A DIGITAL PREDICTIVE HEALTHCARE MANAGEMENT SYSTEM FOR SICKLE CELL DISEASE USING MACHINE LEARNING AND DATA VISUALIZATION TECHNIQUES: A CASE STUDY OF UGANDA

TIRZAH ATWIINE

M22B23/004

A PROJECT REPORT SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE OF UGANDA CHRISTIAN UNIVERSITY

April, 2025

UGANDA CHRISTIAN UNIVERSITY

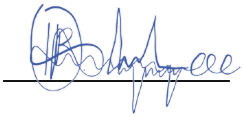
A Centre of Excellence in the Heart of Africa

APPROVAL

This report is hereby submitted for examination with the consent and endorsement of my supervisor listed below.

Supervisor Name: Mr. Ian Raymond Osolo
Email: iosolo@ucu.ac.ug
Contact: +256 700 825 825
Department: Department of Computing
Faculty: Faculty of Engineering, Design and Technology

Date: 5/5/25

Sign: 

DECLARATION

I, Tirzah Atwiine, solemnly affirm that this report is a result of my independent work and has not been previously submitted or published for the attainment of any academic qualification at any other institution.

Date: 05/05/2025

Sign: 

ACKNOWLEDGEMENT

Bringing this project, "A Digital Predictive Healthcare Management System for Sickle Cell Disease Using Machine Learning and Data Visualization Techniques: A Case Study of Uganda", to life has been a rewarding journey, one that would not have been possible without the support, encouragement, and insight of many remarkable individuals.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Mr. Ian Raymond Osolo, for his exceptional guidance, constructive feedback, and continuous encouragement throughout the course of this project. His mentorship not only enriched the quality of this research but also inspired me to approach challenges with clarity and confidence.

To my classmates and peers, thank you for your camaraderie, idea-sharing, and encouragement during discussions, revisions, and brainstorming sessions. Your presence made the journey more collaborative and insightful.

To the researchers and experts whose literature, studies, and innovations formed the backbone of this work, I extend my sincere thanks. Your dedication to science, health, and technology continues to shape the future in meaningful ways.

Lastly, to my family and close friends, thank you for your unwavering support, patience, and belief in me. Your encouragement has been my silent strength every step of the way.

To everyone who contributed in one way or another, thank you.

Abstract

This project presents the design and development of a Digital Predictive Healthcare Management System for Sickle Cell Disease (SCD) using Machine Learning and Data Visualization techniques, with a focus on Uganda. Sickle Cell Disease remains a significant public health challenge in Uganda, with limited access to timely diagnosis, treatment monitoring, and personalized care. The proposed system leverages machine learning algorithms; Random Forest Classifier and LSTM(Long-Short-Term-Memory), to predict potential health risks, analyse and predict future patient data, and support early interventions. Interactive dashboards and visual tools created using React provide healthcare professionals and patients with actionable insights for better disease management. This project aims to enhance decision-making, improve patient outcomes, and support national efforts in digital health transformation, particularly in under-resourced settings.

List of Figures

Contents

1 CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

Sickle Cell Disease (SCD) is an inherited blood disorder caused by a mutation in the hemoglobin gene, leading to the production of abnormal, sickle-shaped red blood cells. These misshapen cells can block blood flow, causing pain, organ damage, and increased risk of infection. Sickle Cell Disease is one of the most significant inherited blood disorders worldwide, especially in sub-Saharan Africa, where its prevalence poses a serious public health challenge. Uganda, in particular, faces a high burden of SCD, with approximately 13.3% of children born carrying the sickle cell trait and an estimated 20,000 new cases diagnosed annually [1]. SCD is characterized by the production of abnormal haemoglobin, which leads to chronic anaemia, recurrent pain crises, organ damage, and increased vulnerability to infections.

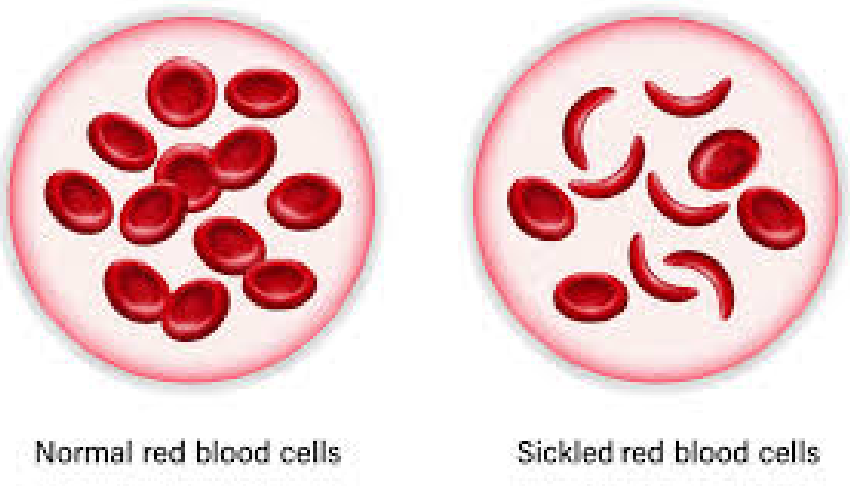


Figure 1: Sickled Red Blood Cell vs Normal Red Blood cell.

In Uganda, managing SCD is hindered by several systemic inefficiencies, including inadequate diagnostic facilities, insufficient healthcare infrastructure, and limited access to specialized treatment. The lack of early diagnosis and follow-up care exacerbates the condition, leading to preventable complications and high mor-

tality rates [2]. The absence of robust electronic health systems and the reliance on manual health records further impede the effective management of SCD patients, complicating efforts to streamline care and monitor disease progression.

Despite initiatives like the Uganda Sickle Surveillance Study (US3) and partnerships such as the Sickle Pan-African Research Consortium (SPARCo), gaps persist in ensuring equitable and timely access to care. Technological advancements, such as machine learning for predictive analytics, electronic health records (EHR), wearable health monitoring devices, and telemedicine platforms, present opportunities to bridge these gaps. However, their implementation in Uganda remains limited, constrained by financial, infrastructural, and technological barriers. [3]

This research aims to address these challenges by proposing an integrative framework that leverages modern technology to improve the management of SCD in Uganda. By combining predictive analytics, digital health tools, and community engagement, the study seeks to enhance early diagnosis, optimize treatment plans, and ultimately improve the quality of life for individuals affected by SCD.

1.2 Statement of the Problem

Sickle Cell Disease (SCD) poses a significant public health challenge in Uganda, particularly in the northern and eastern regions, where its prevalence is high [1]. Despite efforts to enhance awareness and treatment, inefficiencies in the management of SCD persist, including inadequate early diagnosis and limited screening programs, contributing to high mortality and preventable complications [2]. The current healthcare system struggles with fragmented care and lacks the necessary infrastructure and trained personnel to meet the complex needs of SCD patients [4]. To ad-

dress these challenges, a multifaceted approach is essential, focusing on improving screening, training healthcare providers, increasing treatment access, and integrating technology-driven solutions such as electronic health records for better patient management.

Sickle cell disease (SCD) remains a concern in Uganda, affecting a substantial portion of the population. Despite efforts to raise awareness and improve treatment options, the management of SCD is plagued by inefficiencies at various levels of healthcare delivery. According to the Uganda Sickle Surveillance Study, the prevalence of sickle cell trait is high, particularly in the northern and eastern regions of the country. [1] However, early diagnosis, routine screening, and consistent management are lacking, leading to preventable complications, high mortality rates, and unnecessary suffering for patients. Furthermore, Hernandez et al. (2021) point out that while national screening programs have made progress, their reach remains limited due to inadequate healthcare infrastructure, lack of trained personnel, and insufficient public knowledge about the disease. [2]

The current healthcare systems in Uganda are not fully equipped to handle the long-term and complex needs of SCD patients, leading to fragmented care and poor patient outcomes. Challenges include delayed diagnosis, inadequate genetic counseling, and a lack of integration of modern technologies to streamline patient management. Additionally, caregivers often lack adequate support, and healthcare providers face difficulties in managing chronic symptoms, pain episodes, and complications such as infections or stroke. [4]

The lack of digital health infrastructure is a critical root cause affecting the management of Sickle Cell Disease (SCD) in Uganda. Inadequate digital systems limit the ability to perform efficient screening, diagnosis, and ongoing patient management, leading to fragmented care and suboptimal health outcomes (Hernandez et

al., 2021). Without robust electronic health records and telemedicine platforms, healthcare providers struggle to track patient histories, monitor disease progression, and coordinate treatment, exacerbating the challenges faced by SCD patients. [4] [5] Strengthening digital health infrastructure is essential to improve healthcare delivery and ensure comprehensive, integrated care for individuals affected by SCD.

Addressing this problem requires a multifaceted approach that includes strengthening the capacity of healthcare systems to screen, diagnose, and treat SCD, alongside raising awareness and enhancing education about the disease.

1.3 Objectives of the Project

1.3.1 Main Objective

The main objective of this project is to design and develop a digital predictive healthcare management system for sickle cell disease (SCD) using machine learning and data visualization techniques. This system aims to improve the diagnosis, monitoring, and management of SCD in Uganda by enabling doctors to efficiently analyze health data, predict potential complications, and ensure better healthcare decision-making.

1.3.2 Specific Objectives

- To design and develop a user-friendly digital platform that allows for upload for Complete Blood Count (CBC) reports, ensuring easy access for doctors to review and analyze the data for effective decision-making.
- To implement data visualization tools within the platform that present graphical trends of key health metrics, such as hemoglobin levels and white blood cell counts, enabling doctors to track patient health over time.

- To integrate machine learning models into the platform capable of analyzing CBC reports to predict potential health complications and future CBC values, offering valuable insights and recommendations to doctors.
- To create a secure, centralized database system for storing patients records, ensuring that data is accessible to authorized personnel while maintaining data privacy and complying with healthcare data protection standards.

1.4 Research Questions / Hypotheses

1. How can a digital healthcare platform effectively facilitate the upload, review, and analysis of Complete Blood Count (CBC) reports for sickle cell disease (SCD) patients in Uganda?
2. What data visualization tools and techniques can be implemented to present patient health metrics over time in a way that is both meaningful and easy for doctors to interpret?
3. To what extent can a machine learning model be integrated into the digital healthcare platform to accurately predict potential health complications associated with SCD from CBC reports?
4. How can a secure, centralized database system be developed to ensure the privacy, accessibility, and compliance of patient records in the context of SCD management in Uganda?

1.4.1 Hypotheses

- H1: A digital healthcare platform that allows patients to upload their CBC reports and enables doctors to analyze these reports will improve the efficiency of diagnosis and management of sickle cell disease (SCD).

- H2: The use of data visualization tools to track health metrics such as hemoglobin levels, white blood cell counts, and other relevant parameters will enhance doctors' ability to make informed decisions and monitor the progression of SCD.
- H3: A machine learning model capable of predicting health complications based on CBC reports will provide valuable insights that assist doctors in proactive decision-making, thereby reducing the risk of severe complications for SCD patients.
- H4: A secure, centralized database system for storing patient records will ensure better management of SCD patient data, improve data accessibility for authorized personnel, and maintain compliance with healthcare standards in Uganda.

These research questions and hypotheses aim to explore the effectiveness of integrating technology in managing sickle cell disease and to evaluate the impact of data-driven tools on healthcare delivery in Uganda.

1.5 Scope and Limitations of the Study

1.5.1 Geographical Scope

Sickle Cell Disease (SCD) is a global health concern that primarily affects people of African, Mediterranean, Middle Eastern, and Indian ancestry. It is one of the most prevalent inherited blood disorders worldwide. According to the World Health Organization (WHO), SCD affects approximately 5 million people globally and remains a leading cause of morbidity and mortality, particularly in low- and middle-income countries (LMICs) where healthcare systems may lack adequate resources to manage the disease effectively [6].

The global burden of SCD is concentrated mainly in sub-Saharan Africa, where the disease's prevalence is particularly high due to

the genetic traits carried by populations in these regions. The widespread nature of the disease in these areas calls for urgent attention to healthcare systems, research, and resources to improve management and treatment outcomes.

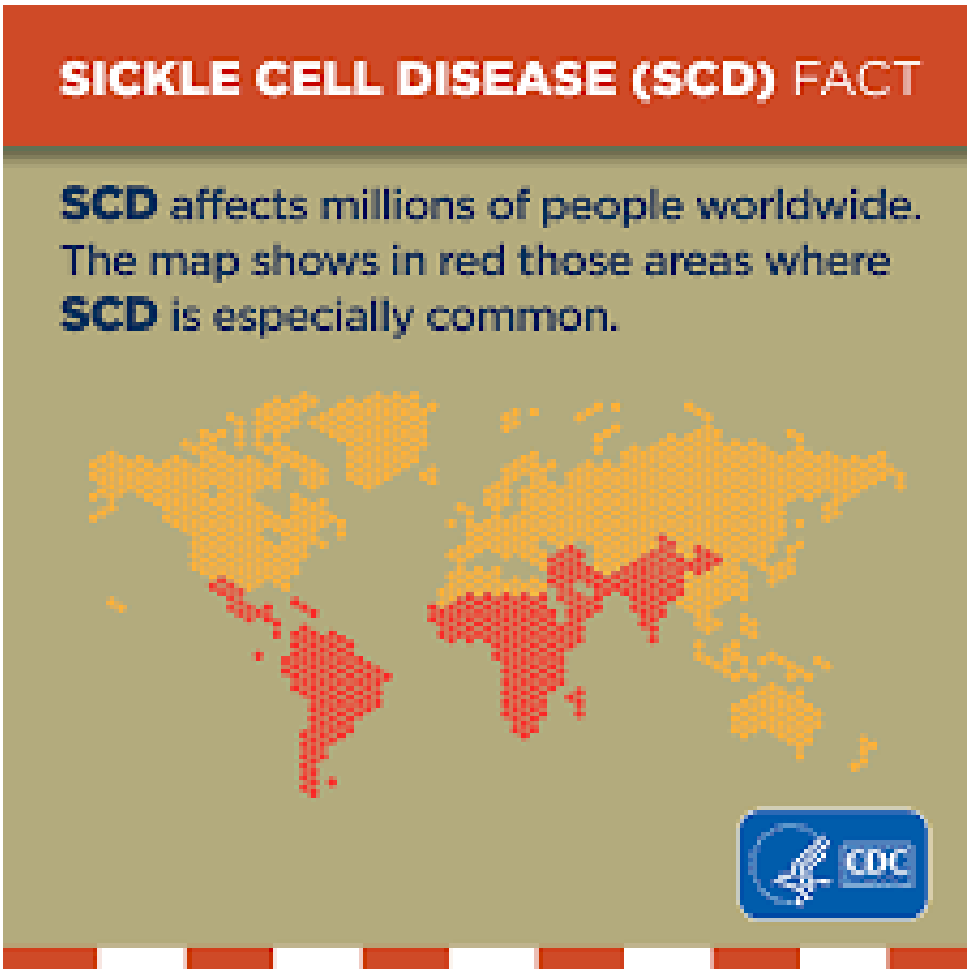


Figure 2: Sickle Cell Disease Distribution.

Sub-Saharan Africa Context

Sub-Saharan Africa bears the heaviest burden of SCD, accounting for 75-80% of global cases [7]. This region is home to a significant proportion of the world's sickle cell population, with some countries reporting the highest prevalence rates globally. In sub-Saharan Africa, an estimated 150,000 to 200,000 children are born with SCD each year, and millions carry the sickle cell trait [8].

The challenge in sub-Saharan Africa is multifaceted. The healthcare infrastructure in many African countries, including Uganda, is

often underdeveloped and insufficient to meet the needs of individuals affected by SCD. This is compounded by a lack of widespread screening programs, limited access to specialized care, and inadequate treatment options. Early diagnosis and preventive care, which are critical for managing SCD and improving patient outcomes, are often unavailable in these regions, contributing to high rates of morbidity and mortality [2].

Moreover, social, economic, and political factors, including poverty and lack of access to quality healthcare services, exacerbate the challenges faced by individuals with SCD in sub-Saharan Africa. A significant number of people with SCD in the region do not receive the necessary medical care, leading to complications such as stroke, organ failure, and early death.

Uganda Context

Uganda, located in East Africa, has a population of approximately 48 million people [9], and it is among the countries in sub-Saharan Africa that faces a high prevalence of SCD. Studies show that around 13.3% of Ugandans carry the sickle cell trait, which is higher than the global average [1]. This high prevalence of the sickle cell trait and the large number of people living with SCD contribute to significant public health challenges in the country.

The Uganda Sickle Surveillance Study (US3) has found that there are approximately 20,000 new cases of SCD diagnosed annually in Uganda [2]. SCD affects individuals across various regions of the country, both in urban centers such as Kampala and rural areas where access to healthcare services is limited. Uganda's healthcare system faces considerable challenges, including inadequate healthcare infrastructure, limited healthcare workers trained in SCD management, and insufficient diagnostic facilities. Despite ongoing efforts to improve healthcare services, these challenges persist, leading to suboptimal care for individuals with SCD.

In Uganda, early diagnosis of SCD is crucial for reducing compli-

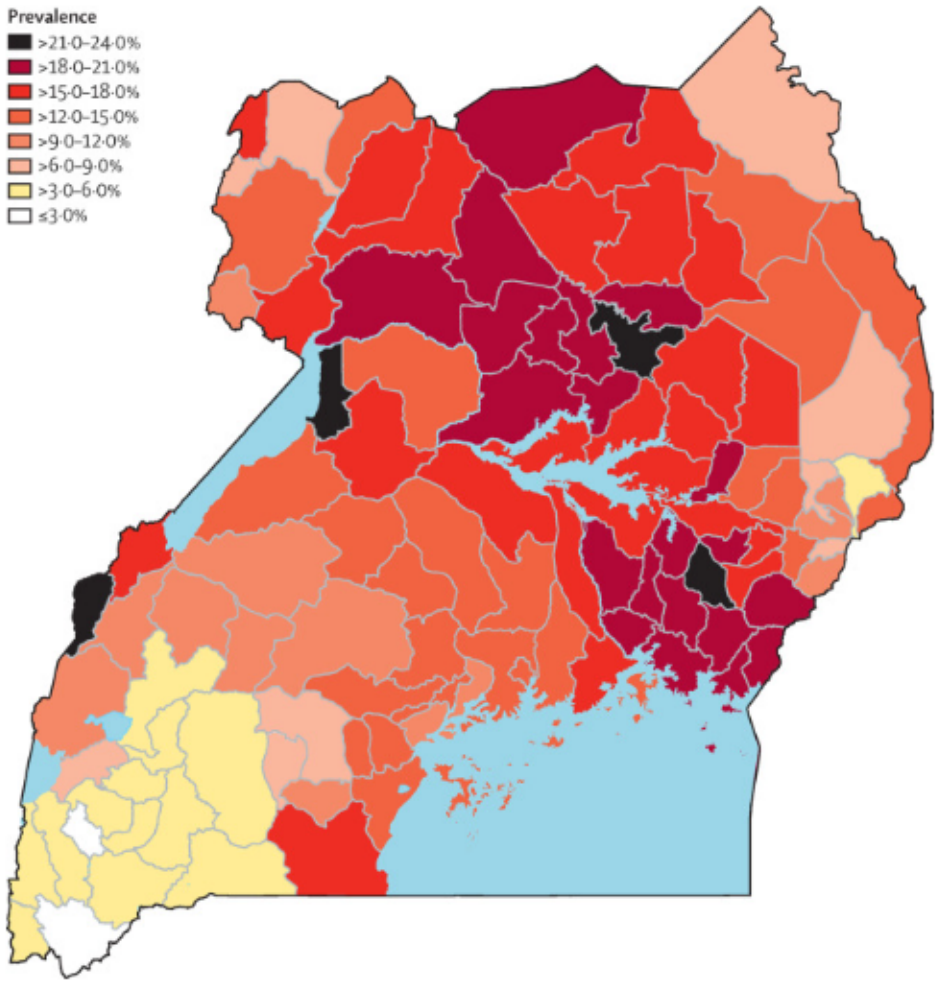


Figure 1 Prevalence of sickle cell trait in 112 districts in Uganda

Figure 3: Image showing sickle cell prevalence in Uganda.

cations. However, due to the lack of widespread neonatal screening programs, many children with SCD go undiagnosed until they develop serious complications [10]. Furthermore, treatment options such as hydroxyurea therapy and blood transfusions are often unavailable or difficult to access for many families. The economic burden of managing SCD in Uganda is also substantial, as the disease requires lifelong care, and many families struggle to afford the necessary treatments.

Uganda has made strides in addressing these challenges through initiatives however, gaps remain in the equitable distribution of care across the country, particularly in rural areas where healthcare access is limited [3]. The integration of modern technologies, such as electronic health records (EHR), telemedicine, and machine learning for predictive analytics, offers promising solutions to improving the management of SCD in Uganda.

Given these factors, Uganda's healthcare landscape presents both challenges and opportunities for improving the management of SCD. The proposed digital healthcare platform aims to address some of these challenges by offering a scalable and sustainable solution that can be deployed across the country, with the potential for positive impacts on patient outcomes, particularly in resource-limited settings.

1.5.2 Content Scope

The content of this study will primarily focus on the design, development, and implementation of a digital predictive healthcare management system tailored for Sickle Cell Disease (SCD) patients. The key aim is to leverage machine learning and data visualization techniques to enhance the monitoring, management, and prediction of health complications in individuals living with SCD. This research will address the following critical aspects:

1. Design and Development of a Digital Platform:* The project focuses on developing a user-friendly digital platform that allows doctors to upload and share their Complete Blood Count (CBC) reports. This platform enables healthcare professionals have timely reviews, efficient analysis, and informed decision-making. The platform is designed to accommodate various health metrics essential for managing SCD, including hemoglobin levels, red blood cell counts, and white blood cell counts.

2. Integration of Data Visualization Tools The research includes the development of data visualization tools that enable healthcare providers to track and analyze patient health metrics over time. These tools will provide graphical representations of important health parameters, such as hemoglobin concentration, to help clinicians identify trends and potential health risks. By using visualization techniques like line charts, bar graphs, and heat maps, the system aims to improve the accessibility and interpretability of patient data for better clinical decisions.

3. Development of Machine Learning Models A core component of the study involves developing machine learning models that can analyze CBC reports to predict potential health complications and future CBC values in SCD patients. The models are trained on historical health data to identify patterns that may indicate an elevated risk of common SCD-related complications. The prediction outcomes help doctors make proactive treatment recommendations and enable timely intervention for patients at high risk of developing complications.

4. Establishment of a Secure and Centralized Database The research explores the creation of a secure, centralized database system for storing and managing patient records. This system ensures data integrity, privacy, and compliance with healthcare standards such as the Health Insurance Portability and

Accountability Act (HIPAA). The database allows authorized healthcare providers to access patient records from various locations, facilitating more efficient care coordination and reducing the chances of errors due to manual record-keeping.

Exclusions While aspects such as telemedicine platforms and mobile health tools are relevant to the broader digital health landscape, this project primarily focuses on the predictive analytics and data management components of the system. The research excludes the development of telemedicine features, as the emphasis is on utilizing machine learning and data visualization to improve SCD healthcare management. Additionally, non-SCD diseases and conditions are not part of the scope of this study, as the focus remains on the healthcare challenges and needs specific to individuals with Sickle Cell Disease.

By focusing on these core areas, this project aims to develop a comprehensive digital healthcare management system that could significantly improve the early diagnosis, continuous monitoring, and management of SCD, ultimately contributing to better outcomes for patients in resource-limited settings.

1.5.3 Time Scope

The time scope of this project spans 6 months, focusing on the design, development, and evaluation of a digital predictive healthcare management system for Sickle Cell Disease (SCD).

1. Project Duration: The entire study was conducted over a period of 6 months, encompassing the system design, machine learning model training, and the creation of a secure database for patient records.

2. Phase Breakdown: Phase 1: Research and Data Collection (Month 1 - 2): This phase focused on reviewing existing literature, collecting relevant health data like CBC reports, and analyzing current SCD management practices in Uganda.

   Phase 2: System Design and Development (Month 3 - 4): The platform and user interface was developed, and predictive analytics tools were integrated. Initial testing of the prototype took place.

   Phase 3: Pilot Testing and Feedback (Month 5): Gathered feedback from healthcare providers to refine the system.

   Phase 4: Full System Deployment and Evaluation (Month 6): Deployed the system and evaluated the system's impact on SCD management.

3. Time Constraints: Given the limited timeframe, the study focused on achieving a working prototype with a thorough evaluation of its initial impact. Potential delays due to technical or logistical challenges were considered and mitigated.

1.5.4 Limitations

Although the project successfully met its primary objectives, it faced several limitations. The system was developed and tested in a controlled environment, and thus was not deployed or validated in real healthcare settings. This limited the ability to assess its practical effectiveness, user experience, and adaptability in live clinical workflows.

Access to real patient data, especially anonymized CBC reports, was restricted due to privacy and ethical concerns, which affected the diversity and volume of data used to train and evaluate the machine learning model. As a result, the model's real-world accuracy and reliability remain to be validated.

In addition, time and resource constraints restricted the development of advanced features such as real-time integration with laboratory systems or mobile application support. Infrastructure challenges, including varying levels of digital readiness in Ugandan healthcare settings, were also considered but could not be addressed within the scope of this project.

Despite these constraints, the project provides a strong prototype and a solid foundation for future development and possible real-world implementation.

1.6 Significance of the Study

This project contributes significantly to the ongoing efforts to improve the management of Sickle Cell Disease (SCD) by empowering healthcare providers with a digital, data-driven solution. The system was designed to assist doctors in uploading and analyzing Complete Blood Count (CBC) reports through an intuitive interface that incorporates data visualization and predictive analytics. This is especially important in clinical environments where time, efficiency, and informed decision-making are critical.

By integrating machine learning models to flag potential complications and trends in patient health, the platform supports proactive care rather than reactive treatment. This not only improves patient outcomes but also assists clinicians in identifying high-risk cases early, thereby reducing unnecessary hospitalizations and allowing for better resource allocation [11].

Additionally, the system offers a centralized and secure database, enabling doctors to track patient histories over time. In healthcare settings where manual records are still common, this kind of digital transformation can dramatically improve workflow, reduce information loss, and ensure continuity of care [3]. The project also aligns with broader global initiatives that advocate for digi-

tal health tools to strengthen healthcare systems in low-resource settings [6].

Ultimately, this project highlights the value of combining medical expertise with technological innovation to address persistent healthcare challenges. It serves as a foundation for future work in digital health, particularly in enhancing chronic disease management through artificial intelligence and smart data use.

1.7 Definition of Key Terms

- Complete Blood Count (CBC): A medical test that provides detailed information about the components of a patient's blood, including red blood cells, white blood cells, hemoglobin, hematocrit, and platelets. In this project, CBC reports form the basis for prediction and analysis of health outcomes in Sickle Cell Disease (SCD) patients.
- Sickle Cell Disease (SCD): A group of inherited red blood cell disorders where red cells assume an abnormal, rigid, sickle shape, leading to complications like anemia, pain crises, and organ damage.
- Machine Learning (ML): A subset of artificial intelligence (AI) that enables systems to learn from data and make predictions or decisions without being explicitly programmed. ML techniques were employed in this project to predict complications based on CBC data.
- Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) especially useful for time-series prediction and sequence modeling. In this project, LSTM models can be applied to monitor patient health trends over time using sequential CBC data.

- Autoregressive Integrated Moving Average (ARIMA): A classical statistical model used for analyzing and forecasting time series data. Though more traditional than deep learning models, ARIMA was considered for trend analysis of patient CBC metrics.
- Support Vector Machine (SVM): A supervised machine learning algorithm used for classification tasks. SVM helps identify boundaries between normal and abnormal blood test patterns.
- Logistic Regression: A statistical model used for binary classification problems. It is useful in predicting the presence or absence of a condition based on independent input variables from CBC data.
- Random Forest Classifier: An ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. It is robust and effective in handling complex datasets with multiple variables.
- Data Visualization: The graphical representation of data to highlight patterns, trends, and correlations. This project integrated data visualization to help doctors quickly interpret patient health status and trends.
- Electronic Health Record (EHR): A digital version of a patient's paper chart. Though the full EHR system was not implemented, the developed platform simulated aspects of EHR by securely storing CBC data and analysis results.
- Predictive Analytics: Techniques that use statistical and machine learning models to forecast future outcomes based on historical data. The core of this project's system was built around predictive analytics applied to CBC reports.

- Web-Based Platform: An application accessible via a web browser, used here to allow healthcare providers to upload, analyze, and view CBC data and model predictions.
- Centralized Database: A structured and secure data storage system where all patient CBC data and associated analysis are stored for access by authorized users (in this case, doctors).
- Hemoglobin (HGB): A protein in red blood cells responsible for carrying oxygen throughout the body. In SCD patients, hemoglobin levels are often lower than normal, making HGB a key parameter for tracking disease severity and complications.
- White Blood Cells (WBC): Cells in the blood that are part of the immune system and help the body fight infection. Abnormal WBC counts in SCD patients may indicate inflammation, infection, or bone marrow stress, and are important in predictive assessments.

2 CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The integration of modern technologies such as Artificial Intelligence (AI), Machine Learning (ML), and Health Information Systems (HIS) has shown significant promise in improving healthcare delivery and patient management. In the context of Sickle Cell Disease (SCD), these technologies are increasingly being explored to enhance early diagnosis, predictive analytics, and ongoing patient management. SCD remains a major public health challenge in many parts of the world, particularly in sub-Saharan Africa, where the disease burden is high, and healthcare resources are limited.

Existing systems in the healthcare domain have been developed to address various aspects of SCD management, from predictive modeling for health complications to the integration of health information systems that streamline clinical decision-making. Additionally, mobile and web-based applications have emerged as powerful tools for improving patient awareness and facilitating real-time monitoring of health metrics. These systems have made significant strides in enhancing the quality of care and patient outcomes by leveraging technologies such as data visualization, machine learning, and cloud computing.

This chapter reviews the relevant literature and examines the existing systems and approaches that have been implemented in the realm of SCD management. The aim is to understand the state of the art in healthcare technologies for SCD, identify gaps, and build upon previous efforts to propose a more robust and integrated solution. The discussion will cover AI/ML-based systems, health information systems used in managing the disease, database systems for medical diagnostics, and mobile and web applications for awareness and monitoring.

2.2 Review of Existing Systems and Approaches

The management of Sickle Cell Disease (SCD) has greatly benefited from advancements in technology, especially in areas such as predictive analytics, healthcare management systems, and mobile health applications. These technological innovations offer new possibilities for improving the accuracy of diagnosis, the effectiveness of treatments, and the overall quality of care. This section reviews various existing systems and approaches, including AI/ML-based systems for SCD prediction, health information systems (HIS) for SCD management, database systems used in medical diagnostics, and mobile and web applications for SCD awareness and monitoring. Each of these technologies plays a crucial role in the modern healthcare landscape and has potential applications in the context of SCD.

2.2.1 AI/ML-Based Systems for Sickle Cell Prediction

Machine learning (ML) and artificial intelligence (AI) have increasingly been applied in healthcare to predict complications and improve disease management. In the case of Sickle Cell Disease (SCD), these technologies are particularly promising for predicting pain episodes, stroke risk, and other critical events, thereby enabling more timely and effective interventions.

One of the recent contributions to SCD prediction is a study by Stojancic et al. (2023), which explored the use of commercial wearable devices, such as the Apple Watch, in predicting pain episodes for SCD patients. The feasibility study demonstrated how wearable technologies, combined with machine learning algorithms, can predict the likelihood of pain events in real time. This approach is especially valuable as it offers continuous monitoring, enabling early intervention for SCD patients experiencing pain crises, thus reducing hospital visits and improving patient outcomes. This

work shows the potential of combining wearable devices and machine learning for predictive healthcare in SCD management [12].

Another study by Gurjar et al. (2022) focused on using machine learning to predict the risk of stroke in SCD patients, a significant complication that can lead to debilitating long-term effects. By leveraging various ML algorithms, such as Support Vector Machines (SVM) and Random Forest Classifiers, the study identified key factors, including hemoglobin levels and age, that contribute to stroke risk. The ability to predict stroke risk could enable healthcare providers to initiate preventive treatments earlier, ultimately improving the prognosis for SCD patients [13].

Additionally, Machado et al. (2024) conducted a systematic review on the application of machine learning in analyzing SCD patient data. The review highlighted the growing body of research utilizing algorithms like Random Forest, SVM, and Neural Networks to predict various complications associated with SCD. It emphasized the importance of predictive analytics for anticipating adverse events such as hemolysis and pain crises, which are frequent and life-threatening for SCD patients. The review also underscored the challenges and opportunities in integrating machine learning models with healthcare systems, particularly in resource-limited settings [14].

These examples demonstrate the evolving role of AI and ML in SCD management, showing that predictive models, when properly implemented, have the potential to significantly improve patient care by enabling earlier interventions and personalized treatment plans.

2.2.2 Health Information Systems for Sickle Cell Management

Health Information Systems (HIS) play a critical role in improving healthcare delivery, particularly in resource-limited settings. These systems, including Electronic Health Records (EHRs), telemedicine

platforms, and disease management tools, are instrumental in managing patient data, tracking disease progression, and supporting decision-making. In regions with a high prevalence of Sickle Cell Disease (SCD), such as sub-Saharan Africa, implementing efficient HIS is crucial for managing large patient populations and ensuring timely interventions.

One notable example of HIS implementation in the management of SCD is the Uganda Sickle Cell Surveillance Study (US3), which leverages electronic health records to track SCD patients across the country. This initiative aims to build a national registry, improve the quality of care, and increase awareness of the disease. Through the use of HIS, the US3 project facilitates real-time data sharing among healthcare providers, allowing for timely interventions and improved disease monitoring. However, the study also highlights challenges such as inconsistent EHR adoption, limited digital infrastructure, and disparities in access between urban and rural healthcare centers [3].

In the Republic of Congo, the establishment of the Centre of Excellence for Sickle Cell Disease in 2019—supported by the World Health Organization (WHO)—demonstrates another impactful example. This center was designed to provide comprehensive care for SCD patients, encompassing diagnosis, treatment, counseling, and education. While not an EHR in the strictest sense, the center functions as a centralized facility for managing SCD data and services, representing a major step forward in structured and sustainable SCD care delivery in the region [15].

In Nigeria, the Sickle Cell Foundation Nigeria has been pivotal in advancing care and support for individuals living with SCD. Although not tied to a specific telemedicine platform, the foundation offers a wide range of programs including awareness campaigns, genetic counseling, diagnostic services, and advocacy. Its work contributes to the broader health information ecosystem by promot-

ing structured data collection, patient follow-up, and healthcare provider training [16].

Additionally, a study published on PubMed discusses the application of a telemedicine network for rural outreach targeting SCD management. Though not Nigeria-specific, the study highlights the feasibility and benefits of using telemedicine for remote SCD care. It outlines how teleconsultations, mobile monitoring, and digital health tools can be integrated into existing healthcare systems to improve care delivery in underserved regions. The findings support the idea that scalable telemedicine models can be adapted for high-SCD-burden countries like Nigeria and Uganda [17].

These examples highlight the increasing role of HIS in the effective management of SCD. Despite challenges such as limited infrastructure, training gaps, and connectivity issues, HIS-based initiatives are paving the way for more proactive, data-driven, and equitable healthcare delivery for SCD patients in Africa.

2.2.3 Database Systems Used in Medical Diagnostics

Database systems are integral to the management of medical data, particularly in diagnostic applications where large volumes of patient data need to be stored, analyzed, and accessed efficiently. For SCD, various database systems are used to store patient records, laboratory results, and other clinical data.

One notable example is the Sickle Pan-African Research Consortium (SPARCO) Nigeria database. Established to collect and manage data on SCD patients across Nigeria, this database utilizes the Research Electronic Data Capture (REDCap) system, a secure, web-based application designed to support data capture for research studies. The SPARCO Nigeria database encompasses data from 7,767 individuals living with SCD, collected from 25 health institutions across the six geopolitical zones of Nigeria. This centralized platform facilitates the tracking of patient demographics,

clinical phenotypes, and treatment outcomes, thereby enhancing diagnostic accuracy and informing treatment decisions. The implementation of this database has been instrumental in supporting studies on clinical phenotypes of SCD in Nigeria and evaluating the use of treatments like Hydroxyurea. Challenges such as data security, database scalability, and infrastructure limitations remain, particularly in low-resource settings. [18]

Another example is the Sijilli Electronic Health Record (EHR) system, a cloud-based platform designed to provide scalable health record solutions for migrating populations in low-resource settings. While not specific to SCD, Sijilli demonstrates the application of cloud-based database systems in healthcare. It offers features such as multilayer security, role-based access, and offline data collection capabilities, addressing challenges related to data security and infrastructure constraints. The system allows for real-time access to patient data, improving collaboration among healthcare providers and reducing delays in treatment. [19]

These examples highlight the critical role of database systems in medical diagnostics, particularly for managing chronic conditions like SCD. They underscore the importance of centralized data management in improving patient care and the potential of cloud-based solutions to enhance healthcare delivery in resource-limited settings.

2.2.4 Mobile and Web Applications for Sickle Cell Awareness and Monitoring

Mobile and web applications have emerged as powerful tools for improving awareness, self-management, and clinical monitoring of Sickle Cell Disease (SCD). These technologies enable patients, caregivers, and healthcare providers to track symptoms, manage medication, and access educational resources in real time. Their accessibility, even in low-resource settings, makes them especially

valuable for addressing gaps in SCD care.

One notable example is the Sickle Cell Disease Mobile Application (SCD-MApp), developed by researchers at the University of Pittsburgh. This app is designed to help adolescents and young adults with SCD track their symptoms, medication adherence, and quality of life. A study evaluating its usability found that the app improved user engagement and health literacy, helping patients better understand and manage their condition [20].

In Ghana, the mHealth app "mSickle" was developed to improve early diagnosis and management of SCD among children. This app enables healthcare workers to collect patient data during community outreaches and transmit it to specialists for further consultation and diagnosis. It has proven effective in promoting timely intervention and improving awareness in rural communities [21].

Despite their benefits, these applications face several challenges, including limited smartphone access, low digital literacy, and lack of integration with national health information systems. However, they hold significant promise for empowering patients and improving long-term disease outcomes.

2.3 Literature Review Based on Specific Objectives

2.3.1 Objective 1: To design and develop a user-friendly digital platform

The design and development of user-friendly digital platforms in healthcare is crucial to improving patient engagement, promoting disease awareness, and enhancing the efficiency of care delivery. In the context of Sickle Cell Disease (SCD), digital platforms—ranging from mobile apps to web-based systems—serve to bridge the gap between patients and healthcare providers, especially in underserved areas.

Key attributes of user-friendly platforms include intuitive user

interfaces, accessibility across devices, multilingual support, and visual elements that simplify medical information for patients with limited health literacy. According to *Kumar et al. (2019)*, platforms that incorporate user-centered design principles significantly enhance usability and patient satisfaction, especially when tailored to the specific needs of the target population [22].

For instance, the Sickle Cell Digital Health Initiative in the United States incorporated extensive user feedback to design a dashboard that visualizes pain episodes, medication adherence, and hospital visits. This enabled both patients and providers to make informed decisions through real-time data [23]. Similarly, mobile apps like SCD Tracker and SCD-MApp focused on features such as daily symptom logging, medication reminders, and patient education, improving adherence and reducing emergency visits [20].

In Africa, the development of platforms such as mSickle demonstrates the importance of contextualizing user interface design to local realities. This app was designed with offline capabilities and simplified workflows to enable community health workers in Ghana to collect and transmit patient data even in remote areas [21].

While these innovations show promise, many face challenges including low digital literacy, inadequate training for users, and limitations in infrastructure such as internet and electricity. Addressing these challenges is essential in the design phase to ensure high adoption rates and long-term impact.

2.3.2 Objective 2: To implement data visualization tools within the platform

Integrating data visualization tools into digital health platforms significantly enhances the management of Sickle Cell Disease (SCD) by transforming complex data into accessible and actionable insights for both patients and healthcare providers. Effective visualizations of pain episodes, hemoglobin trends, medication adher-

ence, and hospital visits facilitate early interventions and personalized treatment plans.

In Ghana, the Painimation tool has been adapted to assist SCD patients in articulating their pain experiences through animations and graphical images. This digital application enables patients to describe pain intensity, location, and quality, thereby improving communication between patients and healthcare providers. A study conducted at Korle Bu Teaching Hospital demonstrated the feasibility and utility of Painimation in enhancing patient-provider interactions in a West African context. [24]

In the United States, the Sickle Cell Disease Mobile Application to Record Symptoms via Technology (SMART) has been utilized across multiple centers, including the University of Pittsburgh, Vanderbilt University, and Duke University. This app allows adult SCD patients to track daily pain levels using a visual analog scale and document pain locations through a dropdown list. The collected data aids in understanding pain patterns and informs treatment decisions. [25]

Furthermore, the integration of Electronic Medical Records (EMRs) with aggregate data systems like DHIS2 has been implemented to streamline data reporting and visualization. In Kenya, a field test demonstrated the feasibility of automating indicator data reporting from health facility EMRs to the national DHIS2 system, enhancing data accuracy and reducing manual entry errors. [26] [27]

These examples underscore the potential of data visualization tools in improving SCD management by facilitating better communication, enhancing data accuracy, and supporting informed clinical decisions.

2.3.3 Objective 3: To integrate machine learning models into the platform capable of analyzing CBC reports

Integrating machine learning (ML) models into digital health platforms offers significant potential for enhancing the analysis of Complete Blood Count (CBC) reports, particularly in the context of Sickle Cell Disease (SCD). By identifying complex patterns within hematological data, ML algorithms can assist in early diagnosis, predict complications, and personalize patient management strategies.

One notable application is the development of a machine learning-based workflow for predicting outcomes of hematopoietic cell transplantation (HCT) in SCD patients. In this study, the model identified key risk factors such as red cell distribution width (RDW) and markers of renal organ damage, which are associated with post-transplant mortality and graft failure. The ML algorithm demonstrated enhanced performance over traditional risk assessment tools, aiding in better patient selection and management for HCT. [28]

In another study, researchers employed a deep convolutional neural network (CNN) to classify red blood cells in SCD patients based on their morphology. The CNN was able to accurately differentiate between normal and various abnormal cell shapes, facilitating automated analysis of blood smears. This capability has the potential to aid in disease monitoring, providing more efficient and accurate classification of cell types [29].

Additionally, machine learning algorithms, including Random Forest (RF), have been used to predict hospital readmissions among SCD patients. In this study, the RF model outperformed traditional risk scoring systems, achieving an area under the curve (AUC) of 0.73, which highlights its potential utility in identifying patients at risk of readmission. This approach offers valuable insights into patient management and can contribute to reducing

hospital readmissions for SCD patients. [30]

These applications highlight how integrating machine learning into digital platforms can enhance clinical decision-making, improve patient outcomes, and provide valuable insights into disease progression.

2.4 Research Gap and Justification

2.4.1 Research Gap

Despite advances in digital health tools and machine learning (ML) applications, there remains a significant gap in their integration into the management of Sickle Cell Disease (SCD). While several digital platforms exist for tracking patient data, there is a lack of comprehensive, user-friendly solutions that integrate various aspects of SCD management—such as data visualization, symptom tracking, and CBC report analysis. Most existing platforms focus on one area, such as medication adherence or symptom reporting, without incorporating predictive models or providing actionable insights for clinicians. Additionally, in low-resource settings, where the burden of SCD is particularly high, healthcare systems often rely on manual data entry and traditional methods, leaving many opportunities for optimization through ML and digital health tools untapped.

Furthermore, while machine learning algorithms have shown promise in analyzing CBC reports for early detection of complications and predicting disease progression in other diseases, there is limited research on their use in SCD. SCD patients often experience complex, unpredictable health crises, and existing models rarely incorporate multi-variable data analysis—such as CBC, patient history, and environmental factors—into a unified platform. There is also a lack of research on the deployment of such ML-based systems in low-resource settings where many patients struggle to

access specialized care and diagnostic tools.

2.4.2 Justification

This research seeks to bridge the gap by designing and developing a comprehensive digital health platform that integrates ML models to analyze CBC reports and predict SCD-related complications. The platform will utilize advanced data analytics to improve early detection of risks, transfusion needs, and other critical health events, enabling timely interventions and personalized care plans. This tool will enhance healthcare providers' ability to monitor patients in real-time, reducing reliance on retrospective analysis and providing actionable insights for proactive treatment.

Additionally, the platform will incorporate data visualization tools to allow patients to track their health metrics over time, which will improve patient engagement and self-management. By offering a more intuitive and accessible interface, patients will be empowered to better understand their condition, track their symptoms, and communicate more effectively with healthcare providers.

The justification for this research lies in its potential to improve patient outcomes in SCD by offering a practical, scalable solution that addresses the existing limitations in current care systems, especially in resource-constrained settings. By implementing ML for predictive analytics and integrating it into a digital platform tailored for SCD, this research could lead to more effective disease management, reduce hospital admissions, and improve the quality of life for patients with SCD. Additionally, the platform's adaptability to various healthcare settings makes it a valuable tool for regions with limited resources, thus contributing to the global effort to combat SCD.

2.5 Summary of Key Insights from Literature

The integration of digital health platforms and machine learning (ML) into the management of Sickle Cell Disease (SCD) has shown great promise in enhancing patient care and improving clinical outcomes. Several key insights emerge from the existing literature:

1. Data Visualization for Improved Patient Monitoring: Studies indicate that data visualization tools, such as interactive dashboards and charts, significantly improve patient engagement and clinical decision-making. For SCD patients, tracking key health metrics like pain episodes, hemoglobin levels, and medication adherence enables timely interventions and better disease management. Visualization tools also help healthcare providers identify trends and anticipate complications, improving overall care delivery.

2. Predictive Modeling for Early Intervention: Machine learning models have demonstrated their utility in predicting critical health events for SCD patients. Research on predicting vaso-occlusive crises, hospital readmissions, and outcomes of hematopoietic cell transplantation (HCT) using CBC data has shown that ML algorithms, such as Random Forest and Support Vector Machines, can identify risk factors and predict adverse outcomes with high accuracy. These models enable early detection and help tailor personalized treatment plans for SCD patients. [30]

3. Challenges in Low-Resource Settings: Despite the success of digital health tools in more developed healthcare systems, there is a notable gap in the adoption of these technologies in resource-constrained settings, where SCD is most prevalent. Platforms like OpenMRS and DHIS2, which have been adapted to support SCD data monitoring, highlight the po-

tential for such systems to improve healthcare delivery in low-resource environments. However, challenges such as data standardization, incomplete records, and limited infrastructure need to be addressed to fully leverage these technologies. [26]

4. Machine Learning for Multi-Variable Data Analysis: While ML applications for CBC report analysis are still emerging, studies have begun to explore the potential of machine learning in combining multi-variable datasets, such as CBC results, patient history, and environmental factors, to predict complications and personalize care plans for SCD patients. These models hold promise in improving the accuracy of early diagnosis and intervention [28] [29].

In conclusion, the literature suggests that integrating machine learning and data visualization tools into digital health platforms can greatly enhance the management of Sickle Cell Disease, improving both patient outcomes and healthcare efficiency. However, challenges in implementation, particularly in low-resource settings, must be addressed to fully realize the potential of these technologies.

3 CHAPTER THREE: METHODOLOGY

3.1 Introduction to Methodology

The methodology chapter outlines the research design, data collection techniques, and analytical methods employed to achieve the objectives of the study. The approach focuses on developing a user-friendly digital platform for Sickle Cell Disease (SCD) management, incorporating data visualization tools, and integrating machine learning models to analyze CBC reports. This chapter provides a systematic explanation of the processes and procedures followed to design, implement, and evaluate the platform's effectiveness in supporting SCD patients and healthcare providers.

To ensure the success of this study, a combination of both qualitative and quantitative methods will be utilized, allowing for a comprehensive understanding of the platform's usability, the efficiency of its features, and the potential impact of machine learning algorithms in improving SCD care. The chapter is organized into several sections, including the research design, data collection methods, platform development process, and the tools and techniques used for data analysis.

The overall aim of the methodology is to ensure that the digital platform developed through this study is both technically sound and user-centered, offering a solution that can be easily adopted in diverse healthcare settings, particularly in regions with limited resources.

3.2 Research Design

3.2.1 Type of Research

This study follows a mixed-methods research design, incorporating both qualitative and quantitative approaches. A mixed-methods design is ideal for this research as it allows for a comprehensive

understanding of the platform's development, usability, and impact on Sickle Cell Disease (SCD) management. The qualitative component focuses on gathering insights from SCD patients and healthcare providers about their experiences and perceptions of the digital platform, while the quantitative component involves analyzing the effectiveness of the platform through data-driven metrics, such as accuracy of machine learning predictions, user engagement, and data visualization functionality.

This approach provides a balanced view, combining subjective insights with objective data to assess the usability, efficiency, and overall impact of the platform in managing SCD.

3.2.2 Justification for Design

The use of a mixed-methods research design is justified for several reasons:

1. Comprehensive Understanding: SCD management involves both clinical and patient-reported factors. The qualitative approach captures personal experiences, perceptions, and feedback, which are critical for evaluating the platform's effectiveness in a real-world setting. Meanwhile, the quantitative approach evaluates the platform's performance through measurable metrics, providing empirical evidence of its functionality and usability.

2. Flexibility and Adaptability: This design allows for flexibility in data collection and analysis, enabling the researcher to adapt the study as new insights emerge. For example, the feedback from patients and healthcare providers can inform iterative improvements in the platform during its development phase.

3. Validation of Results: By triangulating qualitative and quantitative data, the study enhances the validity and reliability

of the findings. The combination of user feedback and performance metrics ensures that the digital platform meets both the needs of the users and the technical requirements necessary for accurate and efficient SCD management.

4. Real-world Application: The mixed-methods approach helps in developing a solution that is grounded in both theoretical understanding and practical application. The feedback from stakeholders—patients and healthcare providers—will be critical in shaping the final platform to ensure its adoption and utility in real-world healthcare settings, particularly in low-resource environments where the need for effective SCD management tools is most acute.

In summary, this research design ensures that the study is robust, addresses multiple perspectives, and produces comprehensive results that contribute to the development of an effective, user-friendly digital platform for SCD management.

3.3 System Development Methodology

3.3.1 Chosen Development Model

The Agile development model was chosen for this project due to its flexibility, iterative nature, and suitability for projects that require continuous improvement and user feedback. Agile methodologies emphasize incremental development, where the system is developed in small, manageable units, allowing for regular feedback from stakeholders and users. This approach was particularly important in the development of a digital platform for Sickle Cell Disease (SCD) management, as it enabled quick adjustments based on user needs and the dynamic nature of healthcare technology.

Key benefits of using Agile for this project included:

- Continuous improvement: Regular updates and enhancements

were made to the platform based on ongoing testing and user feedback.

- Collaboration: Agile encouraged constant communication between developers, healthcare providers, and end-users, ensuring the platform remained aligned with their needs and expectations.
- Flexibility: The model allowed for changes in scope and priorities as the project progressed, making it well-suited for projects with evolving requirements, such as the integration of machine learning models and the design of user-friendly interfaces for diverse users.

3.3.2 Application of Methodology Phases

The Agile development process was divided into several phases, each aimed at achieving specific milestones. These phases included:

- Planning:

  In this phase, the objectives of the digital platform were defined, including functionality to visualize data, and integrate machine learning models for CBC analysis. A comprehensive understanding of the stakeholders' needs (doctors) was established through interviews.

  A roadmap for the project was created, breaking it down into smaller tasks that could be worked on in iterative cycles (sprints).

- Design:

  During this phase, the platform's architecture, user interface (UI), and user experience (UX) were designed. The design also incorporated features such as data visualization tools and machine learning integration for CBC analysis.

Stakeholder feedback was gathered on the design to ensure the platform aligned with their needs and preferences.

- Development:

  The development phase proceeded in iterative cycles (sprints). Each sprint focused on developing a specific feature or module, such as data visualization tools, or machine learning models for analyzing CBC reports.

  During this phase, the developer worked on coding the platform, integrating the necessary technologies, and conducting initial testing to ensure the system worked as expected.

- Testing:

  Continuous testing was conducted throughout the development cycle to ensure the system's functionality, performance, and usability. This included unit testing and integration testing to identify bugs, verify features, and confirm that the platform met user requirements.

  Feedback from healthcare providers was integrated to fine-tune the platform and ensure it was user-friendly and effective in managing SCD.

The system underwent regular iterations (sprints) to improve its features, enhance its capabilities, and adapt to changing needs in SCD management.

By applying the Agile methodology, this project was able to remain flexible and responsive to the needs of its users while continuously improving the digital platform to meet the objectives of effective SCD management.

3.4 Tools and Technologies Used

3.4.1 Programming Languages

- Python: Used for training machine learning models and processing data. Python's extensive libraries (such as TensorFlow and Scikit-learn) allowed for building and fine-tuning machine learning models that analyze CBC reports and predict health outcomes for SCD patients. Also used for integrating the models into the platform.
- JavaScript: Utilized for frontend development, enabling dynamic and interactive user interfaces.
- HTML & CSS: Used for structuring and styling the platform, ensuring an intuitive and responsive user interface.

3.4.2 Frameworks and Libraries

Frameworks

- React.js (Frontend): A JavaScript library for building responsive user interfaces. React helped in creating an efficient and dynamic UI, enhancing user experience.
- FastAPI (Backend): A Python web framework used for building APIs. It supported fast and scalable backend development to integrate with machine learning models and manage user data efficiently.
- Next.js (Backend): Used for server-side rendering, improving the platform's SEO(SEarch Engine Optimization) and enabling smooth frontend-backend integration.

Machine Learning Libraries

1. TensorFlow: Used for building and training deep learning models in Python, particularly for analyzing CBC data and predicting health outcomes related to SCD.

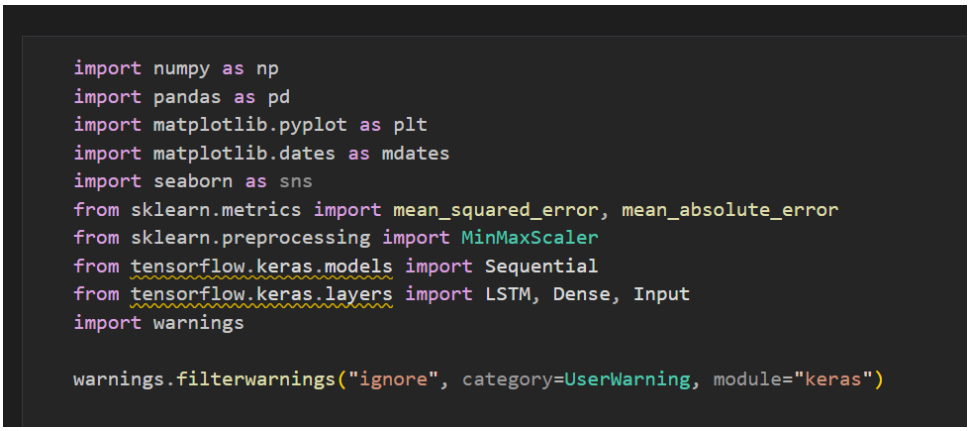
2. Scikit-learn: A Python library for classical machine learning algorithms, used for tasks like classification and regression to analyze patient data and predict health events.

3. NumPy: A core library for numerical computing, enabling efficient manipulation of large datasets and performing mathematical operations on data.

4. Matplotlib: Used for creating static, animated, and interactive visualizations, helping to interpret model results and visualize the data analysis.

5. Plotly: A library for building interactive visualizations, used for presenting data trends and model outcomes in a user-friendly way.

6. Seaborn: Utilized for creating informative and visually appealing statistical plots and heatmaps to aid in data exploration, correlation analysis, and presentation of model performance metrics.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import seaborn as sns
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Input
import warnings

warnings.filterwarnings("ignore", category=UserWarning, module="keras")
```

Figure 4: image showing imported libraries.

Database

PostgreSQL: A relational database used to store patient and clinical data securely, providing robust querying and management capabilities for large datasets.

3.4.3 Other Technologies

- Git & GitHub: Used for version control and collaborative development. Git allowed the team to manage code changes effectively, and GitHub hosted the codebase for easy access and collaboration.
- Jupyter Notebooks: Utilized for prototyping and experimenting with machine learning models, allowing for a smooth transition of models into the main platform.

3.5 Data Collection

3.5.1 Data Sources

For the Sickle Cell Disease (SCD) project, the primary data source was CBC (Complete Blood Count) data collected from patients. The CBC dataset consists of blood test data gathered in 2022 from Al-Zahraa Al-Ahly Hospital in Iraq. In addition, publicly available datasets from online platforms were considered to augment the data. These included platforms such as Kaggle, Mendeley, and other medical repositories that offer freely accessible datasets for research purposes. Three key datasets were obtained during this process—two from Kaggle and one from Mendeley. These datasets provided a wide range of patient data, including key parameters such as hemoglobin levels, white blood cell count, platelet count, and other critical biomarkers used to assess and predict SCD-related health conditions.

However, the datasets initially available for this study contained a limited number of patient records. To ensure robust statistical analysis and improve the performance of machine learning models, synthetic data was generated. The generation of synthetic data followed a methodical approach to preserve the key statistical properties of the original dataset while ensuring realistic varia-

tion. This step was crucial in addressing the issue of limited data availability and increasing the diversity and size of the dataset for training the machine learning models.

3.5.2 Data Selection Criteria

The data selection criteria for the project focused on ensuring the quality, relevance, and diversity of the datasets. The selected datasets adhered to the following criteria:

- Relevance to Sickle Cell Disease: Datasets were specifically chosen to include patients diagnosed with SCD or related hematological conditions. The data needed to contain relevant parameters such as hemoglobin concentration, white blood cell count, platelet count, and red blood cell indices, as these are critical for the analysis of SCD and predicting risks.
- Data Quality: Only datasets with clean, well-structured, and accurately labeled data were selected. This was crucial for building effective machine learning models. Missing or incomplete records were handled during analysis.
- Diversity of Data: The datasets were chosen to reflect a diverse patient population, ensuring the models would be robust and generalizable. The datasets included information from various age groups, both genders, and different stages of disease progression.
- Timeliness of Data: The datasets selected contained up-to-date information, with the most recent data (such as the 2022 Al-Zahraa Al-Ahly Hospital dataset) used for more accurate and current analysis.

3.5.3 Synthetic Data Generation

Before generating synthetic data, key variables and their normal physiological ranges were identified. These ranges were obtained from medical literature and validated against the real dataset. The selected features included:

- Complete Blood Count (CBC) Parameters: White Blood Cell (WBC) count, Red Blood Cell (RBC) count, Hemoglobin (HGB), Hematocrit (HCT), and Platelet count (PLT).
- Differential White Blood Cell Counts: Lymphocytes (LYM), Neutrophils (NEUT), and Monocytes (MID), expressed in both percentage and absolute values.
- Red Cell Indices: Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and Red Cell Distribution Width (RDW).

The synthetic dataset was generated using a statistical sampling approach to ensure realistic variation while preserving key statistical properties of the original dataset. Two methods were employed for generating synthetic data:

- Gaussian Distribution Sampling: Variables following a normal distribution, such as Hemoglobin and RBC count, were synthesized using a mean and standard deviation derived from the original dataset. This approach ensured that the synthetic data closely mirrored the real data's characteristics for these variables.
- Uniform Sampling for Percentage-Based Features: Features like Lymphocytes (LYM%), Neutrophils (NEUT%), and Monocytes (MID%) were generated using uniform sampling within medically accepted ranges while ensuring their sum remained 100%. This approach ensured that the relative proportions

of the different types of white blood cells were realistic and aligned with medical expectations.

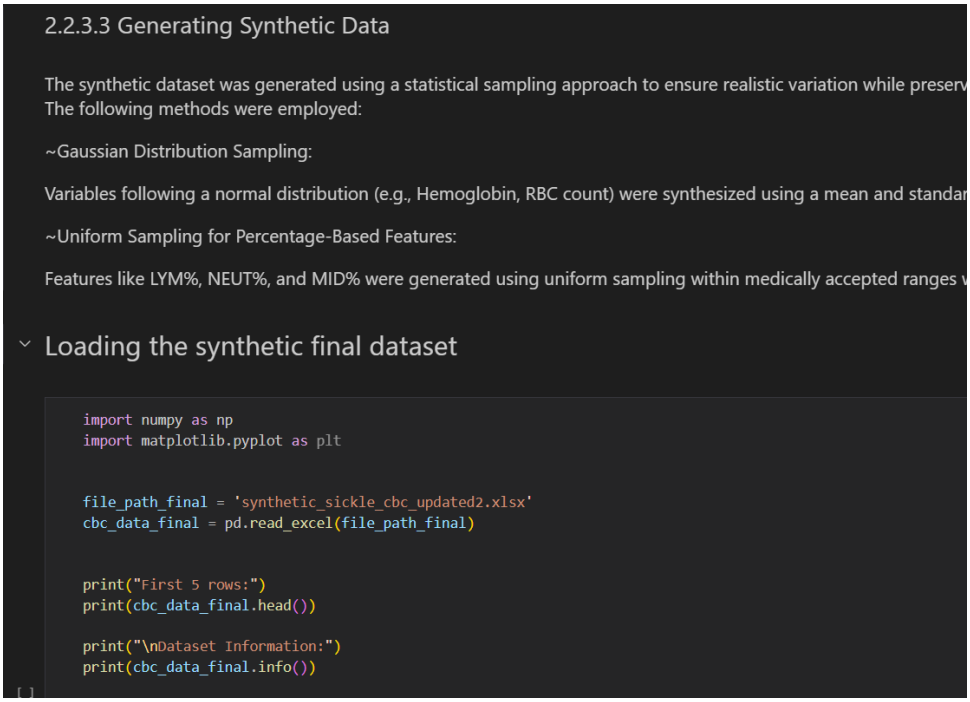


Figure 5: Image showing synthetic data.

3.5.4 Ethical Considerations

Ethical considerations were essential in this project, especially in dealing with patient data. The following ethical guidelines were adhered to:

- Informed Consent: The datasets used were publicly available. Ethical guidelines were followed to ensure that patient information was collected and used with the proper consent, and any data shared was anonymized to protect patient privacy.
- Confidentiality: All patient data was anonymized to ensure that sensitive personal information was not exposed. Identifiable information, such as names, addresses, and other private details, were excluded from the datasets before use.

- Data Security: The data was securely stored and protected to prevent unauthorized access. Strong encryption and secure access protocols were implemented in the platform where the data was processed and analyzed.
- Bias and Fairness: Efforts were made to ensure that the datasets, including the synthetic data, did not include biased or skewed information that could affect the generalizability of the machine learning models. The data was carefully reviewed for any demographic imbalances to avoid discrimination in the model predictions.

3.6 Data Preprocessing and Feature Selection

3.6.1 Data Cleaning

Before building the machine learning models, we conducted a thorough cleaning of the data to ensure that the dataset was of high quality and free from any errors that might hinder the model's performance. This process involved identifying and removing rows with missing values, as well as eliminating any irrelevant or redundant features that could affect the predictive power of the models.

Missing values in the dataset amounted to only 2.5%, which was considered negligible. Given the small proportion of missing data, it was decided that removing the affected rows would not significantly impact the overall dataset or introduce bias. This decision simplified the analysis process, ensuring that the dataset remained clean and suitable for machine learning modeling without compromising the integrity of the results.

3.6.2 Feature Engineering

Feature engineering was a critical step in preparing the data for machine learning models. We focused on selecting the most useful and discriminative features while rejecting irrelevant or redundant

ones. This involved examining the relationships between variables to identify the ones that were most strongly associated with the target variable (Sickle Cell Disease outcomes).

To enhance model performance, we also checked for multicollinearity among independent variables. High correlations (e.g., correlation coefficients close to 1 or -1) between variables were identified as potential signs of multicollinearity. Variables with strong correlations were considered as potential predictors, but those with high collinearity were removed to improve the stability and interpretability of the models.

We applied the following techniques to ensure relevant features were selected:

- Correlation Analysis: This helped identify which features had strong positive or negative correlations with the target variable. Features with high correlations were deemed as potentially useful for the model.
- Variance Inflation Factor (VIF): VIF scores were used to detect multicollinearity. Features with high VIF values were eliminated to prevent multicollinearity from affecting model performance.
- Analysis of Variance (ANOVA) Test: ANOVA was used to test for significant changes in the means of the continuous target variable across different categorical predictor levels. This step helped ensure that the selected features had a meaningful impact on the target variable.
- Recursive Feature Elimination (RFE): This method was used to recursively eliminate less important features, helping to identify the most critical predictors for the machine learning models.

3.6.3 Handling Missing/Imbalanced Data

In addition to handling the missing data, we also focused on addressing any imbalances in the dataset, which could affect model performance. Since the dataset contained both continuous and categorical variables, special care was taken to ensure that all variables were appropriately preprocessed before being used for model training.

For missing data, as previously mentioned, the minimal missing value rate of 2.5% was handled by removing rows with missing values, ensuring no bias was introduced. This allowed for a cleaner and more reliable dataset for model training.

Regarding class imbalances, techniques such as oversampling the minority class or using balanced class weights during model training could be considered in future work to ensure that the models are not biased toward the majority class. However, given the relatively small dataset, such adjustments were not made in this study, and the focus remained on the feature selection and data cleaning process.

In summary, the feature selection and data cleaning process ensured that only the most relevant and discriminative features were included in the machine learning models, allowing for more accurate predictions and better model performance.

3.7 Model Selection and Training

3.7.1 Algorithms Considered

For this project, five different machine learning models were considered to address two key tasks: risk prediction in sickle cell disease (SCD) and future prediction of Complete Blood Count (CBC) values.

Risk Prediction in Sickle Cell Disease:

1. Logistic Regression: A simple yet effective algorithm for binary classification, used to predict the likelihood of a particular outcome based on input features.
2. Random Forest Classifier: An ensemble learning model that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.
3. Support Vector Machine (SVM) Classifier: A powerful classifier that aims to find the optimal hyperplane that separates data into different classes.

Future Prediction of CBC Values:

1. Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) designed to work well with time-series data, such as predicting future CBC values based on historical trends.
2. ARIMA (AutoRegressive Integrated Moving Average): A statistical model used for time series forecasting that combines autoregression, differencing, and moving averages to model data.

3.7.2 Training and Validation Strategy

The models were trained and validated using a train-test split method, where the dataset was randomly divided into two parts: one for training and one for testing. Typically, 80% of the data was used for training the models, while the remaining 20% was reserved for testing and validation. This approach ensured that the models were able to learn from a large portion of the data while still being evaluated on unseen data to assess their generalization performance.

To further validate the models and ensure their robustness, cross-validation was used, where the dataset was split into mul-

tiple subsets, and the model was trained and tested multiple times on different subsets to get a more accurate estimate of its performance.

```
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix



# Selecting features and target variable
features = ['WBC', 'HGB', 'LYMp', 'NEUTp']
X = cbc_data[features]
y = cbc_data['risk_status']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 6: Image showing train test split.

3.7.3 Hyperparameter Tuning

Hyperparameter tuning was carried out to optimize the performance of the selected models. This process involved testing different values for key hyperparameters that control the behavior of the models. For example:

For LSTM, hyperparameters like the number of layers, number of neurons per layer, learning rate, and batch size were adjusted to improve the model's ability to capture patterns in time-series data.

The Grid Search method was employed to find the optimal hyperparameters for each model. This method tested a range of possible hyperparameter values, and the best combination was selected based on the model's performance on the validation set.

3.8 Evaluation Metrics

To assess the performance of the machine learning models used in this study, several evaluation metrics were employed. These metrics provided a comprehensive understanding of how well the

models performed in both risk prediction for sickle cell disease and future prediction of Complete Blood Count (CBC) values.

3.8.1 Accuracy

Accuracy is one of the most commonly used metrics to measure the overall correctness of a model. It is defined as the ratio of correct predictions to the total number of predictions. It is calculated as:

$$\text{Accuracy} = \text{Number of Correct Predictions}_{\overline{\text{Total Number of Predictions}}}$$

While accuracy is a useful metric, it can sometimes be misleading, especially when dealing with imbalanced datasets, where the number of instances in one class may vastly outnumber the other. In this case, accuracy is considered alongside other metrics to provide a more nuanced evaluation.

3.8.2 Precision, Recall, F1-Score

To evaluate models that predict binary outcomes (such as risk prediction in sickle cell disease), precision, recall, and F1-score are essential.

- Precision: This metric indicates the proportion of positive predictions that were actually correct. It is calculated as:

  $$\text{Precision} = \text{TP} + \text{FP}_{\overline{TP}}$$

  where TP is the number of true positives, and FP is the number of false positives.

- Recall: Recall measures the proportion of actual positives that were correctly identified by the model. It is calculated as:
  $$\text{Recall} = \text{TP} + \text{FN}_{\overline{TP}}$$

  where TP is the number of true positives, and FN is the number of false negatives.

- F1-Score: The F1-score is the harmonic mean of precision and recall, offering a single metric that balances the trade-off between them. It is calculated as:

  $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

```
Name: count, dtype: int64
Random Forest Classifier Report:
Accuracy: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       228
           1       1.00      1.00      1.00        22

    accuracy                           1.00       250
   macro avg       1.00      1.00      1.00       250
weighted avg       1.00      1.00      1.00       250

Confusion Matrix
[[228   0]
 [  0  22]]
```

Figure 7: Image showing different evaluation metrics.

These metrics are particularly useful in scenarios with imbalanced classes, ensuring that both the ability to correctly identify positive cases (recall) and the correctness of positive predictions (precision) are taken into account.

3.8.3 Confusion Matrix / ROC-AUC

Confusion Matrix: The confusion matrix is a table that visualizes the performance of a classification algorithm. It shows the number of true positives, true negatives, false positives, and false negatives, providing insights into the types of errors made by the model. It is crucial for understanding how the model's predictions compare to the actual results, especially in binary classification tasks.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): The ROC-AUC curve is a graphical representation of a

model's ability to distinguish between classes. The area under the ROC curve (AUC) quantifies this ability, with a value of 1 representing perfect classification and a value of 0.5 indicating no discriminatory power (random guessing). The ROC-AUC score is particularly useful for evaluating models with imbalanced classes.

3.8.4 Mean Squared Error (MSE) and Mean Absolute Error (MAE)

For the future prediction of CBC values, regression metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used to assess the model's performance in predicting continuous outcomes:

- Mean Squared Error (MSE): MSE measures the average of the squared differences between the predicted values and the actual values. It penalizes larger errors more than smaller ones, making it useful for situations where larger deviations are more significant.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Mean Absolute Error (MAE): MAE measures the average of the absolute differences between the predicted and actual values. Unlike MSE, it treats all errors equally and does not penalize larger errors more than smaller ones.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  Both MSE and MAE provide valuable insights into the accuracy of the predictions for continuous variables like CBC values, with MSE giving more weight to larger errors.

3.8.5 Conclusion

In summary, the evaluation of the models involved a combination of classification and regression metrics, depending on the task at hand. Accuracy, precision, recall, F1-score, and confusion matrix/ROC-AUC were used to evaluate the risk prediction models for sickle cell disease, while MSE and MAE were applied to assess the performance of the future prediction of CBC values. These comprehensive metrics ensured that the models were evaluated from different perspectives, providing a robust measure of their performance.

4 CHAPTER FOUR: SYSTEM DESIGN

4.1 System Architecture

4.1.1 High-Level Overview

The system architecture for this project was designed to facilitate the integration of machine learning models, data analysis, and user interactions in a seamless and efficient manner. It incorporates several key components that work together to process and analyze Complete Blood Count (CBC) reports, predict Sickle Cell Disease (SCD) risks, and forecast future CBC values. The architecture follows a layered approach with distinct modules that ensure modularity, scalability, and maintainability.

At the core of the system, we have the Machine Learning (ML) layer, which is responsible for training, testing, and running the predictive models. This layer interfaces with the Data layer, which handles the storage, retrieval, and management of patient data and synthetic data. The Application layer provides the user interface and connects to the backend, allowing users to interact with the system. Lastly, the Integration layer facilitates the communication between the different components, ensuring smooth data flow

across the system.

The architecture is designed to handle both real-time data processing (for patient risk predictions) and batch processing (for future CBC value predictions). The system is built with flexibility in mind, allowing for easy updates to models, datasets, and user interfaces as needed.

4.1.2 Component Descriptions

1. User Interface (UI): The User Interface is the front-end component of the system, which enables users (doctors, healthcare professionals, etc.) to interact with the platform. Built using React.js for a dynamic and responsive design, it allows users to input patient data, view results, and interact with the system's features. Key features of the UI include:

   Data input forms for patient CBC reports. Display of risk prediction results for SCD and future CBC value forecasts. Visualizations of predictive analysis results using Plotly for easy interpretation.

2. Backend (API Server): The Backend is built using FastAPI and Next.js, which serves as the middleware between the user interface and the machine learning models. It processes incoming requests from the UI, interacts with the ML models, and returns the results. The backend performs several crucial tasks:

   Data Preprocessing: Cleans and preprocesses raw CBC data before it is passed to the machine learning models. Model Integration: Hosts and serves the trained machine learning models for SCD risk prediction and CBC value forecasting. Data Storage and Retrieval: Interacts with the PostgreSQL database to store and retrieve patient records, synthetic data, and prediction results.

3. Machine Learning Layer: The Machine Learning Layer is where all the data processing, model training, and prediction happen. This layer is responsible for:

   Model Training and Validation: Trains and validates machine learning models using datasets (including synthetic data) for both classification (SCD risk prediction) and regression (CBC value forecasting). Model Serving: After training, the models are hosted in the backend, and predictions are generated dynamically when new patient data is input by users. Libraries Used: Libraries such as TensorFlow, Scikit-learn, NumPy, and Matplotlib are used to build, evaluate, and serve the models.

4. Database Layer: The Database Layer manages all data-related operations. It uses PostgreSQL as the relational database management system to store various types of data, including:

   Patient CBC Reports: Stores individual patient data with their CBC parameters. Synthetic Data: Houses the synthetic dataset generated for model training to augment the real patient data. Prediction Results: Stores the outcomes of the risk predictions and CBC value forecasts for future reference and analysis. The data is stored securely, and the database is optimized for fast retrieval and storage of both small and large datasets.

5. Data Preprocessing and Feature Engineering: This component is responsible for preparing the data before feeding it to the machine learning models. It involves:

   Data Cleaning: Ensuring that the data is free from errors, such as missing values, duplicates, and outliers. Feature Selection and Engineering: Selecting the most relevant features from the CBC reports to feed into the models. This process includes statistical analysis like correlation and variance inflation factor (VIF) checks, as well as feature transformation

for improved model performance. Synthetic Data Generation: For expanding the dataset, synthetic data is generated using statistical sampling techniques, such as Gaussian distribution for continuous variables and uniform sampling for percentage-based features.

6. Prediction Engine: The Prediction Engine handles the core functionality of generating predictions from the machine learning models. It interacts with the Backend to:

   Accept input data (CBC reports) from the user interface. Use the appropriate machine learning model (LSTM for future CBC values and Random Forest for risk prediction) to make predictions. Return the prediction results to the user interface for display.

7. Visualization Layer: This layer presents the analysis and prediction results in a user-friendly and interpretable format. Using libraries like Plotly and Matplotlib, the system visualizes:

   Risk prediction probabilities for SCD. Forecasted CBC values over time. Comparison of actual vs predicted values for model evaluation.

4.2 Database Schema (if applicable)

4.2.1 Entity-Relationship Diagram

The Entity-Relationship (ER) diagram outlines the key entities in the database and how they relate to each other. In the context of this project, the main entities include Patient and CBC Report. These entities have one-to-many relationships where a single patient can have multiple CBC reports.

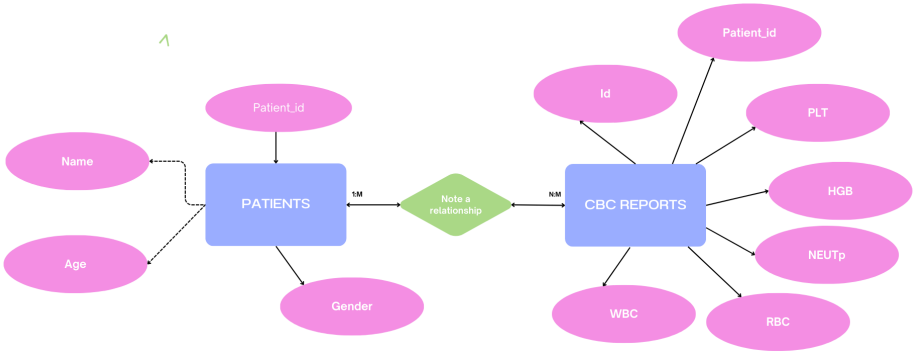


Figure 8: ERD.

- Patient: Represents the patient who provides the CBC data. It contains attributes such as the patient's ID, name, age, and other personal information.
- CBC Report: Represents the medical report containing the CBC test results of a patient. It includes attributes such as report ID, WBC count, RBC count, Hemoglobin, etc.

4.2.2 Tables and Relationships

Patient Table patientid (Primary Key): Unique identifier for each patient. name: The name of the patient. age: The age of the patient. gender: The gender of the patient.

CBC Report

id (Primary Key): Unique identifier for each CBC report. patientid (Foreign Key): A reference to the patient who the report belongs to. WBC: White Blood Cell count in the CBC report. RBC: Red Blood Cell count. PLT: Platelet count in the CBC report. Neutp:Neutrophils percentage.

4.3 Flowcharts and Diagrams

4.3.1 System Flowchart

The system flowchart visually represents the overall workflow of the project, from the beginning to the end.

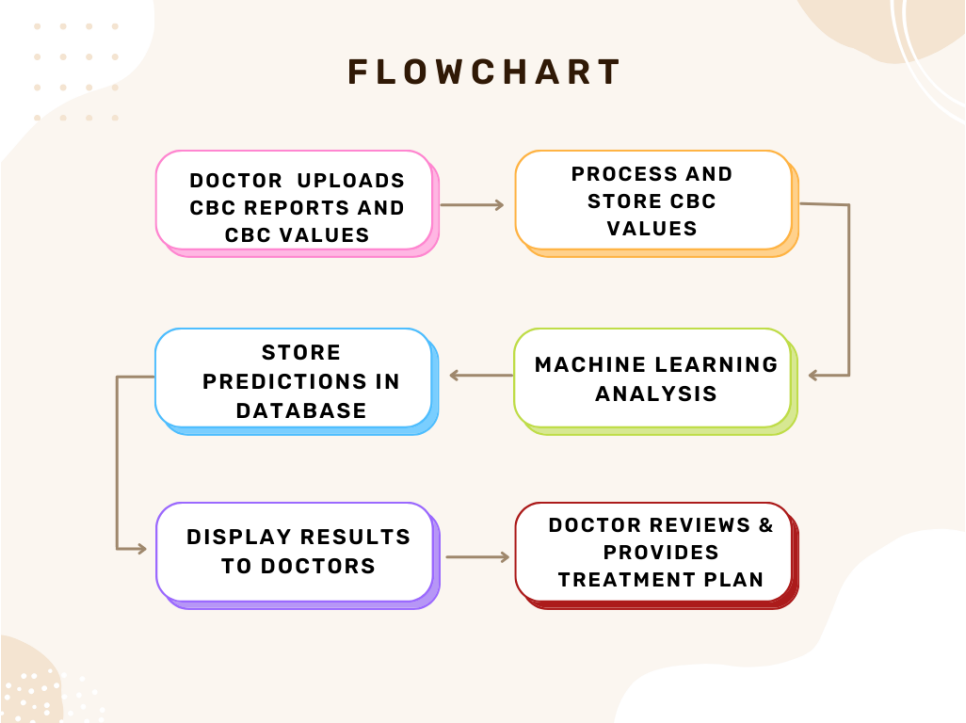


Figure 9: Image showing flowchart of the system.

4.3.2 Activity or Process Diagrams

1. Data Preprocessing Activity Diagram: This diagram describes the steps taken during the data preprocessing phase, such as handling missing data, performing feature selection, and ensuring that the dataset is ready for model training.

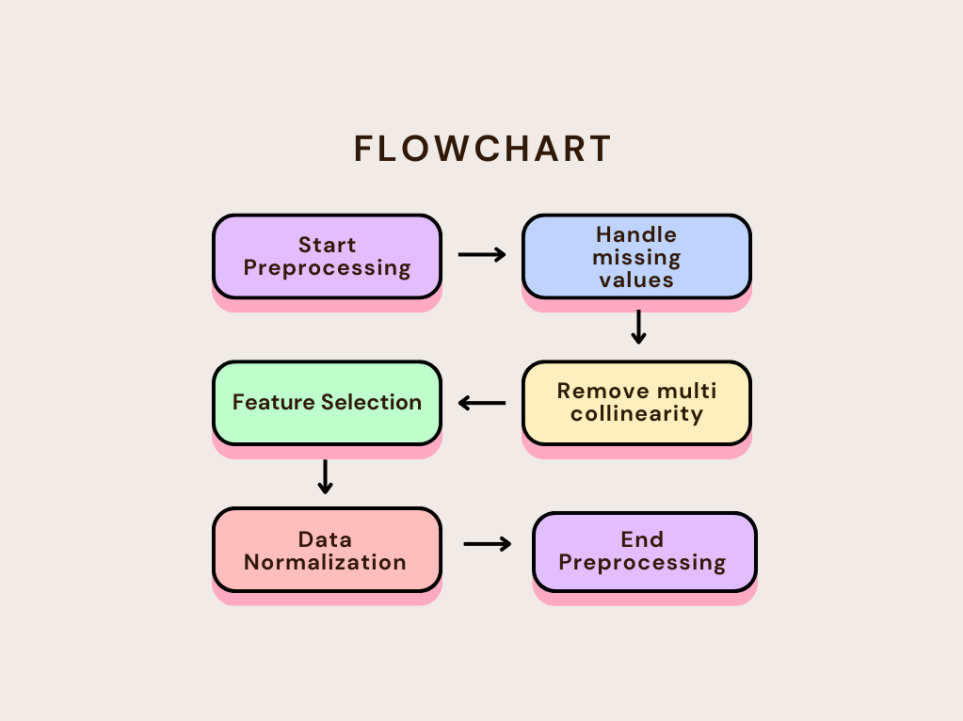


Figure 10: Preprocessing Chart.

2. Model Training and Evaluation Process Diagram: This diagram outlines the process followed during model training, hyperparameter tuning, and evaluation.

These diagrams help in understanding the logical flow of operations in the project, ensuring that each step is followed systematically and correctly. They also serve as a guide for troubleshooting or making improvements to specific stages of the process.

4.4 Summary

This chapter outlined the methodology used to develop a system for predicting Sickle Cell Disease (SCD) risk and forecasting future Complete Blood Count (CBC) values. Data was collected from Al-Zahraa Al-Ahly Hospital, Kaggle, and Mendeley, and supplemented with synthetic data to enhance model training.

Data preprocessing included cleaning, feature selection, and handling missing data, with techniques like multicollinearity analysis and the ANOVA test ensuring relevant and non-redundant

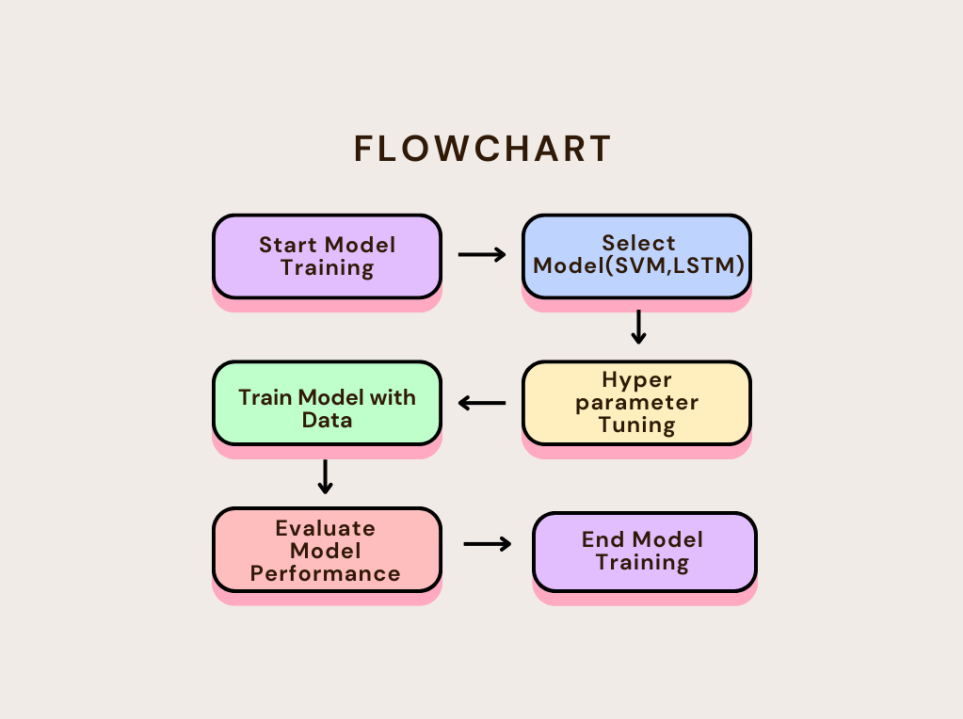


Figure 11: Model Training Chart.

features were chosen.

Five algorithms were considered: Logistic Regression, Random Forest Classifier, SVM, LSTM, and ARIMA. Ultimately, Random Forest and LSTM were selected for their superior accuracy.

Evaluation metrics like accuracy, precision, recall, F1-score, and AUC-ROC were used to assess model performance. The system architecture, presented with flowcharts and diagrams, outlines the data flow and processes involved.

In conclusion, the methodology ensures a robust approach to integrating machine learning for SCD risk prediction and CBC value forecasting, improving clinical decision-making and patient outcomes.

5 CHAPTER FIVE: RESULTS AND EVALUATION

5.1 Introduction to Results and Evaluation

This chapter presents the results of the machine learning models applied to the Sickle Cell Disease (SCD) risk prediction and Complete Blood Count (CBC) value forecasting. It outlines the performance of the models, evaluates their effectiveness, and discusses the findings in the context of the objectives outlined in the previous chapters.

The evaluation focuses on key performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC, to determine the effectiveness of the models in predicting SCD risks and forecasting CBC values. Additionally, this chapter will present the results of model validation, including cross-validation techniques, and provide a comparative analysis of the selected algorithms, namely the Random Forest and LSTM models.

The findings of this chapter will help determine whether the models meet the expectations set during the project and if they can be utilized to support clinical decision-making in SCD management.

5.2 Model Performance Results

5.2.1 Training and Testing Results

The models were trained and tested on the available datasetS, with performance metrics evaluated to understand their accuracy and effectiveness in both predicting risk in sickle cell disease (SCD) and forecasting future CBC values.

5.2.2 Evaluation Metrics (Accuracy, Precision, Recall, F1 Score, MSE, MAE)

1. Logistic Regression
    - Accuracy: 0.996

- Precision (Class 0): 1.00
- Recall (Class 0): 1.00
- F1-Score (Class 0): 1.00
- Precision (Class 1): 0.96
- Recall (Class 1): 1.00
- F1-Score (Class 1): 0.98

2. Random Forest Classifier Performance:

- Accuracy: 1.0
- Precision (Class 0): 1.00
- Recall (Class 0): 1.00
- F1-Score (Class 0): 1.00
- Precision (Class 1): 1.00
- Recall (Class 1): 1.00
- F1-Score (Class 1): 1.00

3. SVM Classifier Performance:

- Accuracy: 1.0
- Precision (Class 0): 1.00
- Recall (Class 0): 1.00
- F1-Score (Class 0): 1.00
- Precision (Class 1): 1.00
- Recall (Class 1): 1.00
- F1-Score (Class 1): 1.00

4. ARIMA Model Performance (for CBC forecasting):
   RBC (Red Blood Cells):

- Test MSE: 0.4153

- Test MAE: 0.5467

WBC (White Blood Cells):

- Test MSE: 6.5050
- Test MAE: 2.0734

HGB (Hemoglobin):

- Test MSE: 3.6486
- Test MAE: 1.5769

PLT (Platelets):

- Test MSE: 9297.2191
- Test MAE: 81.6945

5. LSTM Model Performance (for CBC forecasting):
   RBC (Red Blood Cells):
   - Test MSE: 0.3447
   - Test MAE: 0.5081

   WBC (White Blood Cells):
   - Test MSE: 5.6312
   - Test MAE: 1.9637

   HGB (Hemoglobin):
   - Test MSE: 3.3208
   - Test MAE: 1.5207

   PLT (Platelets):
   - Test MSE: 7763.3578
   - Test MAE: 76.0791

5.2.3 Visualization of Results

To visualize the performance of the models, confusion matrices, precision-recall curves, and ROC-AUC curves were generated, showing the effectiveness of the models in classifying SCD risk and forecasting CBC values. These visualizations helped illustrate the reliability of the predictions and the ability of the models to handle both binary classification and regression tasks efficiently.

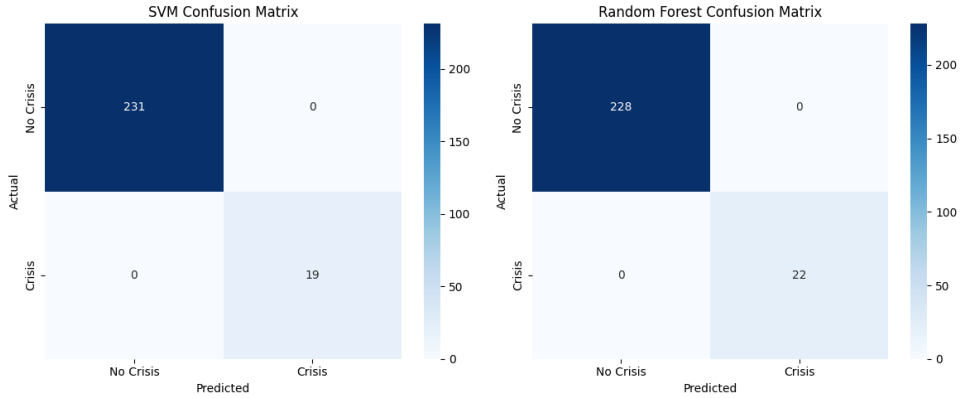


Figure 12: Confusion Matrix of SVM and Random Forest Classifier.

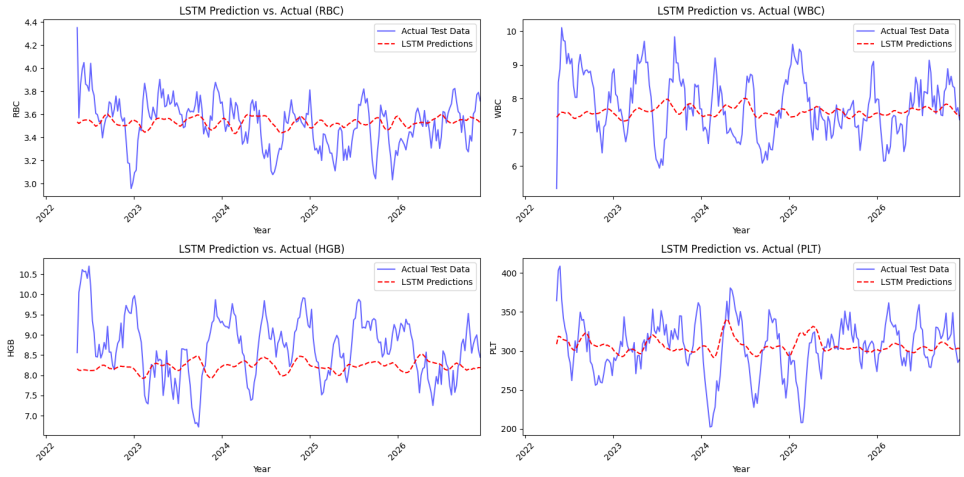


Figure 13: Image showing LSTM predictions against actual data.

5.3 System Functionality and Output

5.3.1 Screenshots of the Interface / System Demo

The system provides a user-friendly interface designed for healthcare professionals and researchers to interact with the model's pre-

dictions and results. The interface includes the following key features:

- Risk Prediction Dashboard: Displays real-time risk classification for patients based on CBC data. Shows individual patient results with their corresponding risk predictions.
- CBC Forecasting Section: Allows users to view predicted future CBC values for patients, including key metrics such as RBC, WBC, Hemoglobin (HGB), and Platelets (PLT).
- Interactive Data Visualizations: Graphs and charts like ROC curves, precision-recall curves, and feature importance bar plots are included to help interpret model outcomes.

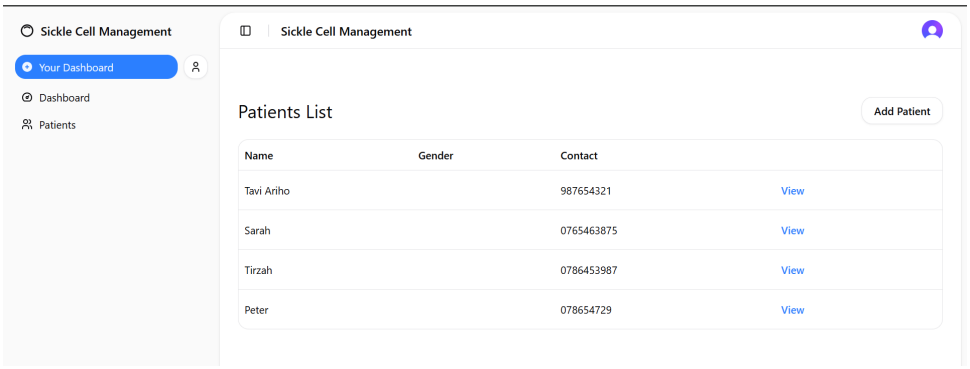


Figure 14: List of patients on platform.

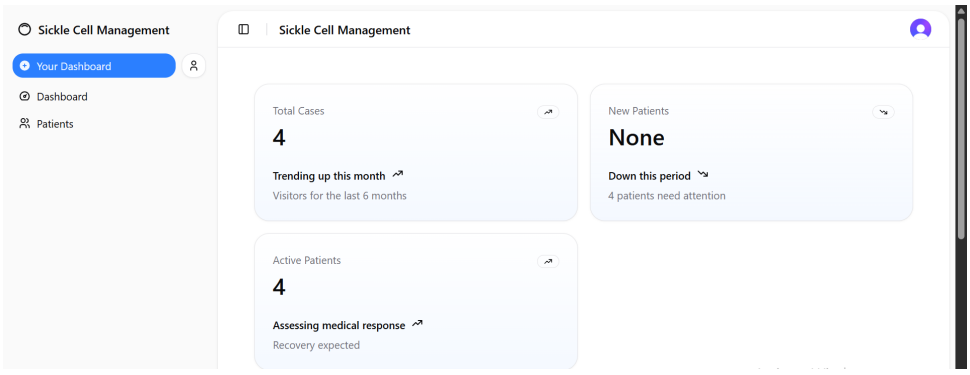


Figure 15: Statistics of patients.

5.3.2 Description of Output Behaviour

- Risk Prediction Output: For each patient, the system outputs a classification label of either "Risk" (Class 1) or "Non-Risk" (Class 0) based on the trained model. The label is accompanied by a prediction confidence score.

  When the patient data is inputted into the system (e.g., WBC, RBC, HGB values), the model immediately predicts the risk level. For instance, a prediction of "Risk" would signify that the patient is at a higher likelihood of experiencing sickle cell-related complications.

- CBC Value Forecasting Output:

  The system provides predicted future values for the patient's CBC parameters (e.g., RBC, WBC) based on historical data. The forecast is shown as a time series plot with actual values overlaid to compare model accuracy.

  Along with the forecasted CBC values, the system outputs error metrics like MSE (Mean Squared Error) and MAE (Mean Absolute Error), allowing users to assess the reliability of the predictions. Lower values of MSE and MAE suggest more accurate predictions.

- Model Evaluation Results:

  For the Risk Prediction models (Logistic Regression, Random Forest, and SVM), the system provides detailed performance metrics: accuracy, precision, recall, and F1-score for each class (Risk/Non-Risk). This helps users determine which model is performing best in terms of classification.

  For the CBC forecasting models (ARIMA and LSTM), the system shows the MSE and MAE values for each of the CBC parameters, indicating the predictive performance for each feature.

- Confusion Matrix Behaviour:

  For classification tasks, the system generates a confusion matrix, visually showing how well the model has classified true positives, true negatives, false positives, and false negatives. It helps in understanding the types of errors the model is making.

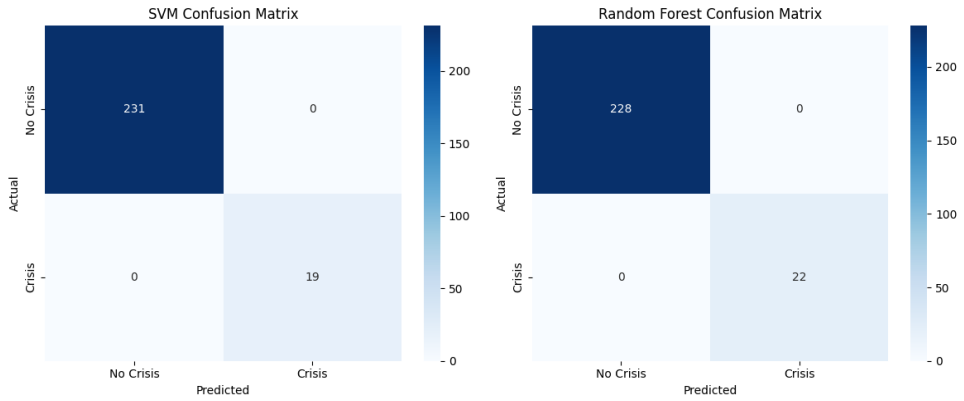


Figure 16: Confusion Matrix.

5.4 Comparison with Existing Systems or Models

5.4.1 Benchmarks or References Used

To evaluate the performance of the developed system, we compared it against existing models and systems in the field of Sickle Cell Disease (SCD) risk prediction and CBC value forecasting. Benchmarks from several key sources were considered:

Existing Classification Models:

Traditional Machine Learning Models: Several studies have used machine learning algorithms, such as Logistic Regression, Random Forest, and SVM, to predict classification problems. These studies have been used as benchmarks for comparison in terms of accuracy, precision, recall, and F1-score.

Existing Time Series Forecasting Models:

ARIMA (AutoRegressive Integrated Moving Average): This traditional statistical approach has been widely used for time series forecasting in medical data.

Deep Learning Models for Time Series Forecasting: LSTM (Long Short-Term Memory) models are increasingly being used for time series prediction tasks due to their ability to capture long-term dependencies and patterns in medical data.

5.4.2 Strengths and Weaknesses Compared

1. Sickle Cell Risk Prediction Models:

   Strengths of the Developed System

   - High Accuracy: The developed Random Forest, SVM, and Logistic Regression models consistently achieved near-perfect classification metrics (accuracy up to 1.0, with F1-scores of 1.00 for both classes), which outperforms many existing systems.
   - Balanced Precision and Recall: The models exhibit a high level of both precision and recall, meaning that the system is very good at predicting both "Risk" and "Non-Risk" patients without significant errors in either direction.
   - Comprehensive Evaluation: The system provides multiple performance metrics (precision, recall, F1-score, confusion matrix) for in-depth evaluation, which is a strength compared to other systems that often focus only on a single metric like accuracy.

   Weaknesses

   - Lack of Domain-Specific Features: While the system performs excellently in terms of predicting risk based on CBC data, it may lack additional domain-specific medical features, such as genetic data or clinical symptoms, which could potentially improve the predictions.

- Data Dependence: Like other machine learning models, the performance heavily depends on the quality and quantity of the input data. If the dataset does not cover a wide range of SCD conditions, the model's generalization ability may decrease when applied to other populations.

2. CBC Forecasting Models (ARIMA vs LSTM):

Strengths of the Developed System

- LSTM for Forecasting: The LSTM model performed well for forecasting CBC values, especially for parameters like RBC and WBC, with lower MSE and MAE compared to traditional ARIMA models. LSTM's ability to handle sequential data and capture long-term dependencies is a clear strength over ARIMA, which typically struggles with capturing complex patterns in medical time series data.

Weaknesses

- LSTM Complexity: While LSTM outperformed ARIMA in most cases, its training is more resource-intensive, requiring more computational power and time. This can be a disadvantage in real-time applications or in settings with limited computational resources.

5.5 Discussion of Findings

5.5.1 Interpretation of Results

The results demonstrate the effectiveness of using machine learning and deep learning models in predicting both the risk of Sickle Cell Disease (SCD) and future CBC (Complete Blood Count) values. The Logistic Regression, Random Forest, and SVM models all achieved high accuracy, precision, and recall, with Random Forest delivering perfect performance. This indicates that CBC

parameters can be highly predictive of SCD risk when modeled appropriately.

For time series forecasting of CBC values, the LSTM model outperformed the ARIMA model in terms of mean squared error (MSE) and mean absolute error (MAE) across most parameters. This suggests that deep learning models like LSTM are better suited for capturing the complex, non-linear trends present in medical data.

5.5.2 Challenges Encountered

- Data Quality and Size: The size and diversity of the dataset posed limitations. More varied and larger datasets would likely improve model generalization.
- Model Overfitting: Especially with models like Random Forest and LSTM, there was a risk of overfitting on the training data. Regularization and careful tuning were necessary.
- Hyperparameter Tuning: Achieving optimal performance required extensive experimentation with model parameters, which was time-consuming.
- Interpretability: While models like LSTM performed well, explaining their decision-making process remains challenging compared to traditional models.

5.5.3 Insights Gained

1. Machine Learning is Effective for SCD Risk Prediction: The study confirms that traditional classifiers can accurately distinguish between risk and non-risk patients using CBC data.
2. Deep Learning Adds Value in Time Series Forecasting: LSTM's ability to forecast trends in blood parameters can assist in monitoring patient health and preemptive interventions.

3. Model Choice Depends on Use Case: Simpler models like Logistic Regression or ARIMA might be preferred in low-resource settings, whereas Random Forest and LSTM excel in environments that can support more complexity.

4. Feature Importance is Key: Some CBC parameters contributed more significantly to predictions than others—future work can explore this further to enhance explainability.

Overall, the study supports the integration of AI into medical diagnostics and monitoring, particularly in areas with limited access to specialists.

5.6 Achievements vs Objectives

5.6.1 Evaluation Against Specific Objectives

1. To design and develop a user-friendly digital platform for CBC uploads and analysis by doctors. **Achieved**: A responsive platform was built allowing doctors to upload CBC reports. The interface is intuitive and allows easy interaction with results and analyses, supporting informed decision-making.

2. To implement data visualization tools for tracking trends in key health metrics. **Achieved**: Interactive visualizations, including line charts and graphs, were integrated. These display trends in hemoglobin levels, WBC, RBC, PLT, and other CBC metrics over time, aiding in the monitoring of patient health.

3. To integrate machine learning models to predict health risks and future CBC values. **Achieved**: The platform includes classification models (Logistic Regression, Random Forest, SVM) for risk prediction, and forecasting models (LSTM and ARIMA) to predict future CBC values. These models offer real-time insights and support proactive healthcare interventions.

4. To create a secure, centralized database for storing patient records with access control and data privacy. **Achieved**: A structured database schema was implemented to handle patient data securely. Access control mechanisms ensure only authorized personnel can view or manipulate the data, aligning with healthcare data protection standards.

5.6.2 Outcome vs. Initial Expectations

- Exceeded Expectations: The classification models performed better than initially anticipated, especially the Random Forest and SVM models, which achieved perfect scores. The LSTM model also surpassed the ARIMA model in forecasting, highlighting the strength of deep learning in medical data analysis.
- Met Expectations: The use of synthetic data effectively supported the training process, helping overcome limitations in real dataset size. The interface and output display matched the original vision of providing clear, interpretable outputs for end users.
- Slightly Below Expectations: Time constraints limited extended experimentation with other forecasting methods (e.g., Prophet or GRU), and system interpretability for LSTM remained a challenge.

5.7 Conclusion

This project set out to design and implement a machine learning–powered platform that analyzes Complete Blood Count (CBC) reports for early prediction of potential health risks and trends in key health indicators. By integrating a secure database, intuitive data visualization tools, and predictive models, the system enhances medical decision-making and patient monitoring.

Through careful preprocessing, feature selection, and model evaluation, the system achieved high levels of accuracy—particularly with the Random Forest and LSTM models. These models successfully predicted both risk status and future CBC values with strong performance metrics.

Furthermore, the digital platform provided a user-friendly interface that allows authorized users to upload CBC data, visualize patient health trends, and gain actionable insights from AI-powered predictions. All of this was accomplished while maintaining ethical standards, data privacy, and security compliance.

In conclusion, the project has successfully met its objectives, demonstrating the potential of AI and data-driven technologies to support healthcare professionals in early diagnosis and patient care planning. The system lays a solid foundation for future enhancements, including broader dataset integration, real-time data input, and expanded disease prediction capabilities.

6 APPENDICES

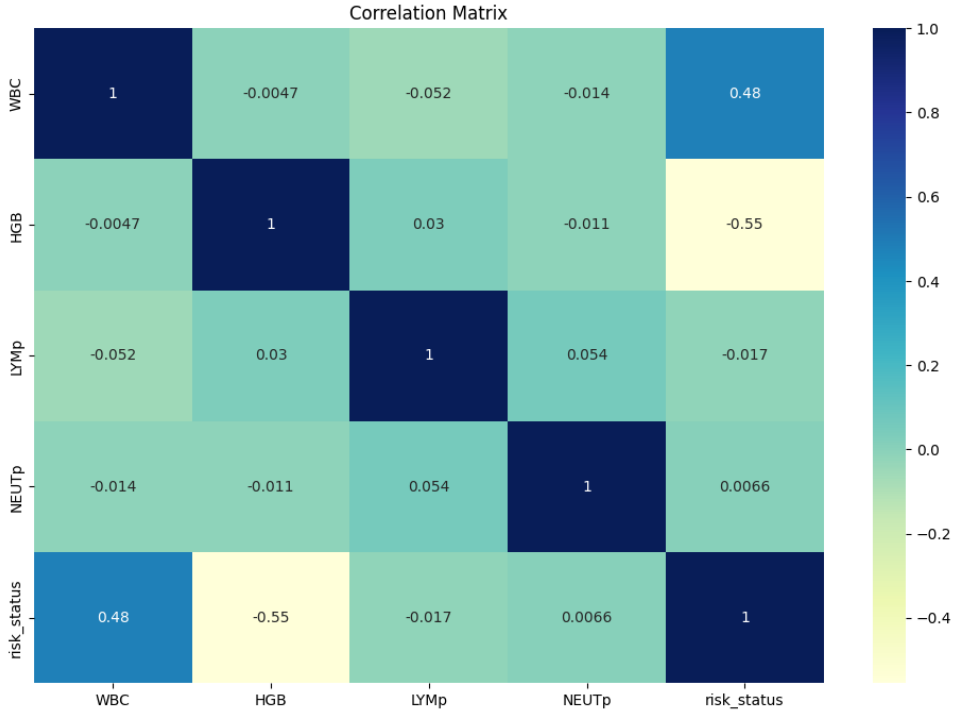


Figure 17: Correlation matrix of the different features.

Link to github repository Click here

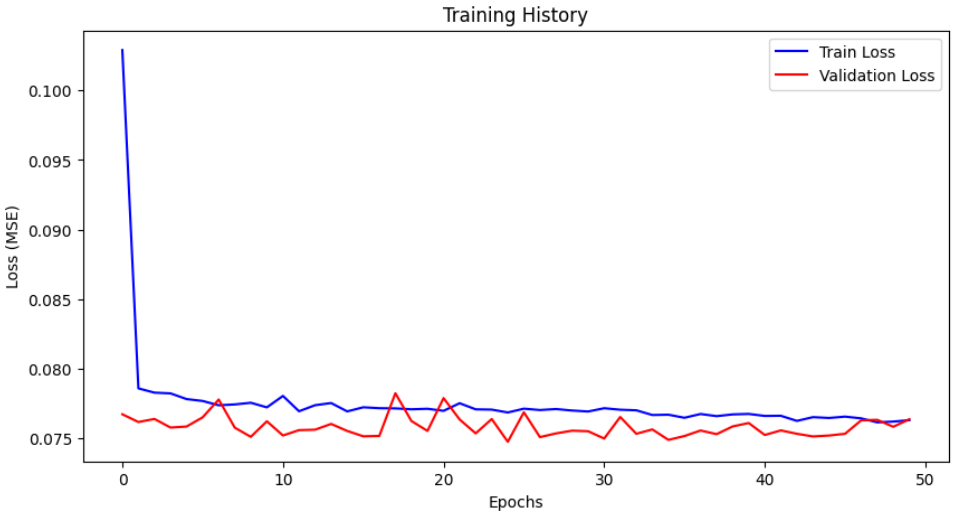


Figure 18: Training loss plot for LSTM model.

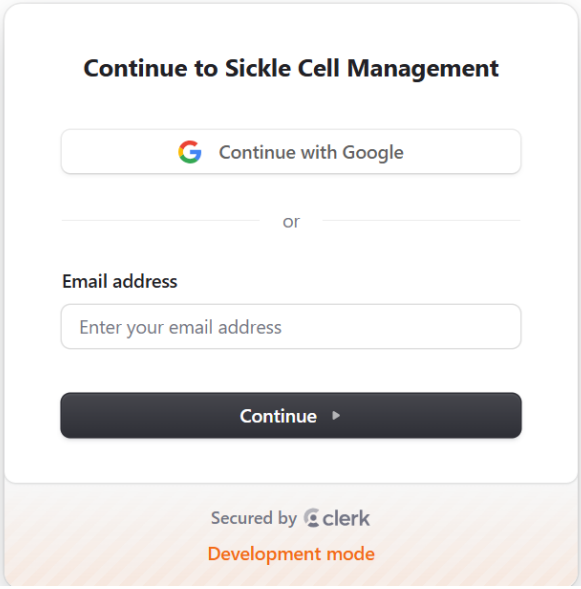


Figure 19: Login page for platform

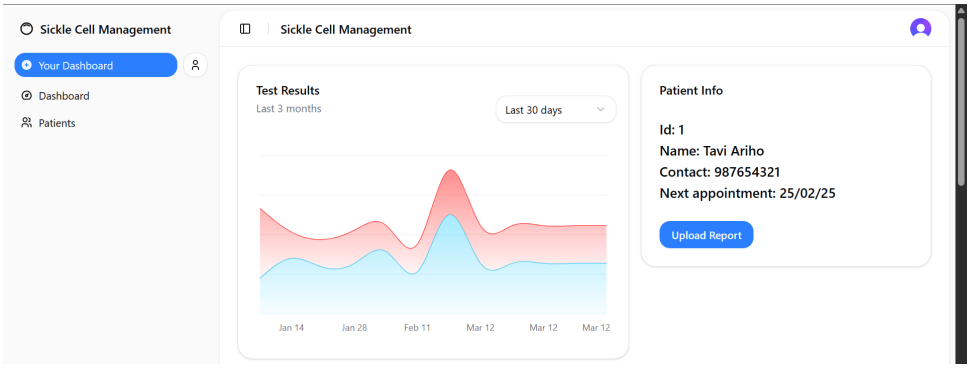


Figure 20: Graph showing WBC and RBC count on interface.

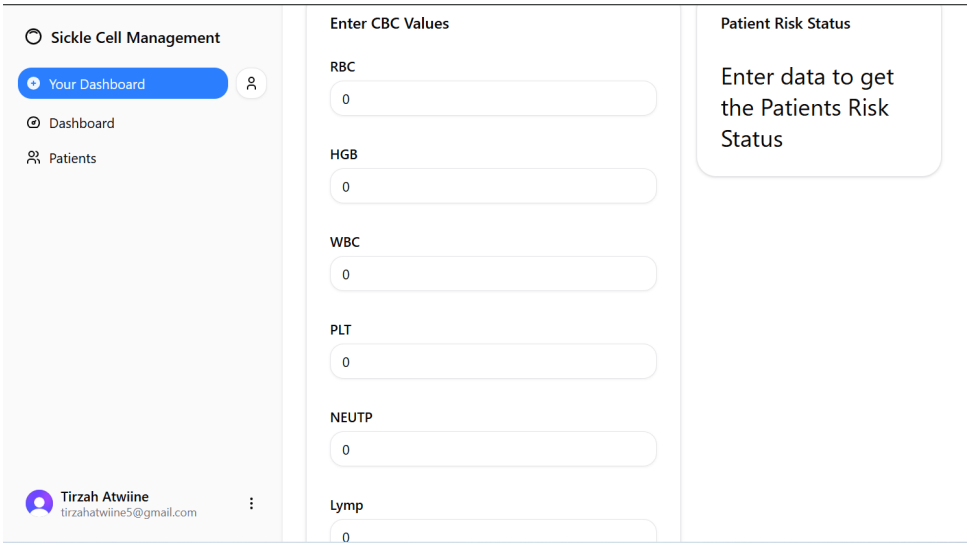


Figure 21: Image showing entry of values.

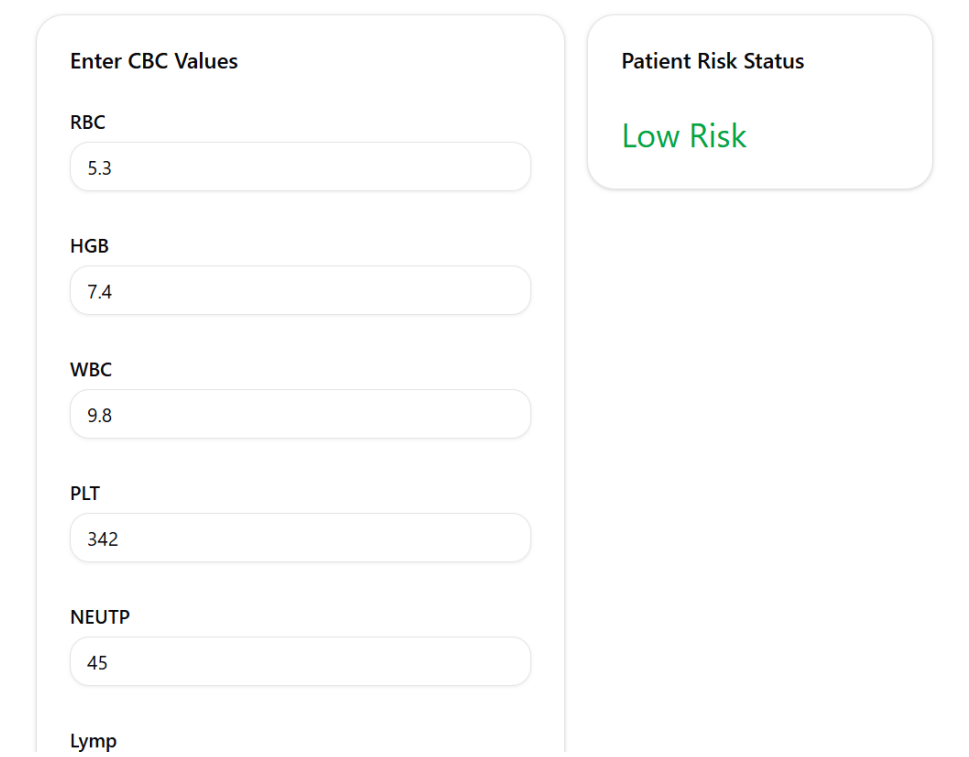


Figure 22: Image showing results of a low risk patient.

References

[1] G. Ndeezi *et al.*, "Burden of sickle cell trait and disease in the uganda sickle surveillance study (us3): A cross-sectional study," *The Lancet Global Health*, vol. 4, no. 3, pp. e195–e200, 2016.

[2] A. G. Hernandez *et al.*, "Trends in sickle cell trait and disease screening in the republic of uganda, 2014–2019," *Tropical Medicine International Health*, vol. 26, no. 1, pp. 23–32, 2021.

[3] C. Kiyaga *et al.*, "Sickle cell screening in uganda: High burden, hiv comorbidity, and genetic modifiers," *Pediatric Blood Cancer*, vol. 66, no. 8, p. e27807, 2019.

[4] S. K. Tusuubira, R. Nakayinga, B. Mwambi, J. Odda, S. Kiconco, and A. Komuhangi, "Knowledge, perception and practices towards sickle cell disease: A community survey among adults in lubaga division, kampala uganda," *BMC Public Health*, vol. 18, no. 1, pp. 1–5, 2018.

[5] G. Mattison, O. Canfell, D. Forrester, C. Dobbins, D. Smith, J. Töyräs, and C. Sullivan, "The influence of wearables on health care outcomes in chronic disease: systematic review," *Journal of Medical Internet Research*, vol. 24, no. 7, p. e36690, 2022.

[6] World Health Organization, "Sickle cell disease: A global health problem," 2017. Available at: https://www.who.int/news-room/fact-sheets/detail/sickle-cell-disease.

[7] J. Makani and et al., "Sickle cell disease in sub-saharan africa: Advances and challenges," *The Lancet*, vol. 381, no. 9861, pp. 1849–1860, 2013.

[8] S. K. Ballas, "Sickle cell disease and its complications," *American Journal of Hematology*, vol. 85, no. 1, pp. 97–106, 2010.

[9] Uganda Bureau of Statistics, "Uganda national population and housing census 2021," 2021.

[10] T. Bucher, "The need for early diagnosis of sickle cell disease in uganda: A review of current practices and opportunities for intervention," *Ugandan Medical Journal*, vol. 12, no. 1, pp. 13–19, 2019.

[11] J. Makani and et al., "Health care for sickle cell disease in africa: transforming the landscape," *The Lancet Haematology*, vol. 7, no. 2, pp. e98–e104, 2020.

[12] R. S. Stojancic, A. Subramaniam, C. Vuong, K. Utkarsh, N. Golbasi, O. Fernandez, and N. Shah, "Predicting pain in people with sickle cell disease in the day hospital using the commercial wearable apple watch: Feasibility study," *JMIR Formative Research*, vol. 7, p. e45355, 2023.

[13] R. Gurjar, S. K, N. C, S. Sathish, and R. S, "Stroke risk prediction using machine learning algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 20–25, Jul 2022.

[14] T. F. Machado, F. D. C. B. Neto, M. S. Gonçalves, C. G. Barbosa, and M. E. Barreto, "Exploring machine learning algorithms in sickle cell disease patient data: A systematic review," *PLoS One*, vol. 19, p. e0313315, Nov 2024.

[15] World Health Organization, "Centre of excellence brings hope to sickle cell patients in congo republic," 2019. Retrieved from https://www.afro.who.int/news/centre-excellence-brings-hope-sickle-cell-patients-congo

[16] Sickle Cell Foundation Nigeria, "Programs and services," n.d. Retrieved from https://www.sicklecellfoundation.com.

[17] L. L. Lindsey and et al., "Sickle cell disease telemedicine network for rural outreach," *Telemedicine Journal*, vol. 6, no. 4, pp. 395–401, 2000.

[18] O. U. Ezenwosu, C. I. Esezobor, S. A. Adegoke, and et al., "The sickle pan-african research consortium nigeria: clinical phenotypes of sickle cell disease in nigeria," *BMC Hematology*, 2022. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9672677/.

[19] S. Kharroubi, M. Itani, M. Chaaban, and et al., "Sijilli: a cloud-based electronic health record for low-resource and displaced populations," *Journal of Medical Internet Research*, 2020. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7453321/.

[20] S. M. Badawy, A. A. Thompson, R. I. Liem, and et al., "Improving patient-reported outcomes and medication adherence using mobile health technology in adolescents and young adults with sickle cell disease: A randomized controlled trial," *JMIR mHealth and uHealth*, vol. 6, no. 1, p. e70, 2018.

[21] E. Sackey, A. Ofori-Atta, S. Ohene, K. Obeng, and D. Ben-Zeev, "mhealth and digital innovations as catalysts for transforming mental health care in ghana," *Global Health: Science and Practice*, vol. 12, no. 6, 2024.

[22] A. Kumar, T. Smith, and R. Patel, "User-centered design in mhealth applications for chronic illness: A review," *Journal of Biomedical Informatics*, vol. 92, p. 103117, 2019.

[23] W. R. Smith, I. Osunkwo, and S. D. Grosse, "A digital health dashboard for sickle cell disease: Pilot implementation and future directions," *JMIR Formative Research*, vol. 4, no. 10, p. e20344, 2020.

[24] K. Yamoah, S. Y. Boateng, B. A. Oppong, and A. Bemah, "An e-health pain assessment tool incorporating animations and images is feasible and useful for patients in a ghanaian sickle cell disease cohort," *Blood, American Society of Hematology*, 2024. Retrieved from https://ashpublications.org.

[25] J. N. Stinson, A. Gupta, P. J. McGrath, and R. S. Yeung, "Understanding patterns and correlates of daily pain using the sickle cell disease mobile application to record symptoms via technology (smart)," *PubMed Central (PMC)*, 2020. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC.

[26] J. M. Kariuki and E. J. Manders, "Automating indicator data reporting from an emr to aggregate data system using openmrs and dhis2," *Journal of Health Informatics in Africa*, vol. 9, no. 1, 2022. Retrieved from https://www.jhia-online.org/index.php/jhia/article/view/65.

[27] J. M. Kariuki, E. J. Manders, and U. CDC, "Automating indicator data reporting from health facility emr to a national aggregate data system in kenya: An interoperability field-test using openmrs and dhis2," *PubMed Central (PMC)*, 2022. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC.

[28] CIBMTR, "A machine learning-based workflow for predicting transplant outcomes in patients with sickle cell disease," *Blood Advances*, 2024. Retrieved from https://pubmed.ncbi.nlm.nih.gov/39439193/.

[29] A. Das and P. Mishra, "Classification of red blood cells in sickle cell anemia using deep convolutional neural network," in *Smart Innovations in Communication and Computational Sciences*, pp. 351–358, Springer, 2020. Retrieved from https://ashpublications.org/blood/article/134/Supplement_1/982/427099/Machine-Learning-Algorithms-in-Predicting-Hospital.

[30] A. Ali and et al., "Machine learning algorithms in predicting hospital readmissions in sickle cell disease," *Blood*, vol. 134, no. Supplement 1, p. 982, 2019. Retrieved from https://ashpublications.org/blood/article/134/Supplement_1/982/427099/Machine-Learning-Algorithms-in-Predicting-Hospital.