NASREC IRB No. 00006465

15th November, 2022

REF NO. NASREC-2022-OCT.-002

Ms. Maambo Mweemba,
The University of Zambia,
School of Natural Sciences,
P.O. Box 32379
LUSAKA

Dear Ms. Maambo,

RE: "ASSISTED ARTIFICIAL INTELLIGENCE MEDICAL DIAGNOSIS SYSTEM FOR HEART DISEASE "

Reference is made to your protocol dated as captioned above. NASREC resolved to approve this study and your participation as Principal Investigator for a period of one year.

REVIEW TYPE	ORDINARY REVIEW	APPROVAL NO. NASREC-2022-OCT..-002
Approval and Expiry Date	Approval Date: 15th November, 2022	Expiry Date: 14th November, 2023
Protocol Version and Date	Version - Nil.	14th November, , 2023
Information Sheet, Consent Forms and Dates	• English.	To be provided
Consent form ID and Date	Version - Nil	To be provided
Recruitment Materials	Nil	Nil
Other Study Documents	Questionnaire.	

Appendix 4: Publications

[1] M. Maambo, J. Phiri, M. Kalumbilo, and L. Jaganathan, "Assisted Artificial Intelligence Medical Diagnosis System for Heart Disease", zictjournal, vol. 6, no. 1, pp. 38–43, Dec. 2022.

[2] The following is my second publication accepted for Springer LNNS series.

AI-Assisted Heart Disease Diagnosis System in Medicine Utilizing Bayesian Classification

Mweemba Maambo[1] and Jackson Phiri[2]

[1] University of Zambia, Lusaka 10101, Zambia
[2] University of Zambia, Lusaka 10101, Zambia
maambomweemba4@gmail.com[1], Jackson.phiri@cs.unza.zm[2]

Abstract. In recent years, the surge in the development of Innovative and impactful applications in the realm of medicine has played a serious role in study. Artificial Intelligence (AI) systems, in particular, have significantly contributed to the expansion of these applications and tools. Addressing one key element in the realm of health concerns globally, this paper focuses on heart disease diagnosis, emphasizing the pivotal role of AI in regulating and improving diagnostic processes. The suggested Artificial Intelligence Medical Diagnosis System employs input medical information sourced from an established dataset available on Kaggle. This information is utilized in an AI application designed with an algorithm for data mining and a basic model, specifically customized for patients in Zambia. The dataset, comprised of medical parameters such as the kind of chest pain, oldpeak, ST-slope, blood pressure, blood sugar (mg/dl), cholesterol (mm/dl), heart rate, exercise-induced angina, resting ECG, age and sex, undergoes a comprehensive pre-processing stage. Subsequently, supervised learning techniques are employed to train and evaluate the predictive model. The Bayesian data mining algorithm is utilized to estimate the likelihood and risk category of heart disease in Zambian patients. Our findings, based on the analysis of the dataset, indicate a remarkable prediction accuracy of 90.97%. This level of accuracy aligns with results produced by other algorithms like Decision Tree, Random Forest, and KNN. The outcomes underscore the effectiveness of the proposed AI-assisted system in predicting heart disease in a Zambian patient population. The application of AI in medical diagnosis, as illustrated by this study, not only showcases its potential in improving diagnostic accuracy but also opens avenues for further research and implementation in diverse healthcare settings.

Keywords: Prediction Model, Artificial Intelligence, Bayesian Classification, Heart Disease, Supervised Learning Techniques, Data Mining Algorithms.

1 Introduction

In contemporary times, heart disease stands out as a formidable health challenge, encompassing a spectrum of medical disorders that directly affect the heart's various components [1]. This pervasive health issue claims the lives of over 17.5 million individuals annually, solidifying its position as one of the world's leading causes of mortality. Projections indicate a distressing surge, with an anticipated rise to 23 million lives claimed by 2030. In Zambia, recent data compiled by the World Heart Federation (WHF) highlights that heart disease is a significant contributor to mortality, accounting for 10 percent of all deaths among individuals aged 30 to 70 [2]. Given these alarming statistics, the imperative to enhance diagnostic capabilities for heart disease becomes increasingly vital to prompt appropriate and timely interventions.

Appendix 5: Flask server code

```
from api_routes import bp1
import joblib
from flask import Flask, render_template, redirect, url_for, request
from flask_bootstrap import Bootstrap
from flask_wtf import FlaskForm
from wtforms import StringField, PasswordField, BooleanField
from wtforms.validators import InputRequired, Email, Length
from flask_sqlalchemy import SQLAlchemy
from werkzeug.security import generate_password_hash, check_password_hash
from flask_login import LoginManager, UserMixin, login_user, login_required, logout_user
import pandas as pd
import pickle
import numpy as np
from sklearn.ensemble import RandomForestClassifier

app = Flask(__name__)

app.register_blueprint(bp1)
app.config['SECRET_KEY'] = 'secret'
app.config['SQLALCHEMY_DATABASE_URI'] = 'sqlite:///database.db'
bootstrap = Bootstrap(app)
db = SQLAlchemy(app)
login_manager = LoginManager()
login_manager.init_app(app)
login_manager.login_view = 'login'

class User(UserMixin, db.Model):
    id = db.Column(db.Integer, primary_key=True)
```

```
    username = db.Column(db.String(15), unique=True)
    email = db.Column(db.String(50), unique=True)
    password = db.Column(db.String(80))


@login_manager.user_loader
def load_user(user_id):
    return User.query.get(int(user_id))


class LoginForm(FlaskForm):
    username = StringField('Username', validators=[InputRequired(), Length(min=4,
max=15)])
    password = PasswordField('Password', validators=[InputRequired(), Length(min=8,
max=80)])
    remember = BooleanField('remember me')


class RegisterForm(FlaskForm):
    email = StringField('Email', validators=[InputRequired(), Email(message='Invalid email'),
Length(max=50)])
    username = StringField('Username', validators=[InputRequired(), Length(min=4,
max=15)])
    password = PasswordField('Password', validators=[InputRequired(), Length(min=8,
max=80)])


@app.route('/login', methods=['GET', 'POST'])
def login():
    form = LoginForm()
    if form.validate_on_submit():
        user = User.query.filter_by(username=form.username.data).first()
        if user:
            if check_password_hash(user.password, form.password.data):
                login_user(user, remember=form.remember.data)
                return redirect(url_for('dashboard'))
```

```
        return render_template("login.html", form=form)
    return render_template("login.html", form=form)


@app.route('/signup', methods=['GET', 'POST'])
def signup():
    form = RegisterForm()
    if form.validate_on_submit():
        hashed_password = generate_password_hash(form.password.data, method='sha256')
        new_user = User(username=form.username.data, email=form.email.data,
password=hashed_password)
        db.session.add(new_user)
        db.session.commit()


        return redirect("/login")
    return render_template('signup.html', form=form)


# @app.route('/')
# @login_required
# def index():
#     return render_template('log.html')
@app.route('/', methods=['GET', 'POST'])
def index():
    form = LoginForm()
    if form.validate_on_submit():
        user = User.query.filter_by(username=form.username.data).first()
        if user:
            if check_password_hash(user.password, form.password.data):
                login_user(user, remember=form.remember.data)
                return redirect(url_for('dashboard'))
```

```
        return render_template("login.html", form=form)
    return render_template("login.html", form=form)

@app.route('/logout')
@login_required
def logout():
    logout_user()
    return redirect(url_for('login'))

@app.route("/dashboard")
@login_required
def dashboard():
    return render_template("index.html")

@app.route("/about")
@login_required
def about():
    return render_template("about.html")


if __name__ == "__main__":
    app.run(debug=False)
```