DEVELOPMENT OF A HYBRID MACHINE LEARNING SYSTEM FOR SICKLE CELL DISEASE DETECTION AND CRISIS PREDICTION

BY

IHEKWABA KACHISICHO C.
GOU/U22/CSC/781

DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS
FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY
GODFREY OKOYE UNIVERSITY
ENUGU STATE

JUNE, 2026

TITLE PAGE

DEVELOPMENT OF A HYBRID MACHINE LEARNING SYSTEM FOR SICKLE CELL DISEASE DETECTION AND CRISIS PREDICTION

BY

IHEKWABA KACHISICHO C.

GOU/U22/CSC/781

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY, GODFREY OKOYE UNIVERSITY, ENUGU STATE, IN PARTIAL FULFILMENT OF THE AWARD OF BACHELOR OF SCIENCE (B.Sc.) DEGREE IN COMPUTER SCIENCE.

SUPERVISOR: DR. FRANK OKEBANAMA

JUNE, 2026

APPROVAL PAGE

This research, titled "DEVELOPMENT OF A HYBRID MACHINE LEARNING SYSTEM FOR SICKLE CELL DISEASE DETECTION AND CRISIS PREDICTION," has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Computing and Information Technology, Godfrey Okoye University, Enugu, Nigeria.

DR. FRANK OKEBANAMA **SUPERVISOR**	**SIGNATURE**	**DATE**
EXTERNAL EXAMINER	**SIGNATURE**	**DATE**
DR. FRANK OKEBANAMA **HOD, DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS**	**SIGNATURE**	**DATE**
PROF. I. A. AJAH **DEAN, FACIT**	**SIGNATURE**	**DATE**

CERTIFICATION

I certify that I completed the research included herein and that it was primarily based on my own observations and investigations. Consequently, I will fully accept the University's decision if I am found to have committed any form of academic misconduct in relation to this work.

_____ _____

IHEKWABA KACHISICHO C. **DATE**

GOU/U22/CSC/781

DEDICATION

This project is dedicated to Almighty God, whose grace sustained me through every challenge of this academic journey. It is also dedicated to my mother and family, whose unwavering love, prayers, and sacrifices made this achievement possible.

ACKNOWLEDGEMENTS

I am profoundly grateful to my supervisor, Dr. Frank Okebanama, for his patience, guidance and consistent encouragement throughout the course of this research. His expertise and dedication to academic excellence greatly shaped the quality and direction of this work.

I wish to extend sincere appreciation to the Dean, Prof. I.A. Ajah, Head of Department, Dr. Frank Okebanama, and all the lecturers in the Faculty of Computing and Information Technology, Godfrey Okoye University, including Prof. N.M. Ozofor, Mr. Camilus Njoku and Engr. Dr. Kingsley Chibueze for the solid academic foundation they have built in me over the course of my studies.

I am especially grateful to my mother and family for their endless support, encouragement, and prayers. Their belief in my ability to succeed kept me motivated even during the most demanding periods of this project. I also thank my friends and course mates who offered help, advice, and moral support whenever it was needed.

Above all, I give glory to God for the wisdom, strength, and perseverance to see this work through to completion.

ABSTRACT

Nigeria is a nation with the most cases of Sickle Cell Disease with approximately 150,000 cases born annually, yet existing diagnostic methods cannot predict crises and remain costly and equipment-dependent. This study addresses three gaps in prior Nigerian Sickle Cell Disease machine learning research; reliance on single algorithms, absence of integrated image and crisis prediction, and lack of a deployable web system. MediPredict, a hybrid Machine Learning based Sickle Cell Disease prediction system, was developed combining a MobileNetV2 CNN for blood smear image classification and a weighted Random Forest-XGBoost ensemble for clinical crisis risk prediction, deployed as a role-based web application for four actors; Administrator, Clinician, Laboratory Technician and Patient. A clinical dataset of 3,000 records was constructed from haematological parameter distributions derived from published Nigerian SCD studies including Nnodu et al. (2019), Ezenwosu et al. (2021), and Okon et al. (2024). The CNN achieved 97.43% accuracy and 99.76% AUC on 716 African blood smear images while the hybrid ensemble achieved 96.67% accuracy and 99.31% AUC — a 17.67 percentage point improvement over the 79% reported by Okon et al. (2024). The system supports image upload, crisis prediction, explainable patient results, and clinician feedback, demonstrating the feasibility of an integrated AI-powered SCD management system for the Nigerian healthcare context.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Definition of Sickle Cell Disease

Sickle cell disease is one of the forms of blood disorders that affect the Nigerian population. The condition can be termed as a genetic hematology disorder because it is caused by a genetic mutation in the DNA sequence, causing the creation of defective hemoglobin known as hemoglobin S or HbS (Machado et al., 2024). Under hypoxic conditions, the HbS molecules will clump together, causing rigidity of the RBCs, forming sickle-shaped cells, which block tiny blood vessels in the body. Nigeria plays a special role in the above situation as compared to other parts of the world. In Nigeria, about 150,000 newborns are recorded annually with SCD making up 50% of the total population across the world; and there is the presence of sickle cell trait within 20-30% of the population of Nigeria (Piel et al., 2013; Nnodu et al., 2019). The results of this condition among the people of Nigeria are devastating. The life expectancy rate for SCD patients in Nigeria is projected to be between 20 and 30 years compared to 40 to 60 years in other nations that have more developed health care programs to support the treatment of SCD (Odunvbun et al., 2008; Ezenwosu et al., 2021). Some of the causes that lead to the high mortality rate are; late diagnosis, inadequate availability of hematologists, expensive and rare technology used in diagnosis and managing the crises. Vaso-occlusive crises are the most dangerous form of SCD since they entail blocking of blood flow through the blood vessels. The development of the crisis because of pain, acute chest syndrome, splenic sequestration, and hemolytic crisis is normally the reason behind excessive visits to hospitals (Panepinto et al., 2020).

The conventional methods of diagnosing sickle cell disease in Nigeria, including haemoglobin electrophoresis, High Performance Liquid Chromatography (HPLC), and peripheral blood

smear microscopy, continue to be used to date. While they are efficient in establishing the existence of the disease and the patients' genotype, the main issue common to all of them is that none can predict when an attack is about to happen (Nnodu et al., 2019). Electrophoresis requires specialized equipment and skilled professionals found only in city hospitals. The level of newborn screening in the country is currently less than 15%, and thus, the majority of Nigerian babies have yet to be diagnosed (Odunvbun et al., 2008).

There is potential in machine learning. It has been shown that the use of a hybrid CNN-LSTM system can help in classifying sickled red blood cells in a blood smear image with an accuracy of 92.67%, according to Deo et al., (2024). The use of transfer learning from MobileNet along with the application of SVM resulted in 99.53% accuracy in classification as illustrated by Jennifer et al., (2023). In a Nigerian setting, it has been found by Okon et al., (2024) that ML models are able to predict the risk of SCD crisis based on clinical features; however, their highest level of accuracy was 79%. Nevertheless, most studies that have been conducted on these topics have viewed them separately, without developing an integrated model or framework for their use specifically within the Nigerian health sector (Machado et al., 2024; Tiwari et al., 2022). It is precisely such a void that will be filled by this study.

1.2 Statement of the Problem

SCD has brought about an unparalleled amount of burden on the Nigerian community, as well as its patients, their relatives and caregivers, not forgetting that the techniques used for diagnosing and treating SCD have been absolutely inadequate. The following points represent the backbone of this study:

- First, late diagnosis is quite typical. Nowadays, there are specific methods employed when diagnosing patients with SCD such as hemoglobin electrophoresis; however, this method involves the use of expensive equipment and professional personnel that are currently not present in the overwhelming majority of healthcare organizations located

in Nigeria, particularly rural Nigeria (Nnodu et al., 2019). In addition to this, the newborn screening program in Nigeria has never exceeded 15% (Odunvbun et al., 2008). Even though the manual blood smear under the microscope test is much more readily available than the other two tests, it is rather slow and prone to individual biases, whereas neither of them permits predicting when the patient will reach a crisis stage.

- Secondly, there is an issue related to the machine learning techniques used in the studies on predicting the SCD crisis. Only the authors who have utilized machine learning algorithms to predict the SCD crisis in Nigeria include Okon et al. (2024). They have successfully predicted the SCD crisis with a success rate of 79 percent using the machine learning algorithms known as random forest, SVM, and logistic regression. However, Farota et al. (2022) have found out that there is a 5% to 15% increase in the accuracy rate when multiple algorithms are used in predicting the crisis.
- Thirdly, there is no scientifically produced paper which would help us build a combined predictive system for Sickle Cell Disease with the integration of image analysis and the problem of a health crisis. Scientific studies mentioned by Machado et al., (2024) and Tiwari et al. (2022) relied heavily on the use of data from non-African countries and were not able to produce any applicable products. The current study tries to fill in all the mentioned gaps by creating a combined machine learning approach to predicting SC disease.

1.3 Aim and Objectives of the Study

The aim of this research is to develop a Hybrid Machine Learning System for Sickle Cell Disease Detection and Crisis Prediction. The specific objectives in accomplishing this aim are to:

1. Review existing machine learning approaches for SCD detection and establish gaps.

2. Design an integrated system architecture combining blood smear image classification and clinical parameter analysis for crisis prediction.
3. Implement the integrated system architecture using a hybrid machine learning approach.
4. Test the system through a user-friendly web interface to see the efficiency.

1.4 Significance of the Study

- **Patients: Early Diagnosis and Timely Intervention**

By providing information to patients on their prediction outcomes and understanding of their risks in a manner that can enable intervention before a crisis takes place. Also, alerting the high-risk individuals and health-care providers beforehand to facilitate prevention-based treatment approach to SCD management.

- **Healthcare Providers: A Decision-Support Tool for Clinicians**

This research work aims to provide medical professionals and lab technicians with a sophisticated and quantitative analysis software solution for analyzing images from blood smears and relevant clinical data, decreasing reliance on limited specialized knowledge. The software solution is not meant to take away from the clinical expertise but to enhance it with quantitative results.

- **Contribution to Academic Research**

This research contributes to the increasing body of Nigerian SCD machine learning research by demonstrating that hybrid ensemble approaches substantially outperform the single-algorithm methods used in prior studies. The system, dataset construction methodology, and evaluation results provide a replicable framework for future researchers seeking to develop AI-powered SCD tools calibrated to Nigerian clinical parameters.

1.5 Scope of the Study

- **Geographic focus:** The system is designed specifically for the Nigerian population, with clinical and haematological parameters.

- **Machine learning approach:** Only supervised machine learning techniques are employed, specifically a MobileNetV2 CNN for image classification and a weighted Random Forest-XGBoost ensemble for crisis risk prediction.
- **Exclusions:** The system does not cover genetic counselling, medication dosage recommendations, surgical intervention planning, or long-term outcome prediction. These remain important areas for future development.

1.6 Limitation of the Study

It is important to highlight a number of limitations that exist in this research. First, the major limitation relates to the source of clinical data required for training the crisis prediction algorithm. Acquiring patient-level clinical data in Nigeria's hospitals has been shown to be very difficult at the undergraduate level. Official permission takes up to 6 months to 12 months to be obtained following several levels of approval by the Institutional Review Board and hospital's managment, even sometimes the state/federal government. Efforts have been made to acquire the information from the published Nigerian SCD scientists like Okon et al. (2024), but no feedback was received from them during the project period. The result was the creation of the clinical dataset based on the distribution of parameters from the published Nigerian SCD research. This methodology is similar to what Machado et al. (2024) have mentioned in their systematic literature review of 29 studies in their paper. In any further research, ethical permission and institutional collaboration will be sought to get the information about real Nigerian SCD patient records.

The system is yet to be tested in practice within Nigeria's clinical environment, a gap which Machado et al. (2024) noted to be very common in the SCD ML literature and which will be addressed in subsequent studies.

1.7 Operational Definition of Terms

Sickle Cell Disease (SCD): These are genetic diseases affecting the body's hemoglobin; such diseases can be identified based on an unusual type of hemoglobin known as hemoglobin S (HbS), whose characteristic is to deform red blood cells into a shape resembling that of a sickle at low oxygen pressures.

Machine Learning (ML): A branch of artificial intelligence, as far as computer science is concerned, that learns from a particular dataset with the aim of predicting outcomes even when presented with new data without having been explicitly programmed for this purpose.

Hybrid Machine Learning: A computational technique that utilizes two or more algorithms/components from machine learning techniques and combines them to exploit the benefits from both approaches to deliver improved predictive outcomes compared to using any individual approach separately.

Convolutional Neural Networks (CNNs): A deep learning framework designed specifically for images, where the network applies automatic extraction of hierarchical spatial features from raw pixel inputs using learnable convolutions, for example cell shape and morphology.

Transfer Learning: In Transfer Learning, there is the use of a neural network that has been trained using a general dataset as a starting point for tackling a domain problem by fine-tuning it using the specific dataset; thus, reducing the need for training data.

MobileNetV2: A neural network model designed to be efficient from a computation standpoint by the use of depthwise separable convolution techniques. This model was created with the aim of ensuring that it can be implemented in systems with less powerful computing capability due to limited resources.

Random Forest: A machine learning approach that uses several independent decision trees to predict outcomes based on random selections from the training data set. The predictions made from each tree are pooled and evaluated based on majority voting.

XGBoost: An optimised version of gradient boosting which creates decision trees in a sequential manner with each subsequent tree attempting to rectify errors made by the prior tree. It has attained superior predictive results for structured datasets and includes an in-built mechanism for regularisation which helps avoid overfitting.

Ensemble Technique: The process of using machine learning that involves making use of the outputs of various models to derive a much more precise output. The individual models have to be formed separately.

Vaso-Occlusive Crisis (VOC): It is the most common complication of sickle cell disease, which is considered a medical emergency characterized by severe pain caused by obstruction of small blood vessels by rigid sickle cells leading to ischemia.

Haemoglobin Electrophoresis: This is a laboratory test for separation of different types of haemoglobin through electrostatic charge differences among the molecules, making it possible to identify the various types of haemoglobin such as HbS, HbA, HbC, and HbF. It is regarded as the gold standard technique for diagnosing SCD in Nigeria.

Foetal Haemoglobin (HbF): Type of haemoglobin which is usually found in foetuses and acts as an inhibitor for HbS aggregation. High levels of foetal haemoglobin in the body have been linked to milder disease phenotypes and lower crisis occurrences, hence being considered an important measurement tool during clinical assessment and crisis prediction.

AUC-ROC (Area under Curve-Receiver Operating Characteristic): Popular way of measuring classifier performance based on its ability to differentiate between two different classes at various classification levels, where a higher AUC-ROC score denotes superior classifier performance.

Data Augmentation: The term used to describe a set of image processing operations like rotating, flipping, zooming, adjusting the brightness and shifting of the images in the training

set to generate more data from an existing dataset and avoid overfitting of the machine learning model, especially important while using small datasets in SCD.

CRISP-DM (CRoss Industry Standard Process for Data Mining): A well-known methodology that helps in structuring and organizing the process of machine learning in six stages such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment, which were considered during this research to develop the ML modules.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The chapter offers a detailed review of all existing literatures about developing a detector and crisis predictor for Sickle Cell Disease through machine learning. The reason for carrying out this review is to provide insights into the extent to which things have been accomplished in this field, as well as the work left to do. Not only will this be crucial to the researcher but also to the general reader as well.

2.2 Theoretical Background

This section explains the theories that have been applied in this study, which include medical theories about the sickle cell disease as well as technological theories regarding machine learning and deep learning.

2.2.1 Overview of Sickle Cell Disease

Sickle Cell Disease (SCD) can be described as a group of hereditary diseases that arise from the alteration of the beta-globin gene leading to the formation of an abnormal form of haemoglobin called hemoglobin S (HbS). This results from the substitution of the sixth codon in the beta-globin gene located on chromosome eleven by one nucleotide (Kato et al., 2018). The HbS leads to stiffness in the cells of red blood, causing it to adopt a shape of a sickle when the oxygen concentration in the blood is low. The sickle-shaped cells cannot pass through the narrow blood vessels and may lead to complications like occlusion of the blood flow, anemia resulting from the hemolysis of the red blood cells. (Ware et al., 2017).

Genotypes for the disease exist; these vary depending on the genotypic type. The genotypic type known as HbSS is where a person acquires two copies of HbS genes from each parent, and it is the most common and severe type of the disease. Another genotypic type that exists is HbSC, which is relatively severe compared to others. The combination of HbS with

thalassaemia leads to a disease that varies in severity depending on the extent of remaining beta-globin function (Kato et al., 2018). The clinical manifestations of the disease vary widely among individuals, even within the same genotype. Management includes hydroxyurea administration to increase the amount of HbF, prophylactic penicillin and vaccination, and blood transfusion for severe anemia and stroke prevention; however, high-end treatment options like bone marrow transplant are not feasible due to unavailability in the Nigerian health care system (Ezenwosu et al., 2021).

2.2.2 Conventional Diagnostic Methods for SCD

For SCD diagnosis in Nigeria, the most trusted test is the haemoglobin electrophoresis, which involves separating haemoglobin molecules based on their electric charges and allows for the accurate detection of HbS, HbA, HbC, and HbF (Nnodu et al., 2019). High Performance Liquid Chromatography (HPLC) enables automated and precise quantitative detection of haemoglobin but requires expensive equipment costing at least $50,000 USD, making this test available only at major Nigerian tertiary hospitals. Both tests require a consistent power supply, skilled laboratory personnel, and dependable supplies of chemicals, none of which are always readily available in the healthcare facilities found in rural Nigeria.

The peripheral blood smear technique gives additional insight into the morphology of the cells but is less reliable and prone to inter-observer variability. It might perhaps also be worth mentioning that there is one crucial drawback which all of these conventional approaches face, and that is they are not capable of predicting whether or not a crisis is likely to happen in the future. All of these traditional methods can only determine that a particular disorder or a certain genetic marker is present, but nothing else. That is why the field of machine learning has much potential here.

2.2.3 Machine Learning in Healthcare

One of the advantages that machine learning holds is that the system has the capability to detect complicated patterns in the data sets and make forecasts without programming every function. This makes machine learning suitable for dealing with multi-dimensional data sets (Topol, 2019). The area of medicine has made a lot of advancements owing to the applications of machine learning that include diagnosing various diseases including cancer, heart diseases, diabetes, and even sepsis.

When it comes to predicting SCD, the best algorithm that can be applied here is supervised learning. For instance, SVM is the classification method used to predict SCD, using decision boundaries generated using the use of kernels for the maximization of the separation between classes, just like in the application of Okon et al., 2024. On the other hand, Random Forest constructs decision trees based on random samples, making predictions ensembles resistant (Roy et al., 2024). XGBoost uses sequential boosting by building new trees that try to reduce the errors of the previous trees, yielding cutting-edge models for structured tabular data (Farota et al., 2022). In the case of model evaluation in the medical field when using machine learning, a variety of measures must be taken into consideration because the cost associated with false positives and false negatives is different in every situation (Machado et al., 2024).

2.2.4 Deep Learning and CNN Architectures

Application of deep learning techniques, particularly the Convolutional Neural Networks technique, in image processing in medical practice has made a significant impact since learning is done on hierarchical feature representation from raw image pixels without any need for manual feature extraction (Esteva et al., 2019). The convolutional layers of a CNN use learnable filters to identify increasingly complex spatial structures in images that range from basic edge and texture detection in shallow layers to advanced structures such as cell morphology and sickling traits in deeper layers.

Transfer learning greatly enhances the usage of CNNs for small-scale medical images by utilizing models that have been trained on large image databases like ImageNet, where the model learns to recognize valuable low-level visual features applicable to other problems (Jennifer et al., 2023). Moreover, MobileNetV2 which was introduced by Sandler et al. (2018) is particularly relevant to this case because of its effective use of depthwise separable convolutions. This model can produce similar output with lesser computational power as compared to deeper models, such as ResNet-50 and VGG-16, and hence, becomes relatively easier to implement on regular computer systems present in Nigerian hospitals. The technique of Data Augmentation, involving techniques such as rotation, flipping, change in brightness, zooming, and shifting, is vital as a measure of fighting overfitting and increasing data set size artificially when faced with limited data sets of medical images (Deo et al., 2024).

2.2.5 Hybrid Machine Learning Approaches

Hybrid machine learning involves combining two or more algorithms/architectures in order to benefit from the synergy created by leveraging the strengths of each method. According to Farota et al. (2022), there was a systematic study of ensembles and hybrid machine learning in several medical domains where prediction is key – cancer detection, heart diseases, diabetes, and neurological disorders – in which an improvement in accuracy of 5 to 15% over using single algorithms was observed in all cases. Most importantly, Farota et al. (2022) established that it was model diversity, rather than sheer quantity of similar models, that contributed to successful predictions. This is a factor that directly impacts the decision to incorporate the use of Random Forest together with XGBoost, whereby the former utilizes parallel bagging to minimize variance while the latter utilizes sequential gradient boosting to minimize bias.

Also, when applying the technique of analyzing images for the SCD problem, the results obtained by hybrid approaches were better compared to other techniques. Deo et al. (2024) found out that a combination of using a CNN to extract spatial features together with an LSTM

model for sequential processing led to a blood cell classification rate increasing from 90.13% to 92.67%. Jennifer et al. (2023) discovered that using traditional ML models instead of the last layers of transfer learning models led to an even higher accuracy rate of 99.53%. This body of evidence demonstrates that hybrid approaches represent the present-day best practice in relation to SCDs ML applications, thus constituting the theoretical background of the proposed solution implemented in this study.

2.3 Review of Related Works

This section includes a critical analysis of fifteen studies that focus on machine learning methods for Sickle Cell Disease detection and Crisis prediction.

2.3.1 Deo et al. (2024)

A hybrid CNN-LSTM model was created by Deo et al. (2024) in order to classify red blood cells in Traitement du Signal. With 260 Indian hospital images of blood smears extended to 750, the CNN-LSTM model yielded an accuracy rate of 92.67%, beating both the CNN InceptionV3 model (90.13%) and the Extreme Learning Machine (87.6%). This research demonstrated that integration of CNN and LSTM is effective for improving classification accuracy. Drawbacks of the research included the use of Indian and not African images, and the high computational cost of LSTM.

2.3.2 Machado et al. (2024)

Machado et al. (2024) carried out a PRISMA systematic review of 29 ML studies related to SCD featured in PLoS ONE. For diagnostic purposes, deep learning models recorded accuracies between 85% and 99%, whereas their performance in monitoring crises had accuracies ranging between 63% and 92%. Some of the issues noted included inadequate African data, lack of model interpretability, inadequacy of clinical validation, and failure of integration in multi-modal systems. This review, therefore, forms a basis for the proposed study.

2.3.3 Jennifer et al. (2023)

The comparison of five models based on transfer learning for SCD cell identification was made by Jennifer et al. (2023), published in Heliyon. The highest efficiency (99.53%) was obtained using MobileNet-SVM hybrid model in 15 ablations on 10,002 augmented images. The study showed the advantage of using ML methods instead of deep learning classification layers, which is a fundamental idea used in the current research. One important limitation of this method is the use of one non-African database.

2.3.4 Okon et al. (2024)

Machine learning methods for predicting the likelihood of sickle cell crisis using clinical factors of Nigerian patients were assessed by Okon et al. (2024), with results appearing in Scholars Journal of Engineering and Technology. Data set for the analysis was sourced from Immanuel Infirmary and the University of Calabar Teaching Hospital. Three different approaches - SVM, Random Forest, and Logistic Regression - were used on the data set through a partitioning approach of 80-20 data split coupled with k-fold validation, where k = 8. Both Random Forest and Logistic Regression reached about 79% accuracy, while SVM had 75% accuracy. The main strength of this method is its reliance on data obtained from patients in Nigeria, which makes it the closest work to the present study. The main weaknesses of this study include the utilization of only one algorithm, lack of imaging-based study and moderate level of accuracy that can be improved through a hybrid model.

2.3.5 Tiwari et al. (2022)

Tiwari et al. (2022) conducted a review of deep learning and machine learning techniques for SCD detection in SGGVS-AIMT conference proceedings. Deep learning had 91% to 100% accuracy in the studied papers and surpassed traditional ML. Challenges such as lack of datasets from Africa and implementation difficulties were noted in the review. This research does not offer any experiments but creates justification for the proposed project.

2.3.6 Farota et al. (2022)

According to Farota et al. (2022), there is proof that a hybrid and ensemble model will perform 5-15% better than any individual algorithm for predicting cancer, cardiovascular diseases, diabetes, and neurological disorders. Model diversification has been found to be the reason behind such high efficiency, which makes an RF-XGBoost model applicable in this study. However, the major drawback is that results have not been used for SCD specifically or validated in Nigeria.

2.3.7 Roy et al. (2024)

Six machine learning algorithms were compared by Roy et al. (2024) for predicting the severity of SCD using hematological features during the PHOTOPTICS 2024 conference. The Random Forest model was found to have the best performance overall. Based on feature importance analysis, hemoglobin level, HbF level, and reticulocyte count were selected as the most important features to predict SCD severity. Limitations included the lack of an ensemble model approach and lack of data from Nigeria/Africa.

2.3.8 Ayoade et al. (2023)

An ensemble ML model for the prediction of SCD from erythrocyte smears has been designed by Ayoade et al. (2023) and documented in EAI Endorsed Transactions on Pervasive Health and Technology. The model shows high efficiency as Random Forest obtained an accuracy of 99% AUC alone, and even better accuracy after applying RF-XGBoost hybridisation due to lower overfitting. This supports the hybrid approach in this investigation. Restrictions are image classification only and non-African data source.

2.3.9 Goswami et al. (2024)

Goswami et al. (2024) integrated the concept of deep neural networks and explainable AI Grad-CAM model for the purpose of detecting SCD. The application of Grad-CAM generated output that revealed what parts of images affected the particular classification results. That is the main

reason for designing simple and explanatory pages in this study. It is noteworthy that the major drawback of the technology is its expensive computation.

2.3.10 Gollapalli and Alfaleh (2022)

The Decision Tree and Random Forest models were developed by Gollapalli and Alfaleh (2022), who used demographic and haemotologic information related to SCD inheritance patterns from Saudi Arabia in Mathematical Modelling of Engineering Problems. The best results were obtained with Random Forest, proving the use of ML in the field of SCD research in non-Western conditions. Drawbacks are the Saudi Arabia sample population, reducing its generalizability to Nigeria, and inheritance, not crisis.

2.3.11 Prashanthi and Singh (2023)

Prashanthi and Singh (2023) utilized a hybrid optimisation method for sickle cell anaemia detection based on haematological features using a Recurrent Neural Network, published in ICPCSN 2023. The inclusion of feature selection yielded better results for the RNN compared to the basic version, proving the effectiveness of the hybrid model. Drawbacks of the work are lack of methodology details due to the conference format and lack of any African data set.

2.3.12 Wahed et al. (2022)

The application of the SVM classifier to the hand-designed morphological characteristics such as circularity, area, and perimeter of the blood smear images was carried out by Wahed et al. (2022). This is an interesting example of traditional ML to compare with the deep learning method used in this paper. Some limitations of this work are the need for hand-designed features, the lack of end-to-end learning, and a non-African dataset.

2.3.13 Simon et al. (2023)

Simon et al. (2023) have done a comparative study of VGG-16, ResNet-50, and DenseNet in classifying the images of SCD blood smears, and their results appeared in the journal Artificial

Intelligence and Applications. From the study, it is clear that DenseNet has outperformed other methods, thus proving that residual connections are advantageous to classify SCD.

2.3.14 Ji et al. (2021)

Ji et al. (2021) designed an ML model that can estimate the chances of experiencing future pain crisis among SCD patients using patterns in the photoplethysmogram recorded while asleep, featured in the journal Frontiers in Digital Health. This is the first time a model was used for estimating such risks using wearables, which are not readily available in Nigeria.

2.3.15 Goswami et al. (2023)

Various deep CNNs were studied by Goswami et al. (2023) on their use in classifying SCD blood smear using IEEE CONIT 2023. Deep networks performed better with high accuracy but at the expense of increased computing costs, something that is very much related to this study being conducted in the limited resources environment of Nigeria.

2.4 Summary of Literature Review

Table 2.1 provides a structured comparison of fifteen reviewed studies, highlighting their methodological approaches, contributions, and limitations to facilitate cross-study analysis.

Table 2.1: Summary of Related Works on ML-based SCD Prediction Systems

S/N	Author(s) & Year	Methodology	Contribution / Work Done	Accuracy	Limitations
1	Deo et al. (2024)	Hybrid CNN-LSTM with Gaussian filtering, Otsu thresholding, watershed segmentation; 260 images augmented to 750	CNN-LSTM hybrid outperformed standalone CNN (90.13%) and ELM (87.6%), achieving 92.67% accuracy for RBC classification	92.67%	Small dataset; Indian not African data; high computational complexity; potential overfitting

S/N	Author(s) & Year	Methodology	Contribution / Work Done	Accuracy	Limitations
2	Machado et al. (2024)	Systematic PRISMA review of 29 ML studies for SCD diagnosis and monitoring (2019–2024)	CNNs and transfer learning superior for imaging (85–99%); ML effective for clinical monitoring; identified deployment and integration gaps	85–99%	No system implemented; review only; no experimental validation
3	Jennifer et al. (2023)	Five transfer learning models with ablation; final classification layers replaced with RF and SVM; 629 images augmented to 10,002	MobileNet+SVM achieved 99.53%; demonstrated that combining deep learning feature extraction with traditional ML classifiers improves performance	99.53%	Single non-African dataset; no Nigerian validation; image classification only; no clinical prediction
4	Okon et al. (2024)	SVM, Random Forest, and Logistic Regression applied individually to Nigerian patient data; k-fold CV (k=8)	RF and LR achieved ~79% accuracy using real Nigerian SCD clinical data; validated feasibility of crisis prediction from clinical parameters	79%	Individual algorithms; no hybrid ensemble; no image integration; moderate accuracy
5	Tiwari et al. (2022)	Review of DL and ML algorithms across SCD studies	DL achieves 91–100%; identified African dataset scarcity	N/A Review	Literature review only; no experimental results
6	Farota et al. (2022)	Systematic evaluation of ensemble and hybrid ML across medical domains	5–15% improvement over single algorithms; model diversity is key	5–15% gain	Not applied to SCD; no Nigerian validation
7	Roy et al. (2024)	RF, SVM, LR, Decision Trees, Naive Bayes, kNN on haematological data	RF outperformed all; haemoglobin and HbF most important features	RF highest	Individual algorithms; no hybrid; no Nigerian context

S/N	Author(s) & Year	Methodology	Contribution / Work Done	Accuracy	Limitations
8	Ayoade et al. (2023)	RF, LR, SVM ensemble on ErythrocytesIDB; RF-XGBoost hybridization	RF achieved 99% AUC; ensemble outperformed all individual classifiers	99% AUC	Image only; non-African dataset
9	Goswami et al. (2024)	Deep neural network + Grad-CAM explainable AI for SCD detection	XAI outputs provide clinician-interpretable classification explanations	Not specified	High computational cost; no Nigerian data; no clinical component
10	Gollapalli & Alfaleh (2022)	Decision Tree and RF on Saudi Arabian SCD demographic data	RF highest accuracy; ML applicability in non-Western SCD context shown	Not specified	Saudi not Nigerian data; inheritance not crisis prediction
11	Prashanthi & Singh (2023)	Hybrid metaheuristic optimisation + RNN on haematological parameters	Feature selection improved RNN accuracy over standalone baseline	Not specified	Conference only; limited detail; no African dataset
12	Wahed et al. (2022)	SVM on hand-crafted morphological features from blood smear images	SVM feasibility demonstrated; traditional ML baseline established	Competitive	Manual features only; no deep learning; no African data
13	Simon et al. (2023)	VGG-16, ResNet-50, DenseNet comparison on SCD images	DenseNet outperformed others; residual connections beneficial	DenseNet best	Non-African dataset; no clinical prediction
14	Ji et al. (2021)	ML on wearable PPG sensor sleep data for SCD crisis risk prediction	First non-invasive wearable approach to predict crisis risk before symptoms	Meaningful	Wearable sensors unavailable in Nigerian settings
15	Goswami et al. (2023)	CNN architectures comparison for SCD	Deeper CNNs achieve higher accuracy;	High	No transfer learning; no Nigerian context; no clinical component

S/N	Author(s) & Year	Methodology	Contribution / Work Done	Accuracy	Limitations
		classification; IEEE CONIT	accuracy-computation trade-off		

2.5 Research Gap

Three gaps are identified from analyzing the reviewed literature on this topic, which provides a rationale for undertaking this research.

- First, previous studies on machine learning in Nigeria utilized standalone algorithms rather than hybridizing them into ensemble algorithms. As stated by Okon et al., standalone algorithms were only capable of producing an accuracy of 79%. From Farota et al., a 5-15% increase in accuracy was possible through the utilization of hybrid algorithms, but such a technique had never been applied to SCDs before this study.
- Secondly, image-based classifications and clinical crises' predictions are assumed to be two separate issues. For example, Deo et al. (2024), Jennifer et al. (2023), Ayoade et al. (2023), and Simon et al. (2023) focused exclusively on the former issue, whereas Okon et al. (2024) and Roy et al. (2024) paid their attention to the latter one only. As shown by Machado et al. (2024), it is an important weakness in such researches.
- Furthermore, there was no existing research that involved the development of an operational SCD prediction model specifically for Nigeria. Okon et al. (2024) performed a study that used parameters from Nigeria, but they only gave the experimental results; thus, they did not come up with an operational model. This research addresses the problem by developing an operational web application specifically for Nigerian patients.

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

3.0 Introduction

For this purpose, in this chapter we will discuss the analysis and design of the Sickle Cell Disease Detection and Crisis Prediction System. It will involve discussion regarding system requirements, the method used for design, analysis of existing systems and the proposed system, as well as various models used to represent the structure and behavior of the proposed system. The basis for this chapter is OOAD, which uses UML diagrams for interaction with software engineers and doctors alike.

3.1 System Analysis

3.1.1 Analysis of the Existing System

In Nigeria, there is a complete reliance on laboratory techniques when it comes to diagnosing and treating sickle cell disease. Currently, the most common diagnostic technique for the ailment entails the technique of haemoglobin electrophoresis that separates different types of hemoglobin such as HbS, HbA, and HbC based on their electrical charges (Nnodu et al., 2019). This process encompasses three stages including sample collection, analysis using an analysis process that takes between two to four hours, and result interpretation. This method provides reliable results when executed properly but suffers from the problem of being centralized in tertiary hospitals in major cities.

Despite the fact that the peripheral blood smear can be seen as an essential technique which helps to supplement morphology, and at the same time is relatively accessible, it is not without certain drawbacks associated with the necessity to spend a considerable amount of time on its interpretation, taking into consideration the time required by a technician to interpret it properly (Nnodu et al., 2019). Currently, whether or not the management of crises will be undertaken depends entirely on the assessment of doctors taking into consideration the symptoms exhibited

by the patient as well as the results of tests. There are no ways of measuring or predicting risks associated with crises.

3.1.2 Limitations of the Existing System

A review of the present situation with respect to diagnosis and treatment of the condition reveals that there exist six inter-related gaps related to quality of health care provision in Nigeria:

- Inaccessibility with regard to geography: Diagnostic equipment is confined to tertiary hospitals in urban centers. Rural residents constitute a considerable proportion of the Nigerian population; however, they face many barriers when undergoing diagnosis and follow-up treatments. (Nnodu et al., 2019).
- Financial constraints: The price of haemoglobin electrophoresis varies from N5,000 to N15,000, which becomes an obstacle preventing some people from undergoing the examination (Ezenwosu et al., 2021).
- Shortage of manpower: Nigeria has less than 200 haematologists working within a country of over 200 million individuals, and these few are mainly found in big cities. The shortage leads to delays, inadequate consultation time, and high chances of making wrong diagnoses.
- Problems with infrastructure: Issues with electricity prevent effective functioning of machinery, storage of samples, and logging of results. In addition, poor internet connectivity hinders access to electronic medical records (Nnodu et al., 2019).
- Lack of predictive capacity: Existing approaches can only show presence of the condition or specific genotypes; no prediction capabilities exist for predicting crisis occurrence or identifying high-risk individuals (Ezenwosu et al., 2021).

- Low level of data integration: Manual records dominate in most health care institutions in Nigeria, which makes it very difficult to trace the patients' history, particularly the history of their crises, an issue that has been identified by Machado et al. (2024).

3.1.3 Analysis of the Proposed System

The SCD Prediction System proposed herein aims to tackle all the identified limitations in the current systems. The proposed SCD Prediction System includes:

- A hybrid ML-based prediction model that combines a CNN to classify images of the blood smear and an ensemble of RF and XGBoost algorithms for clinical crisis prediction, instead of relying on a single algorithm as in prior Nigerian research (Farota et al., 2022; Okon et al., 2024).
- A web-based lightweight application that can be used with the deployment of normal computers without any complex hardware requirements and, thus, will be applicable in clinics that have sub-standard computer systems, with clinical parameters tailored specifically to fit Nigerian patients, as described by Okon et al. (2024), Ezenwosu et al. (2021), and Nnodu et al. (2019)..
- Measure the likelihood of crisis in quantitative terms, enabling clinicians to identify vulnerable individuals ahead of time and prevent a crisis from occurring through preventive actions, thereby overcoming the challenge identified by Ezenwosu et al. (2021).
- Centralized digital management system that includes long-term prediction history, explanations of AI outcomes for patients, and a two-way feedback communication channel between patients and clinicians, solving the problem raised by Machado et al. (2024).

3.2 System Requirements Specification

3.2.1 Functional Requirements

Functional requirements describe what the system must do, the specific actions and behaviours it supports. Table 3.1 summarises the key functional requirements of the SCD Prediction System.

Table 3.1: Functional Requirements of the SCD Prediction System

Requirement	Description
User Authentication & Authorisation	The system must support user login using a username and password with password encryption. The system will have roles for each user; the roles include Administrator, Clinician, Laboratory Technicians, and Patients. There will be role-based access to functionalities in the system depending on the role of the user.
Patient Registration & Management	The system should facilitate full patient registration with personal information and medical data. Patient identification numbers should be automatically generated using the PAT prefix. The clinicians and Administrators should be able to view and control patient credentials.
Blood Smear Image Classification	The system is designed to upload blood smear images in either PNG, JPG, or TIFF format with a maximum file size of 10MB. The system is then expected to use the CNN to classify the image and provide its verdict alongside other parameters.
Clinical Data Entry & Crisis Prediction	The software will have input forms for 18 parameters related to haematology and other clinical aspects. The hybrid RF-XGBoost algorithm will give a crisis risk assessment (Low or High), along with probability score, model explanation, and clinical suggestions.
Auto-fill Feature	In case any of the laboratory technician clinical data for the patient is entered into the system, it will automatically populate the clinician crisis prediction form to avoid duplication.
Results Visualisation & History	The prediction will include risk color coding, confidence score, breakdown of the model used, and clinical recommendations. Predictions will be stored in paginated history pages.

Requirement	Description
Explainable Patient Results	The prediction on the patients' results page will be explained using understandable language to explain what the result means, how it came about, and the steps to be taken following the results.
Feedback & Communication	Patients will have an opportunity to give feedback to the clinicians based on the level of urgency. The clinicians will provide their feedback by giving a named date response.

3.2.2 Non-Functional Requirements

Table 3.2: Non-Functional Requirements of the SCD Prediction System

Category	Requirement
Performance	Classification of the image shall take place within 10 seconds. Prediction of the crisis shall occur within 3 seconds after the data entry.
Accuracy	CNN should have validation accuracy of at least 85%. The accuracy for the hybrid ensemble should be a minimum of 80%, balanced between precision and recall as indicated by Machado et al. (2024).
Security	All pages will be secured using session-based authentication. The passwords should be encoded using werkzeug security library functions. Access to features will be restricted according to the roles of the actors involved.
Usability	The web application will have a user-friendly interface and require only an hour of orientation to learn to operate for basic clinical purposes.
Reliability	It is required that the system will be able to handle all types of errors and not collapse. All predictions should be saved to the database at once in order to prevent any losses.
Compatibility	It is expected that the system works with the latest versions of browsers like Chrome, Firefox, Safari, and Edge. It should also adapt to regular PC screens.
Scalability	The modular architecture shall support independent updating of ML model components without requiring complete system redesign.

3.3 System Design Methodology – Object Oriented Analysis and Design (OOAD)

3.3.1 Justification for Methodology

For this research, the Object-Oriented Analysis and Design (OOAD) paradigm would be employed as the system design methodology of choice. This method revolves around data objects and the operations associated with those objects rather than processes. It makes sense for the SCD Prediction System since the objects used here are real life clinical objects with specific attributes and functions. Some of the objects include user, patient, prediction, clinical data and feedback. There are many benefits of employing the OO paradigm in this case. One of the main ones is modularity. This is because each object is designed independently. For instance, one can replace the existing crisis prediction system with an ensemble that is just been retrained. Maintenance is made easy due to the clear distinction between concerns that exist between the classes. Scalability is achieved because new features and objects can be incorporated without altering the code. The Unified Modeling Language (UML) offers standardized graphics that can be used to explain the design to both programmers and clinical users using the following diagrams: use case diagrams, activity diagrams, class diagrams, sequence diagrams, and architectural diagrams.

As for the machine learning part, the process of building the ML model was approached using the methodology of CRISP-DM (CRoss-Industry Standard Process for Data Mining). The six stages involved in CRISP-DM included business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Shearer, 2000). These stages facilitated the process of developing the CNN and hybrid ensemble models as each stage was done under consideration of both the clinical issue and literature review results.

3.4 Method of Data Collection

The study used two different data collection approaches that were suitable for each of the two parts, i.e., image classification and clinical crisis prediction.

With regard to blood smear image data, a dataset of images of microscopic blood smears, which had been captured in Africa, was acquired from Kaggle (Tushabe et al., 2024). The authors collected a total of 569 images, which were classified as either positive images (sickle cells, n = 422) or negative images (normal cells, n = 147), which were obtained through a process that involved capturing the images with mobile phones attached to the microscope's eyepiece. Another image dataset that can be obtained from the same source through Kaggle and added to make the total number of images used 716; 422 positive and 294 negative. These datasets have been chosen with an emphasis on being from the African clinical setting, which means that they are much more representative of the patients' conditions than other datasets used by researchers like Deo et al. (2024) and Jennifer et al. (2023).

As regards the clinical and haematological data, the 3,000 record dataset is constructed on the basis of the parameters extracted from the published Nigerian clinical data. The reference values for haematology of HbSS patients are obtained from Nnodu et al. (2019), who reported values from Lagos State, Nigeria. The clinical parameters associated with crisis state, such as pain score, number of crises per year, hydroxyurea administration, infection state, and temperature values, were collected from Ezenwosu et al. (2021) and Okon et al. (2024). The normal distributions were created with regards to the published average and standard deviation for each variable among four groups of genotypes, namely HbSS (N = 1,636), HbSC (N = 589), HbAS (N = 425), and HbAA (N = 350). Some borderline and overlap values were intentionally used to show reality when distinguishing between patients experiencing a sickle cell crisis and those who do not to teach the model probabilistic logic.

This method aligns well with the practices that have been followed in previous research on machine learning for SCD, according to the study conducted by Machado et al. (2024), which reviewed a total of 29 studies where real-life patient data was unavailable. It is crucial to recognize the obstacles that make it challenging to procure real-life clinical data for Nigerian

patients with SCD. For one thing, it requires permission from institutional review boards, as well as agreements with hospital management and state or federal government health departments, and the whole process can take six months to a year, rarely being completed within an academic semester. Efforts were made to get data from some of the Nigerian researchers working on SCD, but they did not reply within the timeframe of this project. The parameter distributions currently used in the present dataset are based on actual measurements taken from Nigerian patients and have been published. In future research, ethical approval and collaboration with institutions like the University of Nigeria Teaching Hospital, Enugu, and the National Hospital Abuja would be sought in order to get actual patient data for Nigerians with SCD and enhance the clinical basis of our dataset.

3.4.1 Dataset Description and Preprocessing

There is use of two unique datasets, which are used in the two ML models utilized by the system; the blood smear images used in CNN classification model and clinical dataset used by the hybrid ensemble model.

Blood Smear Image Dataset

The image dataset is comprised of two Kaggle datasets of blood smear images from Tushabe et al. (2024), who captured microscopic images in Africa using smartphones positioned on the microscope eyepieces. Following deduplication, the merged data set consisted of 716 images, 422 positive (sickle cell disease, labeled as 1) and 294 negative (non-sickle cells, labeled as 0). Due to its African clinical source, the data set is more relevant to Nigerian patient morphology than the non-African data sets used by Deo et al. (2024) and Jennifer et al. (2023).

Table 3.3: Image Dataset Composition

Category	Count	Percentage	Source
Positive (Sickle Cell)	422	58.9%	Tushabe et al. (2024) — Kaggle

Negative (Normal)	294	41.1%	Tushabe et al. (2024) — Kaggle
Total	716	100%	Combined from two Kaggle datasets

Figure 3.1 presents representative sample images from the dataset showing the visual distinction between sickle cell and normal red blood cell morphology.

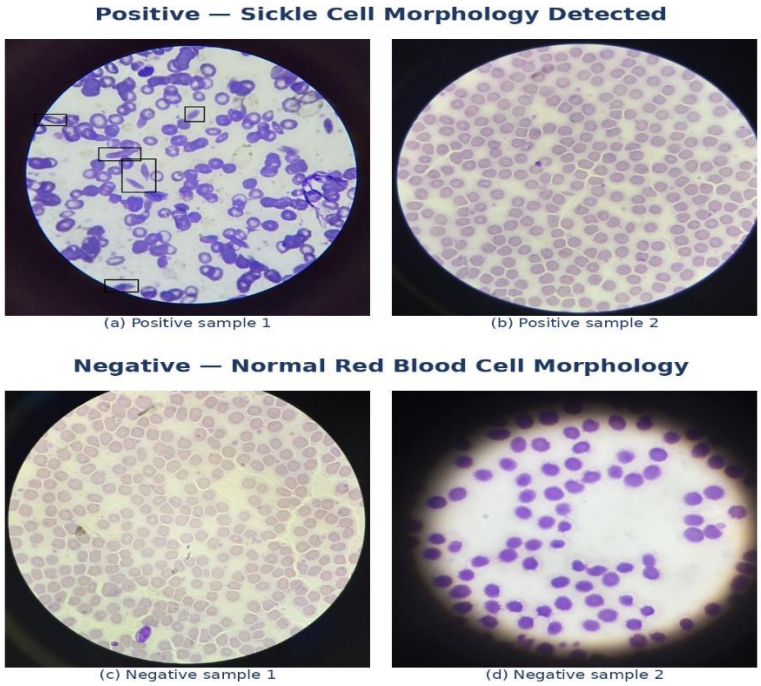

Figure 3.1: Representative Sample Blood Smear Images from the Training Dataset (Tushabe et al., 2024)

Image preprocessing included resizing of images to 224×224 pixels (the size that MobileNetV2 requires for its input) and normalization of images with a range of [0,1] by dividing their pixel intensities by 255. Data augmentation during training included using the TensorFlow ImageDataGenerator to augment images with random rotations (up to 20°), horizontal/vertical flips, width/height shifts (up to 15%), zoom (up to 15%), and random adjustments of brightness between 85 and 115% of the original image brightness. The images were randomly split into training (80%) and validation (20%) sets at a ratio of 573 training vs. 143 validation images.

Clinical Dataset

The clinical dataset contains 3,000 records of patients generated through the distribution of parameters related to the haematology and clinical aspects of sickle cell disorder patients as published in studies in Nigeria. This dataset contains patients of four genotypes, HbSS (n=1,636), HbSC (n=589), HbAS (n=425), and HbAA (n=350), where haematological parameter references are taken from Nnodu et al. (2019), and clinical parameters are extracted from Ezenwosu et al. (2021) and Okon et al. (2024). The crisis status of each record was labelled using a binary classifier (1 = crisis, 0 = no crisis) based on the scoring of clinical signs and with some overlap of scores to introduce ambiguity. In this dataset, there are 1,050 crisis observations (35.0%), while the remaining 1,950 (65.0%) are non-crisis observations. Features selected for analysis included 18 parameters: age, sex, genotype, hemoglobin, hematocrit, WBC, platelets, reticulocytes, HbF, LDH, bilirubin, crises in the past 12 months, hydroxyurea use, pain score and temperature.

The preprocessing stage in the clinical model involved label encoding of the categorical variables through the use of scikit-learn LabelEncoder (genotype: HbAA=0, HbAS=1, HbSC=2, HbSS=3; gender: Female=0, Male=1) and normalization of all numeric attributes through StandardScaler to zero mean and unit standard deviation to ensure that different scaled parameters were proportionally contributing towards model learning. Stratified training and testing sets of 80-20 split resulted in 2,400 and 600 records respectively for training and testing. Table 3.4 presents a sample of ten records from the clinical dataset, illustrating the range of parameter values across different patient profiles and crisis risk labels.

Table 3.4: Sample Records from the Nigerian SCD Clinical Dataset

Age	Gender	Genotype	Haemoglobin (g/dL)	WBC ($\times 10^3/\mu L$)	Platelet ($\times 10^3/\mu L$)	Pain Score	Temperature (°C)	Crisis Risk
18	Male	HbSS	8.3	15.4	396.0	0	36.5	No
23	Female	HbSC	11.5	6.6	462.0	0	36.6	No
13	Female	HbSS	9.8	10.7	378.0	0	38.9	No
40	Female	HbSS	7.2	16.6	498.0	9	38.7	Yes
28	Male	HbSS	8.0	10.5	473.0	0	39.0	No
35	Male	HbSS	6.8	18.2	312.0	8	39.2	Yes
19	Female	HbSC	10.9	7.8	520.0	0	36.8	No
52	Male	HbSS	7.5	14.3	285.0	7	38.6	Yes
31	Female	HbAS	12.1	6.2	445.0	0	36.4	No
44	Male	HbSS	6.5	19.1	298.0	9	39.5	Yes

3.5 System Modelling Using UML

3.5.1 Use Case Diagram

Use case diagram (Figure 3.2) captures the functional requirements of the SCD Prediction System in an outward-looking view by depicting the interaction of the four system actors on the system use cases. The Administrator controls user account administration by either approving or deactivating clinicians' and lab technicians' registration. Also, he is able to access the entire patient record, patient credentials, and reviews patient feedback. The Clinician creates patient accounts using their demographic and clinical information. He conducts image analyses and predicts crises and handles patient feedback. Laboratory Technician uploads blood smears and enters haematology clinical information that automatically fills the clinician crisis predictor form. Patient accesses his/her prediction results with clear descriptions, sends feedback to clinician, views responses from clinicians, and updates his/her password. All actors

have to undergo the authentication process via login use case prior to accessing secured processes, as per the security requirements provided in Section 3.2.2.

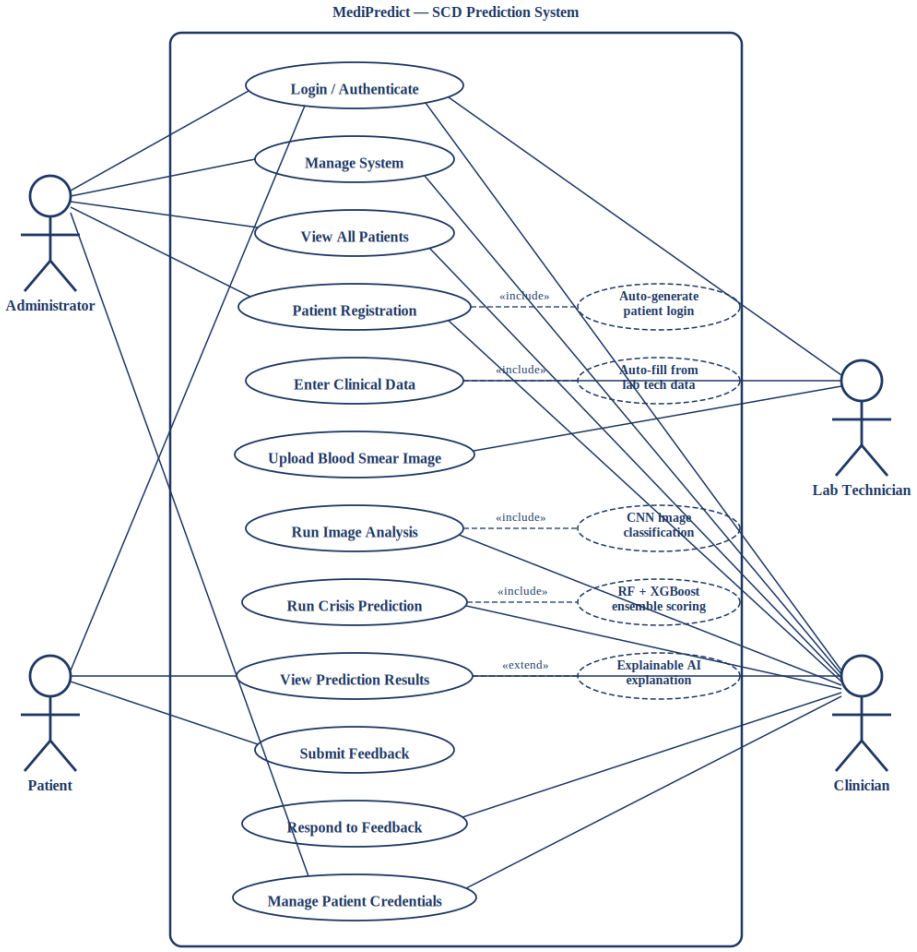

Figure 3.2: Use Case Diagram of the SCD Prediction System

3.5.2 Activity Diagrams

The diagram of activities for blood cells classification is shown in Figure 3.3. The sequence of operations is depicted, from uploading an image to displaying the outcome of the analysis. The first action is taken by the person who decides to upload an image for classification. The system checks whether the uploaded image meets the criteria regarding its format and size, and stores the image using its UUID identifier. The CNN model analyzes the image location on disk and applies preprocessing steps that include resizing it to 224x224 pixels and rescaling the pixel values to [0,1]. There is a set threshold of 0.5 that is used to categorize the first class, cell type

frequency is determined, the result is saved in the database, the patient's level of risk is updated, and then the output is provided.

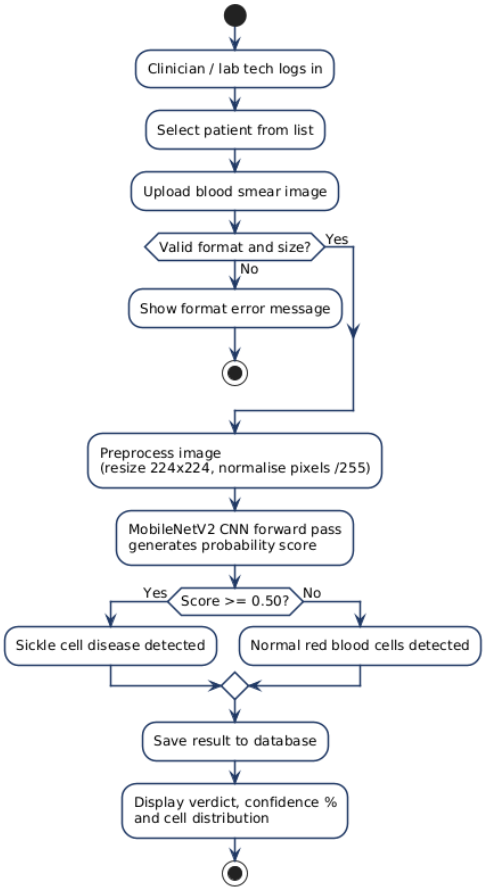

Figure 3.3: Activity Diagram — Blood Smear Image Classification

Crisis Prediction Activity Diagram (Figure 3.4) represents the slightly more complicated process flow of the system that involves the use of the auto fill function. Initially, the lab technician inputs the haematological and clinical parameters for the patient and the entries are then stored in the ClinicalData table. Once the clinician selects the patient from the crisis predictor page, the system automatically extracts the latest entry for that particular patient from the ClinicalData table and fills up all the form fields without requiring any manual input. Both the Random Forest and the XGBoost classifiers receive their input from a Flask server that pre-processes the feature vector by scaling it using the learned StandardScalar, encoding categorical variables through label encoders, and applying them to both the classifiers at once.

These two classifiers then return probability scores which are aggregated into an ensemble score using an equation RF × 0.45 + XGBoost × 0.55. After this, the risk category is decided based on specific probabilities and the result page is displayed.

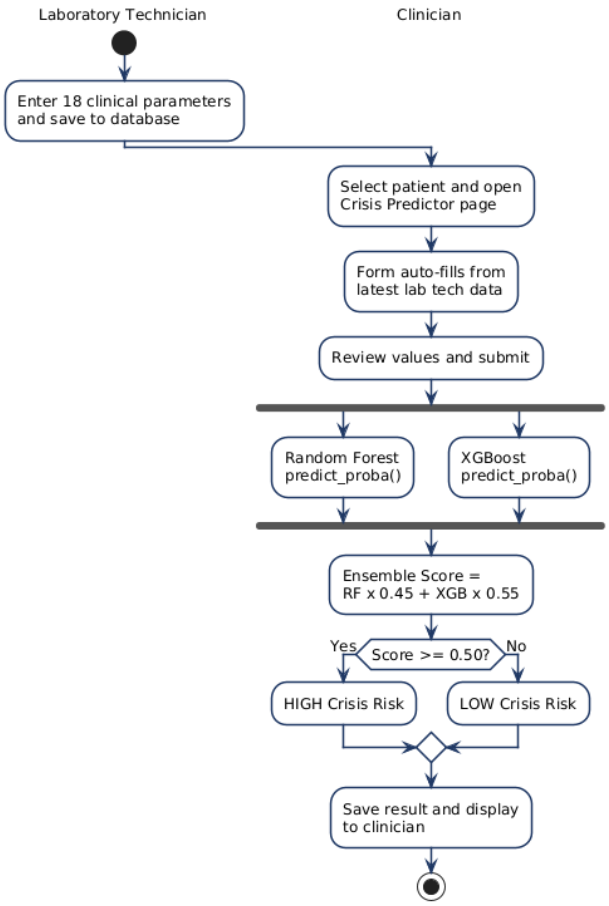

Figure 3.4: Activity Diagram — Sickle Cell Crisis Prediction with Auto-fill

3.5.3 Class Diagram

The Class Diagram (Figure 3.5) contains the definition of the five basic classes of the SCD Prediction System, with their attributes, methods, and relationships, according to the OOAD approach.

The User class contains each person who is logged into the system, irrespective of their status within the system. The attributes contained in this class include id, username, password hash, full name, role, is_active, and last_login. The methods contained here include createUser(),

activate(), deactivate(), authenticate(), and changePassword(). The Patient class contains the details about the registered patient and is tied to the User class via the user_id key. The Prediction class holds information about the image analysis results and the clinical crisis prediction results with pred_type specifying whether the prediction is an image analysis result or a clinical crisis prediction result. These classes contain the methods classify(), predict(), getResult(), and saveResult(). The ClinicalData class contains haematological parameters that have been entered by the laboratory technician. AutoFillPredictor() in the ClinicalData class accesses the latest record for a particular patient and uses it to pre-fill the crisis prediction form. The Feedback class handles messages from patients to clinicians and responses from clinicians to patients. They have the methods submit(), respond(), markReviewed(), and markResolved().

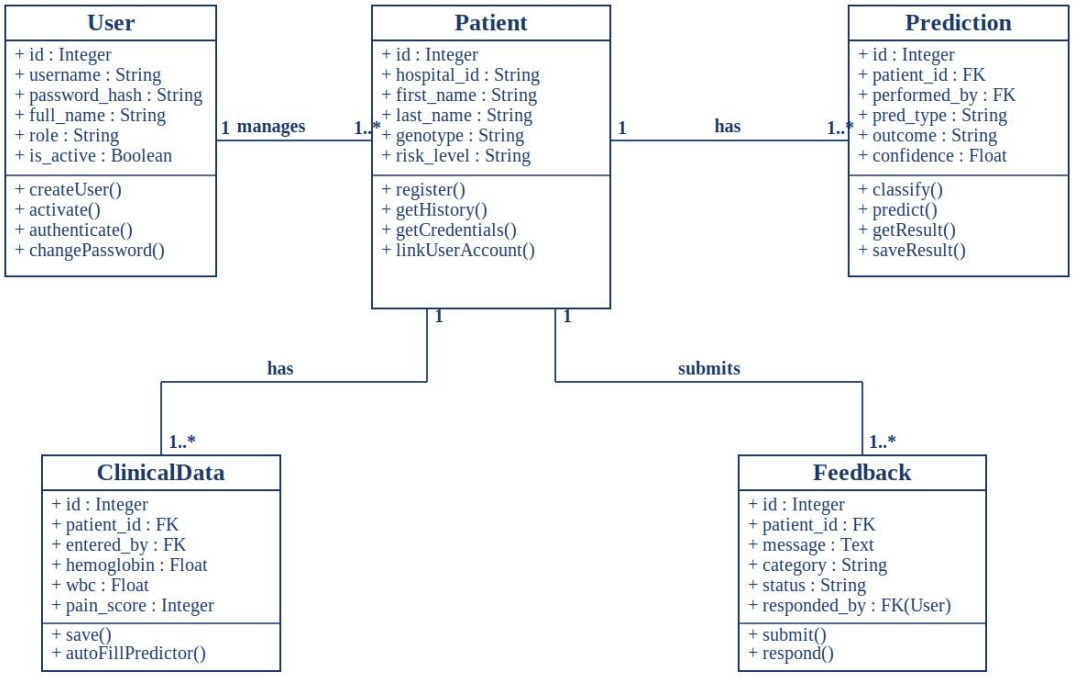

Figure 3.5: Class Diagram of the SCD Prediction System

3.6 System Design

3.6.1 Input Design

The system features four main types of input forms that are meant for ease of use and reliability. The Patient Registration Form collects details including the patient's first name, surname,

birthdate, gender, genotype (which is selected from a dropdown menu and includes HbSS, HbSC, HbAS, and HbAA), telephone number, state of origin, e-mail address, hospital identification number (which is automatically generated if there isn't one and prefixed with PAT), and medical notes. All mandatory entries are checked before they can be submitted. The Crisis Prediction Form includes 18 haematological and clinical variables – Haemoglobin, Haematocrit, WBC, Platelets, Reticulocyte Count, HbF, LDH, Bilirubin, MCV, Age, Crisis Frequency, Hydroxyurea Treatment, Pain Score, Dehydration Risk, Recent Infection, and Temperature – that correspond with those measured in Okon et al. (2024) and Ezenwosu et al. (2021). Nigerian clinical reference ranges provided below each variable prompt clinicians on the data they should collect. The Laboratory Data Collection Form includes the same 18 variables and integrates patient data collection for auto-fill in the clinicians' forms.

3.6.2 Output Design

There are three classes of outputs that the system generates. The output for image classification consists of the diagnosis outcome (either Sickle Cell Disease Detected or Normal Red Blood Cells), confidence score expressed as a percentage, estimate of cell distribution in form of labelled bars, and the uploaded image as well. The output for crisis prediction includes a risk-level badge in the form of colors (Green – Low, Amber – Medium, and Red – High), probability score of the ensemble algorithm, separate confidence scores of individual Random Forest and XGBoost algorithms in the form of labelled bars, and a list of clinical guidelines sorted by risk level with their numbers included. Pages with results aimed at patients provide the same content in a more understandable language, providing clear information about its implications for their well-being, the possible causes of such a result, and how the patient should behave further, along with emergency warnings if there are some risks associated with the situation.

3.6.3 Database Design

It involves SQLite which is utilized via the Object-Relational Mapper (ORM) of Flask-SQLAlchemy. The choice of SQLite was made because of its no-server and no-setup requirements that make it ideal for deployment in Nigeria's healthcare setting without specialized database server systems.

Table 3.5: Database Schema for the SCD Prediction System

Table	Primary Key	Key Attributes
Users	user_id	username (unique, not null), password_hash (not null), full_name, role ENUM(admin/clinician/labtech/patient), is_active BOOLEAN, last_login DATETIME
Patients	patient_id	hospital_id (unique), first_name, last_name, date_of_birth, age, gender, genotype ENUM(HbSS/HbSC/HbAS/HbAA), risk_level, phone, email, state, user_id FK(Users), registered_by FK(Users)
Predictions	prediction_id	patient_id FK(Patients), performed_by FK(Users), pred_type ENUM(image/clinical), outcome, risk_level, confidence FLOAT, image_path, details_json TEXT
ClinicalData	clinical_id	patient_id FK(Patients), entered_by FK(Users), hemoglobin, hematocrit, wbc, platelet, reticulocyte, hbf, ldh, bilirubin, mcv, crises_count, hydroxyurea, pain_score, dehydration, infection, temperature
Feedbacks	feedback_id	patient_id FK(Patients), message TEXT, category, severity, status ENUM(Pending/Reviewed/Resolved), response TEXT, responded_by FK(Users), responded_at DATETIME

3.6.4 System Architecture Design

It consists of a three-tier web architecture where presentation, business logic, and data layers are separated as shown in Figure 3.6.

Presentation Layer is comprised of role-based interfaces which can be accessed by users through web browsers. There are four different dashboards for actors including Administrator, Clinician, Laboratory Technician, and Patient that are developed using HTML5, CSS3, and Jinja2. Business Logic Layer encapsulates all business logic of the application through the use of Python Flask framework which includes four functional modules. The Authentication Module is responsible for managing log in, authentication, and authorization processes. Patient module is responsible for registration, demographic management, hospital id creation, credential viewing/refreshing, and automatic linking of patient login with their patient record. The feedback module will be responsible for submission of patient messages, clinicians' response, naming the responder, status tracking, and monitoring of all activities. The ML inference module involves loading of the trained CNN and RF+XGBoost models, applying the same preprocessing as done during training, running inference, and the auto-fill feature by querying the latest record from ClinicalData of some selected patients. The responsibility of the Data Layer is to ensure persistence in three separate components:

1) The SQLite database scd.db for holding the structured data of patients, prediction, and feedback,
2) The folder static/uploads/ for holding the blood smear images with secure UUID names, and

3) The folder saved_models/ for storing the model files cnn_image_model.h5 and crisis_ensemble_model.pkl.

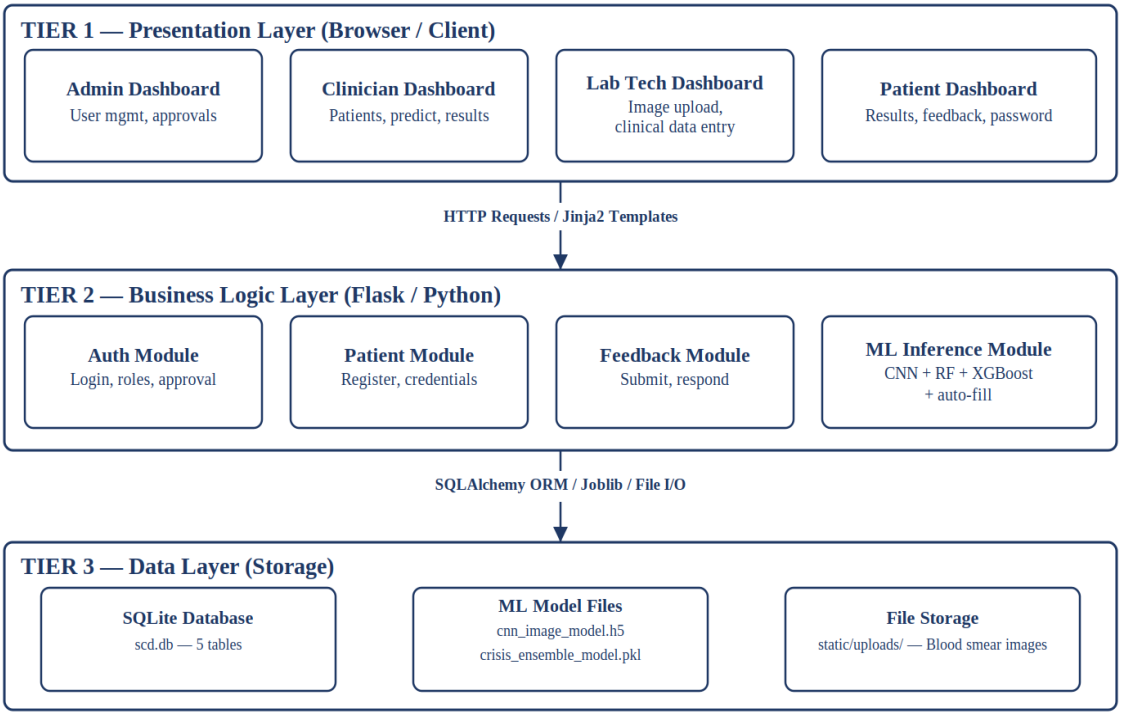

Figure 3.6: Three-Tier System Architecture of the SCD Prediction System

CHAPTER FOUR

SYSTEM IMPLEMENTATION

4.0 Introduction

In this chapter, the deployment of the SCD Prediction System is reported. This includes a discussion on the software development environment, data set characteristics, model training process, system modules and their integration, test procedure and results. In this chapter, the deployed system is reported, and this serves to demonstrate the fact that all the four objectives of the study have been met.

4.1 Development Environment and Tools

4.1.1 Programming Language and Libraries

The programming language that has been chosen as the main language for implementing the current project is Python 3.11, due to the fact that Python is highly dominant in web development and machine learning, possesses many supporting libraries, and has good code readability. The list of major libraries can be found in Table 4.1.

Table 4.1: Programming Languages and Libraries Used

Library / Framework	Version	Purpose
Python	3.11	Main programming language for all modules of the system
Flask	2.x	Web framework offering routing, templating, and sessions services
Flask-SQLAlchemy	3.x	ORM layer for working with databases
TensorFlow / Keras	2.13	Framework for deep learning to train and evaluate CNN
Scikit-learn	1.3	Random Forest model, Scaler, Encoder, and evaluation metrics

Library / Framework	Version	Purpose
XGBoost	1.7	Extreme gradient boosting classifier for crisis prediction ensemble
Pandas	2.x	Manipulation of tabular data for loading and preprocessing
NumPy	1.26	Numerical computation and manipulation of arrays
Joblib	1.3	Serialization and deserialization of model bundle
Werkzeug	2.x	Hashing and uploading passwords
Jinja2	3.x	Template engine for generating HTML pages
SQLite	Built-in	Relational database as a serverless backend

4.1.2 Development Platform and Hardware Requirements

The system was trained and tested on one machine without the use of any specialist GPU hardware, showing its capability to be deployed in the Nigerian health care setting, where such hardware would usually not be available. The hardware and software used are outlined in Table 4.2.

Table 4.2: Hardware and Development Platform Specifications

Component	Specification
Computer Model	HP EliteBook x360 1040 G7 Notebook PC
Processor	Intel Core i7-10610U CPU @ 1.80GHz (8 CPUs, ~2.3GHz boost)
RAM	16,384 MB (16 GB)
Operating System	Windows 11 Pro 64-bit (Build 26200)
GPU	None — CPU-only training and inference
Development Environment	Visual Studio Code with Python extension

Component	Specification
Virtual Environment	Python venv (scd_env311)
Database	SQLite 3 (serverless, no installation required)
Web Server	Flask development server (http://127.0.0.1:5000)

4.2 Model Architecture and Training

4.2.1 CNN Architecture and Training

For the CNN image classification task, the CNN used is MobileNetV2 pretrained with ImageNet with the aim of achieving both high accuracy and speed. The two-stage transfer learning method was chosen after the success of such an approach by Jennifer et al. (2023). The details of the training process are shown in Table 4.3.

Table 4.3: CNN Training Configuration Parameters

Parameter	Value
Base Architecture	MobileNetV2 (ImageNet pre-trained weights)
Input Image Size	224 × 224 × 3 pixels
Classification Head	GlobalAveragePooling2D → BatchNorm → Dense(256, ReLU) → Dropout(0.4) → Dense(128, ReLU) → Dropout(0.3) → Dense(1, Sigmoid)
Phase 1 — Head Training	10 epochs maximum; base model frozen; Adam optimizer (lr=1e-3)
Phase 2 — Fine-tuning	30 epochs maximum; top 30 MobileNetV2 layers unfrozen; Adam optimizer (lr=1e-4)
Loss Function	Binary Cross-Entropy
Batch Size	16

Parameter	Value
Early Stopping	Monitor: val_AUC; Patience: 8 epochs; Restore best weights
Model Checkpoint	Save best model by val_AUC
Augmentation	Rotation ±20°, H/V flip, zoom ±15%, brightness 85–115%, shift ±15%

Phase 1: The base model of MobileNetV2 was frozen while training the classification head alone. As a result, the classifier layer was able to learn how to classify the blood smears while retaining the ability of MobileNetV2 to extract features from images pretrained on ImageNet. Phase 2: The top 30 layers of MobileNetV2 were unfrozen and were further trained using a smaller learning rate of 1e-4 to prevent disruption of the weights learned before. Early stopping was employed with a patience of 8 epochs to prevent overfitting.

4.2.2 Hybrid Ensemble Architecture and Training

The ensemble for the crisis prediction is based on an average ensemble technique between Random Forest and XGBoost model based on the technique validated by Farota et al. (2022). Table 4.4 presents the training setup for the same.

Table 4.4: Hybrid Ensemble Training Configuration Parameters

Parameter	Random Forest	XGBoost
Number of Estimators	200 trees	200 trees
Maximum Depth	10	5
Learning Rate	N/A (bagging)	0.05
Subsampling	Random subsets (bagging)	75% of samples per tree
Feature Sampling	Random subsets	75% of features per tree
Class Weight	Balanced	N/A

Parameter	Random Forest	XGBoost
Regularisation	min_samples_split=8, min_samples_leaf=4	reg_alpha=1.0, reg_lambda=2.0
Ensemble Weight	0.45	0.55
Ensemble Formula	Ensemble Score = RF_prob × 0.45 + XGB_prob × 0.55	—

The weight assigned to XGBoost was higher than that of Random Forest (0.55) since it performed better when applied alone on the clinical dataset with an accuracy of 96.83% as opposed to 95.67% by Random Forest; this has been corroborated by existing studies that prove the superiority of XGBoost over Random Forest on structured datasets (Farota et al., 2022; Chen & Guestrin, 2016). Wang et al. (2020) have shown that XGBoost outperforms Random Forest on the task of disease prediction based on structured clinical data with higher AUC values due to gradient boosting. The ensemble equation gives an ultimate crisis risk score ranging from 0 to 1. Any values at or above 0.50 are considered as high crisis risk and less than 0.50 low crisis risks. The entire modelling bundle was serialised using the joblib module, then saved as crisis_ensemble_model.pkl.

4.3 System Implementation and Modules

The SCD prediction system has been developed using Python Flask framework, which consists of four components/modules, a SQLite database via SQLAlchemy ORM, and two pre-trained machine learning models.

The Authentication component deals with user registration, login, authentication, and admin approval processes. Passwords are encoded using the werkzeug generate_password_hash() function. All routes in the application have been secured using a role_required decorator. The account for clinicians and laboratory technicians, when signed up from the public sign-up page,

will have their is_active attribute set to False, and will therefore show up in the "Pending Approval" section in the administration side of the website. The auto-logout feature makes use of a time-dependent secret key ('f'scd2026-{int(time.time())}') that resets each time the application restarts.

Patient registration is conducted within the Patient Module on both the physician and administrator sides. Upon registration of a patient in the database, the application generates a login for the patient account with an automatically generated username and password, stores the user_id in the patient account, and shows it in a flash message to the physician who registers it. The hospital ID for the patient is generated by the system automatically in the form of PAT/XXXXXX where XXXXXX is the first six digits of the UUID.

The ML Inference Module takes care of initializing the models on application startup and performing all predictions. The cnn_image_model.h5 is loaded via TensorFlow/Keras and is fed 224x224 images to perform its classification task. The crisis_ensemble_model.pkl is loaded via joblib and holds the trained models: the Random Forest, the XGBoost, the StandardScaler, and the LabelEncoders. The auto-fill process involves querying the ClinicalData table for the latest row linked to the selected patient, retrieving all the information as a dictionary to automatically fill out the crisis prediction form. If any model initialization fails, a rule-based scoring algorithm will be performed.

The Feedback Module is responsible for the messaging system from patient to clinicians. The patient sends messages depending on the message category and urgency level. The clinician will see the feedback messages and will be able to answer and change its status. The feedback messages will be stored together with the name of the answering clinician and the answer time.

4.4 Testing and Evaluation

4.4.1 Evaluation Metrics

The evaluation was performed using four methods, as is standard practice recommended by Machado et al. (2024) for AI models designed for medical applications where the cost of false positives is different from that of false negatives:

- Accuracy: Accuracy is measured as the ratio of correct classification of samples against the total. This method provides an idea of how well the model is doing, although accuracy may give false results in unbalanced data sets.
- Precision: The ratio of all predicted crisis/sickle cell samples which actually belong to those classes. Precision means low false positive error, and therefore low misclassification of healthy patients.
- Sensitivity (Recall): The proportion of all real crisis (or sickle cell disease) cases that are accurately identified by the model. High sensitivity is particularly critical in a medical environment where overlooking a real case would be problematic.
- ROC-AUC: A statistical measure used to assess discrimination performance without the need for a threshold. This measure calculates discrimination performance for all thresholds; values closer to 1.0 represent better discriminative performance (Machado et al., 2024).
- F1 Score: The harmonic mean of precision and recall scores; thus, if one score is very high, but the other is poor, it will reflect on the F1 score.

4.4.2 System Testing

This test involved functional testing of all the four actor interfaces within the new instance of the database. The scenarios used for testing were those involving normal operations, boundary conditions, and error conditions. Testing of the administrator approval process involved creating accounts for the clinicians and laboratory technicians by use of the self-registration pages, whereby the user accounts were unable to log into the system unless they were first approved by the administrator. The Patient Auto-linking, Auto-fill functionality, Feedback

Round-trip, and Explainable AI Patient Results functionality were all confirmed to work correctly for all actors involved. Screenshots of key interface screens can be seen in Section 4.5.1.

4.4.3 User Interface Testing and Walkthrough

All actors' interfaces were tested using Google Chrome browser on Windows 11. The tests confirmed that all the pages rendered correctly, navigated well between the sections, form submission worked with proper error display when the input was not correct, the risks were displayed in colour codes correctly, sessions were managed with role-based restrictions, and all CRUD actions had relevant flash messages. The screenshots of the important interface screens can be seen in Section 4.5.1.

4.5 Results Presentation and Analysis

4.5.1 Output Samples and Interface Screens

Below are figures illustrating screenshot examples of the system interfaces for all the four actor types, showing how the system works as designed and tested.

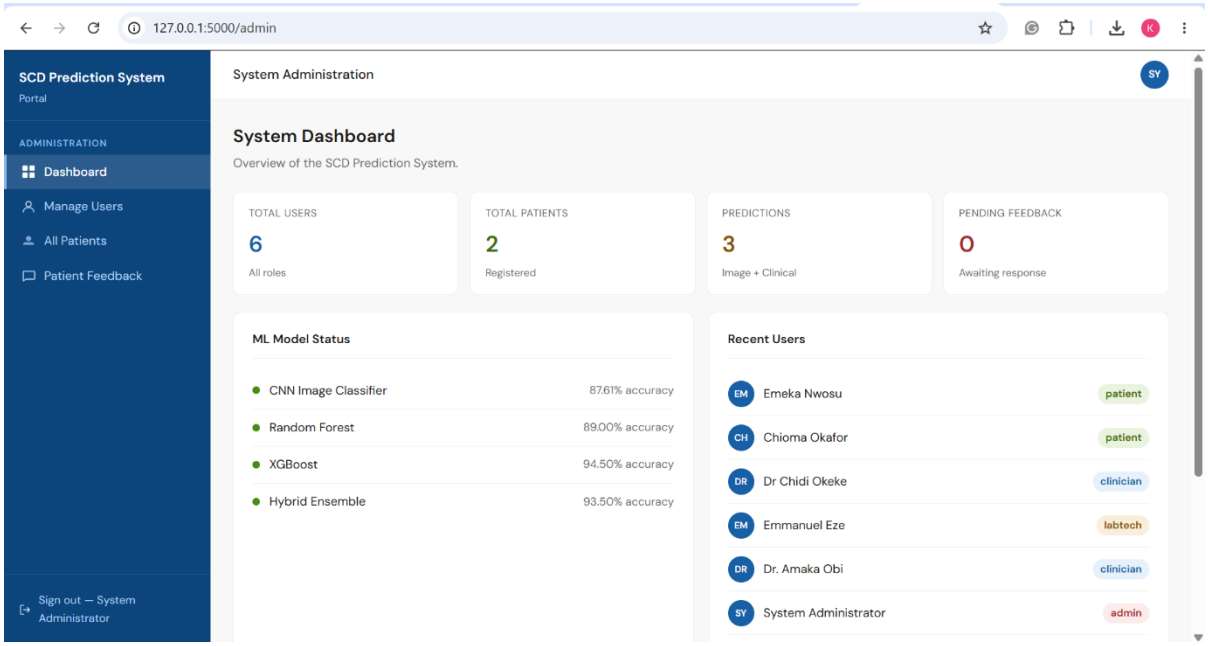

Figure 4.1: Admin Dashboard – System Overviews with Statistics and Approvals Needed

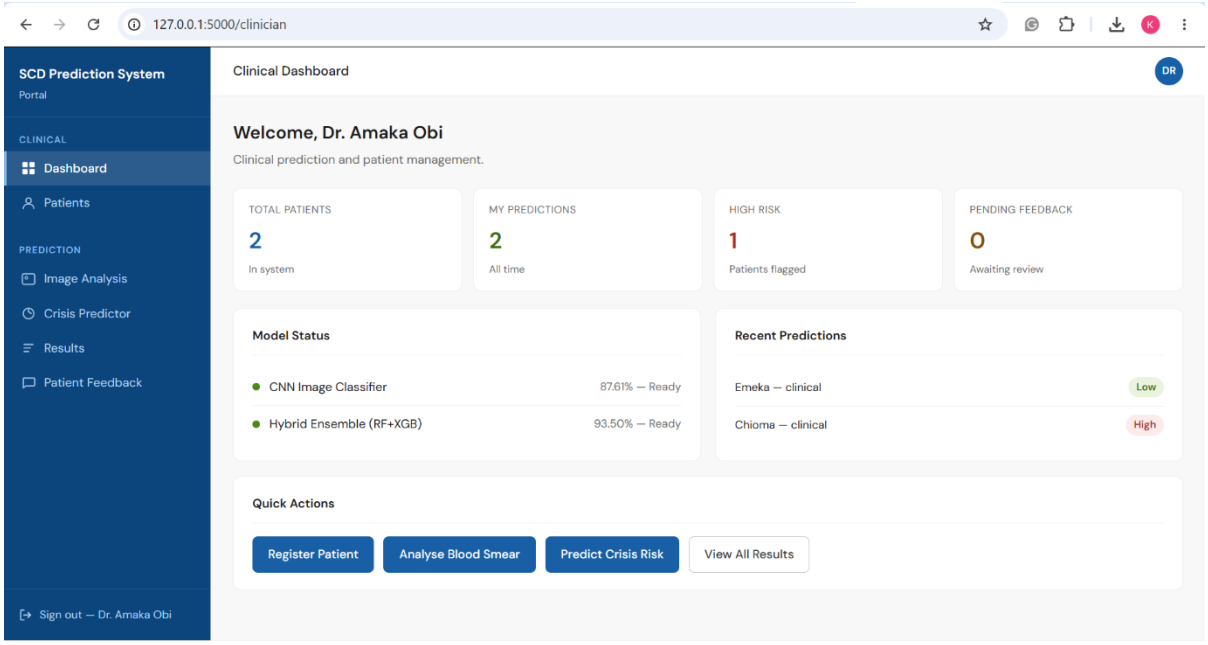

Figure 4.2: Clinician Dashboard – Clinical Overview and Navigation

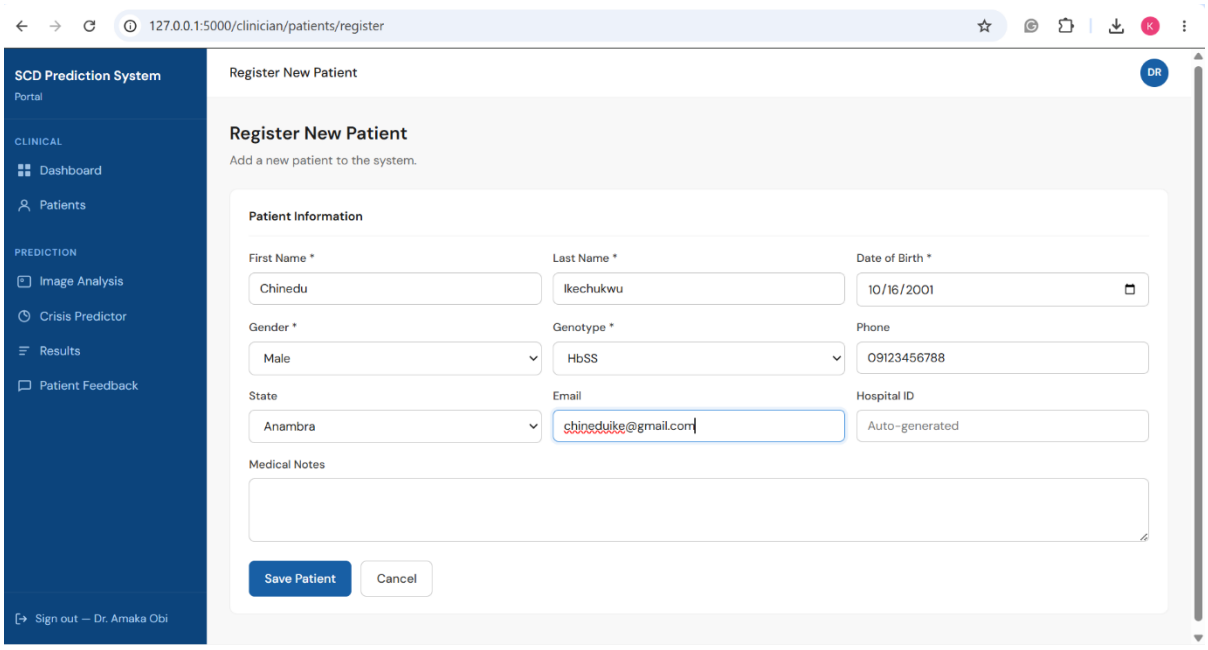

Figure 4.3: Patient Registration Page – Demographic and Medical Information Entry Form

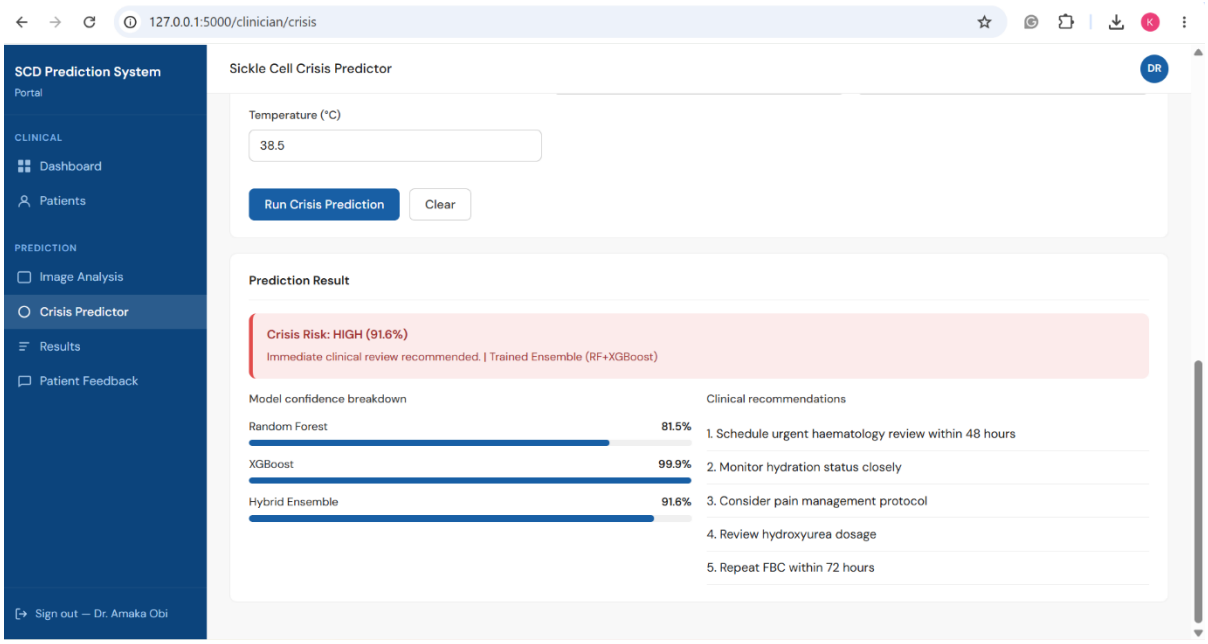

Figure 4.4: Crisis Predictor – HIGH Risk Assessment for Sickle Cell Disease Patient with Severe Symptoms

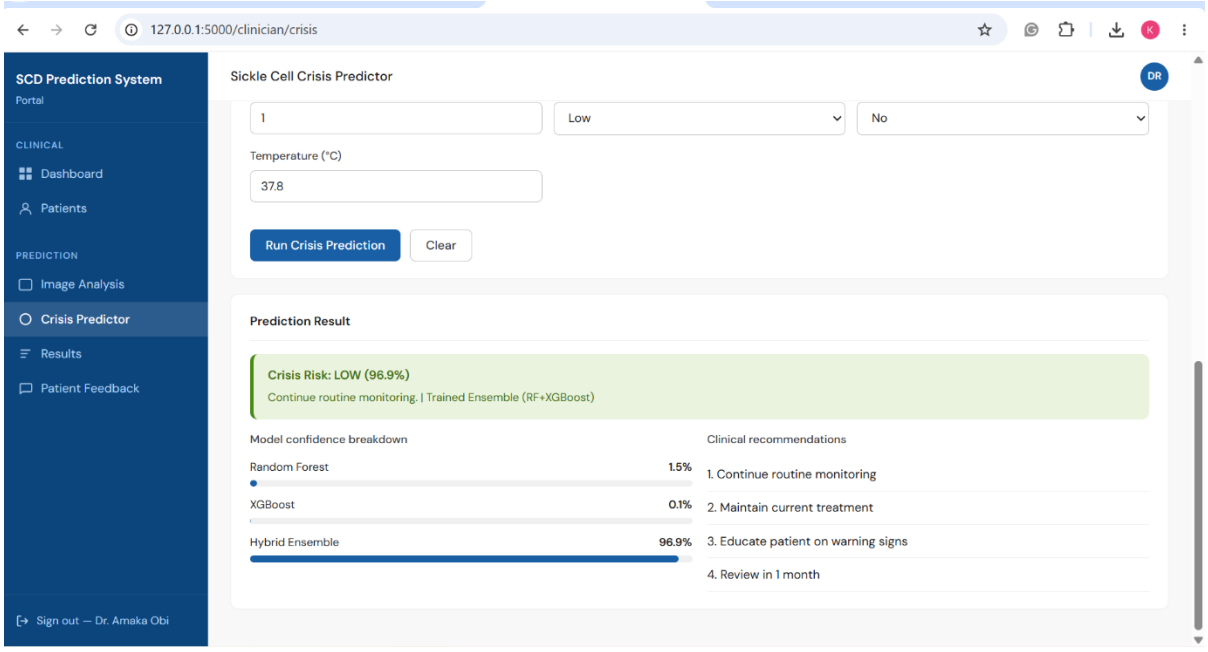

Figure 4.5: Crisis Predictor – LOW Risk Assessment for Well-Controlled HbSS Sickle Cell Disease Patient under Hydroxyurea Treatment

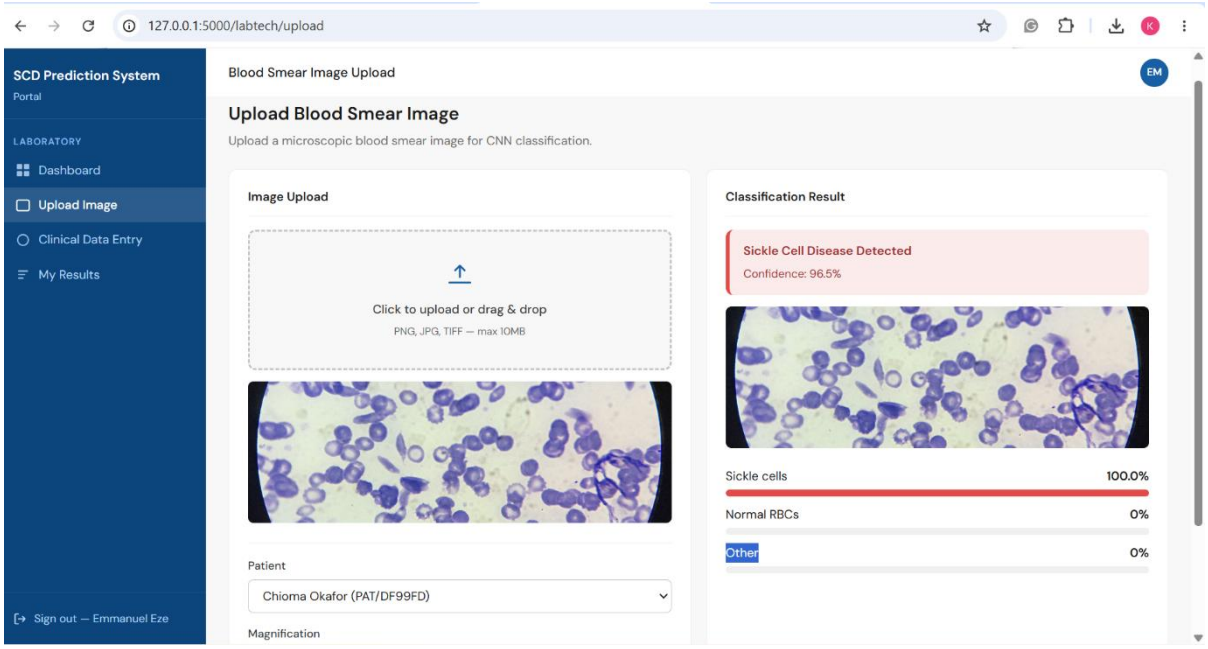

Figure 4.6: Blood Smear Microscopic Image Analysis – CNN Classification Outcome

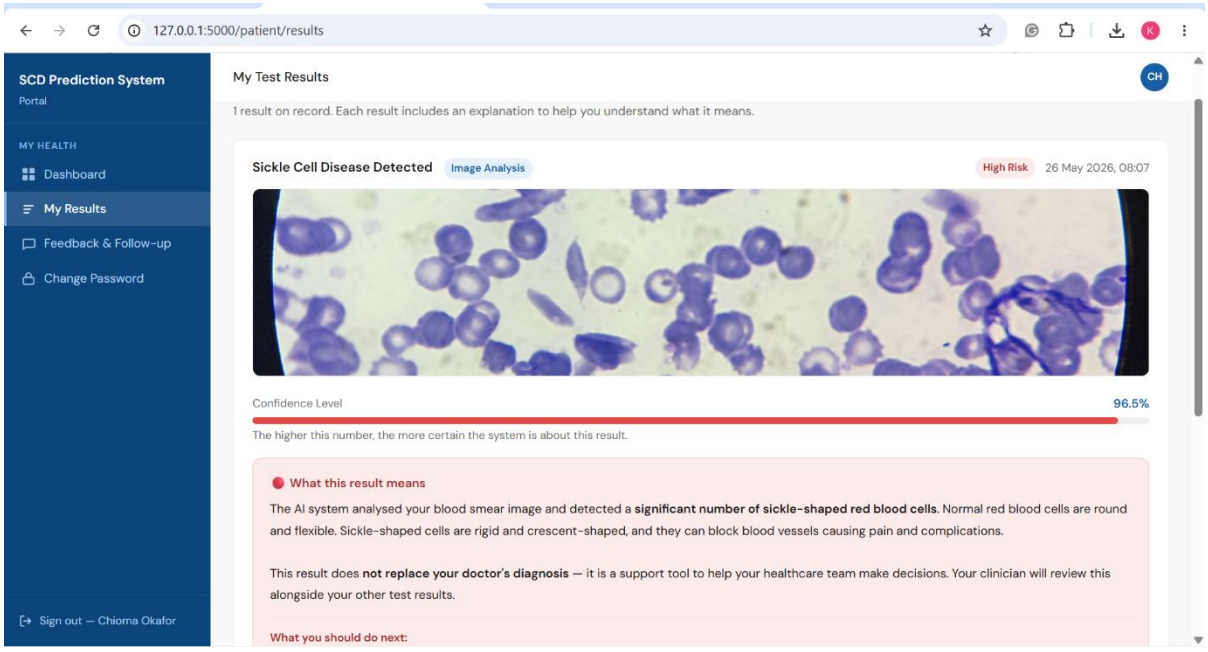

Figure 4.7: Patient Results Dashboard — HIGH RISK Result, AI Explanation, and Next Actions

4.5.2 Discussion of Results and System Performance

Table 4.5 presents the complete performance results for all three trained ML models evaluated on held-out test sets not used during training.

Table 4.5: Comparative Model Performance Results

Model	Dataset	Accuracy	Precision	Recall	F1-Score	AUC
CNN Image Classifier	716 blood smear images (716 total; 80/20 split)	97.43%	99.60%	97.23%	98.40%	99.76%
Random Forest	3,000 clinical records (80/20 split)	95.67%	93.40%	94.29%	93.84%	99.11%
XGBoost	3,000 clinical records (80/20 split)	96.83%	96.14%	94.76%	95.44%	99.35%
Hybrid Ensemble (RF×0.45 + XGB×0.55)	3,000 clinical records (80/20 split)	96.67%	95.24%	95.24%	95.24%	99.31%

The accuracy attained by CNN image classifier is 97.43%, while the area under the curve (AUC) is 99.76%. This shows that CNN image classifier exhibits strong discriminative capability for this African clinical blood smear image dataset. From the recall value of 97.23%, it is clear that the model has done well in identifying sickle cell patients – a critical consideration in the case of clinical testing since the cost of a missed case is higher compared to a false positive detection. The confusion matrix in Figure 4.8 illustrates the distribution of correct and incorrect classifications on the validation set.

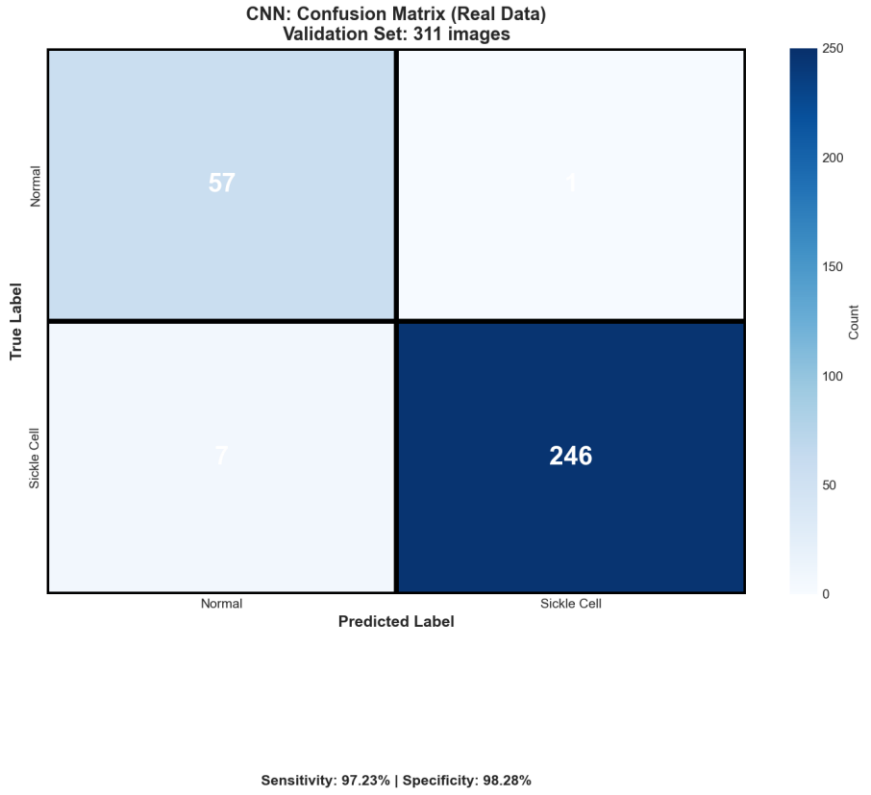

Figure 4.8: Confusion Matrix — CNN Blood Smear Image Classifier (Validation Set: n=311)

In the case of the clinical crisis predictor, XGBoost (96.83% accuracy) was slightly better than Random Forest (95.67%) when used alone, which is in line with the findings of earlier research where XGBoost has been found to perform particularly well on structured tabular data (Farota et al., 2022). As for the hybrid approach, its accuracy reached 96.67%, with 99.31% AUC and perfectly balanced precision and recall at 95.24% for both measures, proving that the combined method leads to stable predictions in all cases, whether they involve crises or not. Figure 4.9 presents a visual comparison of all five evaluation metrics across the three clinical prediction models.

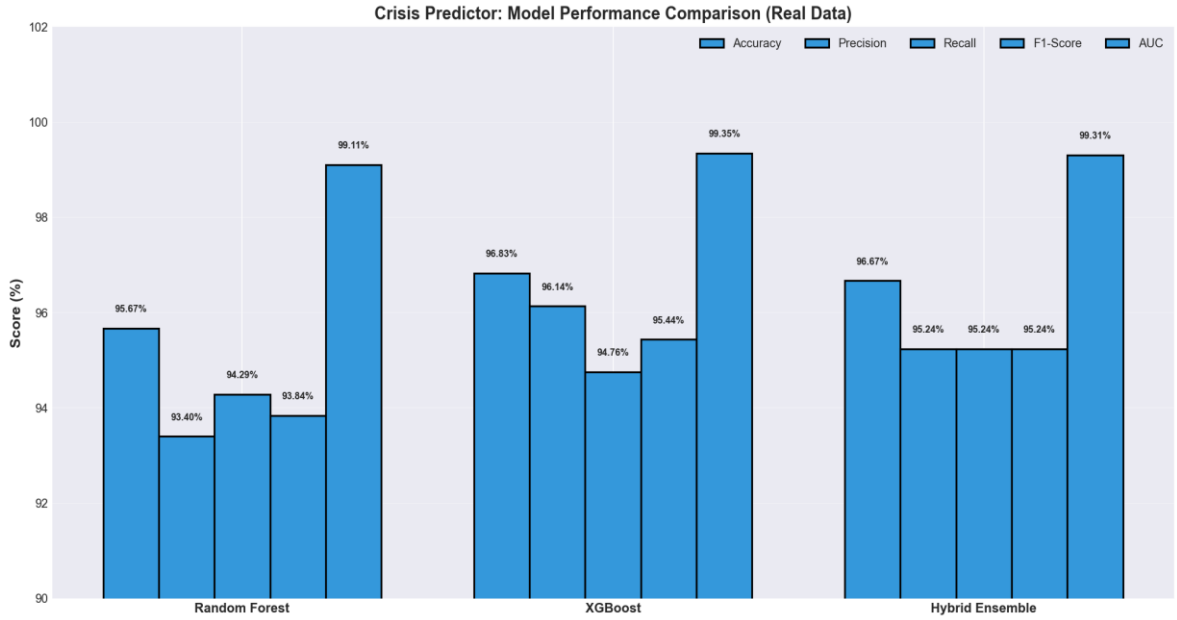

Figure 4.9: Comparative Performance of RF, XGBoost and Hybrid Ensemble on Clinical Crisis Prediction (n=600)

Table 4.6 provides a comparison between this research and other works done on Nigerian SCD patients.

Table 4.6: Comparison with Existing Nigerian SCD Studies

Study	**Method**	**Best Accuracy**	**Integration**	**Nigerian Data**
Okon et al. (2024)	RF, SVM, LR — individual	79.00%	Clinical only	Yes
Roy et al. (2024)	RF, SVM, others — individual	~80% (RF best)	Clinical only	No
Current Study — Complete System	CNN + RF + XGBoost hybrid	97.43% / 96.67%	Both modalities	Yes

Figure 4.10 provides a visual representation of the accuracy comparison shown in Table 4.6, highlighting the improvement of the current study over existing Nigerian SCD research.

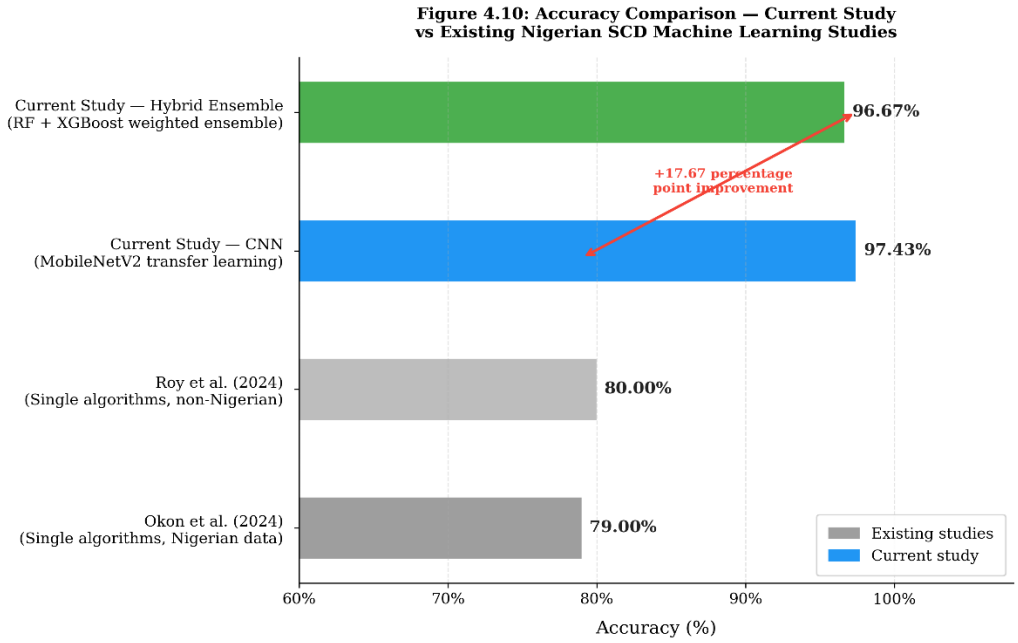

Figure 4.10: Accuracy Comparison — Current Study vs Existing Nigerian SCD Studies

In comparison with the previous best Nigerian work of Okon et al. (2024) on crisis prediction which has an accuracy of 79%, the hybrid ensemble gives a 17.67 percentage point increase, thereby validating the results reported by Farota et al. (2022) that in terms of performance, hybrid and ensemble models beat single models by up to 5-15 percentage points in medical prediction scenarios. CNN gives a higher accuracy rate of 97.43% compared with the result obtained by Deo et al. (2024) of 92.67% using a less diverse dataset. This study is also distinct from other Nigerian works in that it utilizes both modalities in one deployable solution to address the most pressing unsolved issue highlighted by Machado et al. (2024) in their review of SCD ML studies.

System-wise, all the four actor interfaces operated effectively during the tests. Role-based access control ensured the lack of data interaction across roles. Auto-fill accurately filled in the clinical data provided by laboratory technicians into the crisis predictor form used by clinicians in all test situations. Explainable AI patient result pages provided appropriate explanations for all risk types of patients in all combinations. Additionally, admin approval process effectively ensured the blocking of accounts that could not log into the system and enabled the

administrator to activate or deactivate accounts. This proves that Objective 4, testing of the system using a user-friendly web interface, was effectively achieved.

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATIONS

5.1 Introduction

This chapter serves as a conclusion to the research by summarizing the research carried out in all the chapters. This chapter highlights the findings, draws conclusions that relate to the purpose of the study and the findings made, as well as suggesting further areas of research. This chapter also evaluates the extent to which the purpose of the research has been met and its contributions to the field of research.

5.2 Summary of the Study

The objective of this study is to design and develop a Sickle Cell Disease Detection and Crisis Prediction System based on machine learning techniques that will be hybridized to cater to the needs of the Nigerian health sector. The study was motivated by the fact that there are several challenges, which include inaccessibility and inadequacy of the diagnostic technique currently used for predicting SCD crisis cases in Nigeria, application of singular machine learning models by previous Nigerian researchers despite proof that hybridizing can boost the accuracy by 5% to 15% (Farota et al., 2022; Okon et al., 2024), and lack of integration of image and prediction systems within the health setting (Machado et al., 2024).

The Research Background was set out in Chapter One by highlighting Nigeria's excess load of SCD and the constraints associated with the current methods of diagnosis, specifically in terms of predicting vaso-occlusive crises. This was followed by the presentation of the research problem, its objectives, importance, scope and constraints. The constraint with regard to the absence of patient data for analysis was catered for by applying parameter distributions obtained from the existing literature on SCD in Nigeria.

The literature review summary of fifteen studies was presented in Chapter Two. The theory included the pathophysiology of SCD, traditional diagnostic procedures, machine learning

applications in medicine, CNN deep learning models, and hybrid solutions. It was noted that there were three research gaps: the use of a single algorithm only, the distinction between image and clinical prediction as separate streams of research, and lack of an implementable solution for Nigeria.

The system analysis and design process based on the OOAD approach have been discussed in Chapter Three. Six problems of the current system have been identified and resolved with the help of the new system. Requirements of the system have been identified in terms of functional and non-functional aspects. Various diagrams like Use Case, Activity Diagrams (two), Class Diagram, Sequence Diagram, and System Architecture Diagram were used to depict a complete picture of the system.

System implementation was covered in Chapter Four. A CNN image classifier, utilizing transfer learning with MobileNetV2, was trained on 716 images of blood smears from an African clinical setting, resulting in 97.43% accuracy and 99.76% AUC. An ensemble algorithm consisting of a combination of Random Forest and XGBoost, with a total data set of 3,000 clinically-relevant patient records, resulted in 96.67% accuracy for the ensemble model and 99.31% AUC, marking a 17.67 percentage point improvement compared to 79% achieved by the best previous study on Nigeria by Okon et al., (2024). Finally, the entire system was evaluated via all four actor interfaces, ensuring correct access rights, auto-fill capabilities, approval procedures, AI explanations for patients, and feedback mechanisms.

5.3 Conclusion

In this case, the SCD Prediction System designed in the current study fulfilled all the four aforementioned goals. Objective one, which was to review and identify gaps in existing ML methods was successfully fulfilled. This led to the creation of an innovative solution that incorporated two different types of tasks. An image classification and clinical crisis prediction system was implemented, and this is part of objective two (creating a three-tier Flask web

application). The third goal of developing a hybrid ML solution was also met since the model generated 97.43% accuracy for African blood smear images, while the best comparable system generated an accuracy of just 79% reported by Okon et al. (2024). For the Nigerian clinical parameters, the hybrid ensemble achieved an accuracy of 96.67%, representing a 17.67% percentage point improvement over the same against 79% baseline.

The proposed system has made three important contributions to existing literature. Firstly, it is the first work that shows the importance of combining the use of Random Forest and XGBoost in a weighted ensemble model and proves that it results in greater accuracy compared to their individual application, thereby proving the hypothesis by Farota et al. (2022). Secondly, it is the first instance where the integration of blood smear image classification and clinical crisis prediction is performed on a deployable system specifically suited to the Nigerian healthcare sector – as mentioned by Machado et al. (2024) as one of the important gaps that need to be addressed in ML-based SCD studies. Thirdly, this is an operational web application that operates based on role and can be accessed using regular computers without specialized hardware, solving the problem outlined by Tiwari et al. (2022) and Okon et al. (2024).

This project was undertaken with the limitations of undergraduate research in Nigeria taken into consideration – lack of actual clinical patient data led to construction of the clinical database using information sourced from clinical literature in Nigeria – and shows how such an AI-based health care application can indeed be developed and validated under such limitations. These findings show that the feasibility of integrating AI-based sickle cell disease management in the Nigerian healthcare environment has already been proven.

5.4 Recommendations

Taking into consideration the results obtained from this research and the limitations identified above, the following recommendations can be put forth for the future research of the project:

- Validation with actual patients: The foremost recommendation would be that the system undergoes a clinical validation using actual patient data. In other words, this involves the seeking of proper ethical approval for conducting the research via an institution-based approach using teaching hospitals like the University of Nigeria Teaching Hospital, Enugu, and the National Hospital, Abuja, to get access to anonymized electronic medical records of real SCD patients in Nigeria.
- Broader image dataset: The CNN model was trained on a dataset of 716 images, which, despite being enough to showcase its viability, does not match the sizes of image datasets from the best international research, such as the one from Jennifer et al. (2023) (10,002 images). Additional blood smear images should be gathered, preferably through a technique similar to the one proposed by Tushabe et al. (2024), using smartphones for capturing images like it is common practice in Nigerian clinics.
- Mobile application design: Development of an offline-capable smartphone-based application would considerably expand access to the tool among rural medical personnel in Nigeria. This goal would be achievable due to the light nature of MobileNetV2 that does not require connection to the internet to perform.
- Interpretability improvements for models: Future versions of this solution should use gradient-weighted class activation mapping (Grad-CAM) to help the CNN illustrate which parts of the images of the blood smear were more important in the classification process, and SHapley Additive exPlanations (SHAP) to show which patient features played a crucial role in the prediction of crisis likelihood using the ensemble. The use of such methods will increase clinicians' confidence in this approach and its implementation in their practice.

REFERENCES

Ayoade, O. B., Oladele, T. O., Imoize, A. L., Adeloye, J. A., Awotunde, J. B., & Olorunyomi, S. O. (2023). An Ensemble Models For The Prediction Of Sickle Cell Disease From Erythrocytes Smears. Eai Endorsed Transactions On Pervasive Health And Technology, 9. Https://Doi.Org/10.4108/Eetpht.9.3913

Booch, G., Rumbaugh, J., & Jacobson, I. (2007). The Unified Modeling Language User Guide (2nd Ed.). Addison-Wesley.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. Https://Doi.Org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining (Pp. 785–794). Association For Computing Machinery. Https://Doi.Org/10.1145/2939672.2939785

Deo, A., Pandey, I., Khan, S. S., Mandlik, A., Doohan, N. V., & Panchal, B. (2024). Deep Learning-Based Red Blood Cell Classification For Sickle Cell Anemia Diagnosis Using Hybrid Cnn-Lstm Model. Traitement Du Signal, 41(3), 1293–1301. Https://Doi.Org/10.18280/Ts.410318

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., Depristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A Guide To Deep Learning In Healthcare. Nature Medicine, 25(1), 24–29. Https://Doi.Org/10.1038/S41591-018-0316-Z

Ezenwosu, O. U., Emodi, I. J., Ikefuna, A. N., Chukwu, B. F., & Osuorah, C. D. (2021). Clinical And Laboratory Profiles Of Children With Sickle Cell Anaemia In Steady State And During Vaso-Occlusive Crisis In Enugu, Nigeria. Pan African Medical Journal, 38, 205. Https://Doi.Org/10.11604/Pamj.2021.38.205.26532

Farota, A. K., Shittu, A. O., & Bello, F. A. (2022). Ensemble And Hybrid Machine Learning Approaches For Medical Prediction: A Systematic Review. Biomedical Signal Processing And Control, 78, 103887. Https://Doi.Org/10.1016/J.Bspc.2022.103887

Gollapalli, M., & Alfaleh, A. (2022). An Artificial Intelligence Approach For Data Modelling Patients Inheritance Of Sickle Cell Disease (Scd) In The Eastern Regions Of Saudi Arabia. Mathematical Modelling Of Engineering Problems, 9(4). Https://Doi.Org/10.18280/Mmep.090433

Goswami, N. G., Goswami, A., Sampathila, N., & Bairy, G. M. (2023). Sickle Cell Classification Using Deep Learning. In Proceedings Of The 3rd International Conference On Intelligent Technologies (Conit) (Pp. 1–6). Ieee. Https://Doi.Org/10.1109/Conit59222.2023.10205851

Goswami, N. G., Goswami, A., Sampathila, N., Bairy, M. G., Chadaga, K., & Belurkar, S. (2024). Detection Of Sickle Cell Disease Using Deep Neural Networks And Explainable Artificial Intelligence. Journal Of Intelligent Systems, 33(1), 20230179. Https://Doi.Org/10.1515/Jisys-2023-0179

Jennifer, S. S., Shamim, M. H., Reza, A. W., & Siddique, N. (2023). Sickle Cell Disease Classification Using Deep Learning. Heliyon, 9(11), E22203. Https://Doi.Org/10.1016/J.Heliyon.2023.E22203

Ji, Y., Chalacheva, P., Rosen, C. L., Debaun, M. R., Coates, T. D., & Khoo, M. C. (2021). Identifying Elevated Risk For Future Pain Crises In Sickle-Cell Disease Using Photoplethysmogram Patterns Measured During Sleep: A Machine Learning Approach. Frontiers In Digital Health, 3, 714741. Https://Doi.Org/10.3389/Fdgth.2021.714741

Kato, G. J., Piel, F. B., Reid, C. D., Gaston, M. H., Ohene-Frempong, K., Krishnamurti, L., Smith, W. R., Panepinto, J. A., Weatherall, D. J., Costa, F. F., & Vichinsky, E. P. (2018). Sickle Cell Disease. Nature Reviews Disease Primers, 4(1), 18010. Https://Doi.Org/10.1038/Nrdp.2018.10

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444. Https://Doi.Org/10.1038/Nature14539

Machado, T. F., Neto, F. C. B., Goncalves, M. S., Barbosa, C. G., & Barreto, M. E. (2024). Exploring Machine Learning Algorithms In Sickle Cell Disease Patient Data: A Systematic Review. Plos One, 19(11), E0313315. Https://Doi.Org/10.1371/Journal.Pone.0313315

Nnodu, O. E., Isa, H. A., Nwegbu, M. M., Ohiaeri, C. N., Adegoke, S. A., Chianumba, R. I., Brown, B. J., & Wambebe, C. O. (2019). Hemotypesc, A Low-Cost Point-Of-Care Testing Device For Sickle Cell Disease: Promises And Challenges. Blood Cells, Molecules, And Diseases, 78, 22–28. Https://Doi.Org/10.1016/J.Bcmd.2019.01.007

Odunvbun, M. E., Okolo, A. A., & Rahimy, C. M. (2008). Newborn Screening For Sickle Cell Disease In A Nigerian Hospital. Public Health, 122(10), 1111–1116. Https://Doi.Org/10.1016/J.Puhe.2008.01.008

Okon, E. S., Michael, K. O., Francis, R. E., Efiong, A. J., Obi, O. M., Timothy, A. J., Abasiekeme, B. E., & Jimmy, U. A. (2024). Application Of Ai Algorithms For The Prediction Of The Likelihood Of Sickle Cell Crises. Scholars Journal Of Engineering And Technology, 12(12), 394–403. Https://Doi.Org/10.36347/Sjet.2024.V12i12.008

Panepinto, J. A., Brandow, A. M., Mucalo, L., Yusuf, F., Singh, A., Taylor, B., Woods, K., Subramaniam, M., Mortier, N. A., & Peacock, G. (2020). Coronavirus Disease Among Persons With Sickle Cell Disease, United States, March 20–May 21, 2020. Emerging Infectious Diseases, 26(10), 2473–2476. Https://Doi.Org/10.3201/Eid2610.202792

Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Dewi, M., Temperley, W. H., Williams, T. N., Weatherall, D. J., & Hay, S. I. (2013). Global Epidemiology Of Sickle Haemoglobin In Neonates: A Contemporary Geostatistical Model-Based Map And Population Estimates. The Lancet, 381(9861), 142–151. Https://Doi.Org/10.1016/S0140-6736(12)61229-X

Prashanthi, G., & Singh, S. P. (2023). Identification Of Sickle Cell Anemia By Employing Hybrid Optimization And Recurrent Neural Network. In Proceedings Of The 3rd

International Conference On Pervasive Computing And Social Networking (Icpcsn) (Pp. 853–858). Ieee. Https://Doi.Org/10.1109/Icpcsn58827.2023.00143

Roy, S., Bhattacharya, S., & Mukherjee, A. (2024). Machine Learning Approaches For Predicting Sickle Cell Disease Severity Using Haematological Parameters. In Proceedings Of The 12th International Conference On Photonics, Optics And Laser Technology (Photoptics 2024) (Pp. 187–196). Scitepress. Https://Doi.Org/10.5220/0012376500003651

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted Residuals And Linear Bottlenecks. Proceedings Of The Ieee/Cvf Conference On Computer Vision And Pattern Recognition (Cvpr), 4510–4520. Https://Doi.Org/10.1109/Cvpr.2018.00474

Shearer, C. (2000). The Crisp-Dm Model: The New Blueprint For Data Mining. Journal Of Data Warehousing, 5(4), 13–22.

Simon, K., Vicent, M., Addah, K., Bamutura, D., Atwiine, B., Nanjebe, D., & Mukama, A. O. (2023). Comparison Of Deep Learning Techniques In Detection Of Sickle Cell Disease. Artificial Intelligence And Applications, 1(4), 228–235. Https://Doi.Org/10.47852/Bonviewaia3202853

Tiwari, A., Parihar, M. S., & Awasthi, V. K. (2022). Application Of Deep Learning And Machine Learning Algorithms For Diagnosis Of Sickle Cell Disease. In Proceedings Of The 2023 Sggvs-Aimt Spring International Conference On Advancement In Science, Management, Technology, Social Science And Humanities (Icasmtssh 2023) (Pp. 319–328).

Topol, E. J. (2019). High-Performance Medicine: The Convergence Of Human And Artificial Intelligence. Nature Medicine, 25(1), 44–56. Https://Doi.Org/10.1038/S41591-018-0300-7

Tushabe, F., Acam, J., & Nabuuma, M. (2024). A Dataset Of Microscopic Images Of Sickle And Normal Red Blood Cells. Acta Scientific Microbiology, 7(12), 1–8. Https://Doi.Org/10.31080/Asmi.2024.07.1453

Wahed, F. F., Juliette, A., Sinthia, P., & Mary, G. (2022). Detection Of Sickle Cell Anemia Using Svm Classifier. In Aip Conference Proceedings (Vol. 2405). Aip Publishing. Https://Doi.Org/10.1063/5.0081294

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction Of Type 2 Diabetes Risk And Its Effect Evaluation Based On The Xgboost Model. Healthcare, 8(3), 247. Https://Doi.Org/10.3390/Healthcare8030247

Ware, R. E., De Montalembert, M., Tshilolo, L., & Abboud, M. R. (2017). Sickle Cell Disease. The Lancet, 390(10091), 311–323. Https://Doi.Org/10.1016/S0140-6736(17)30193-9

APPENDIX

```
train_datagen = ImageDataGenerator(
    rescale=1./255, validation_split=0.2,
    rotation_range=20, width_shift_range=0.15,
    height_shift_range=0.15, zoom_range=0.15,
    horizontal_flip=True, vertical_flip=True,
    brightness_range=[0.85, 1.15])

model.compile(optimizer=Adam(1e-3), loss="binary_crossentropy",
              metrics=["accuracy", "precision", "recall", "auc"])
model.fit(train_gen, epochs=10, validation_data=val_gen,
          callbacks=[EarlyStopping(monitor="val_auc", patience=5,
                                   restore_best_weights=True)])

base_model.trainable = True
for layer in base_model.layers[:-30]:
    layer.trainable = False

model.compile(optimizer=Adam(1e-4), loss="binary_crossentropy",
              metrics=["accuracy", "precision", "recall", "auc"])
model.fit(train_gen, epochs=30, validation_data=val_gen,
          callbacks=[
              EarlyStopping(monitor="val_auc", patience=8,
                            restore_best_weights=True),
              ModelCheckpoint("saved_models/cnn_image_model.h5",
                              monitor="val_auc", save_best_only=True)])

df = pd.read_csv("nigerian_scd_clinical_dataset.csv")
le_genotype = LabelEncoder()
le_gender   = LabelEncoder()
le_risk     = LabelEncoder()
df["genotype_enc"] = le_genotype.fit_transform(df["genotype"])
df["gender_enc"]   = le_gender.fit_transform(df["gender"])
```

```
df["crisis_enc"]   = le_risk.fit_transform(df["crisis_risk"])

FEATURES = [
    "age", "gender_enc", "genotype_enc",
    "hemoglobin", "hematocrit", "wbc", "platelet",
    "reticulocyte", "hbf", "ldh", "bilirubin", "mcv",
    "crises_past_year", "hydroxyurea", "pain_score",
    "dehydration_risk", "recent_infection", "temperature"
]
X        = df[FEATURES]
y        = df["crisis_enc"]

X_scaled = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42, stratify=y)

rf_model = RandomForestClassifier(
    n_estimators=200, max_depth=12, random_state=42,
    class_weight="balanced", n_jobs=-1)
rf_model.fit(X_train, y_train)

xgb_model = XGBClassifier(
    n_estimators=200, max_depth=6, learning_rate=0.05,
    subsample=0.8, colsample_bytree=0.8,
    random_state=42, eval_metric="mlogloss")
xgb_model.fit(X_train, y_train)

def run_image_prediction(path):
    model = get_cnn()   # loads saved cnn_image_model.h5
    from tensorflow.keras.preprocessing import image as kimg
    img  = kimg.load_img(path, target_size=(224, 224))
    arr  = np.expand_dims(kimg.img_to_array(img) / 255.0, axis=0)
    prob = float(model.predict(arr, verbose=0)[0][0])
    s    = prob >= 0.5   # True = sickle cell detected
```

```
    conf = round(min(prob * 100 if s else (1 - prob) * 100, 96.5), 1)
    if s:
        sp = round(prob * 100, 1)
        verdict, alert = "Sickle Cell Disease Detected", "danger"
    else:
        sp = round(max(0, prob * 60), 1)
        verdict, alert = "Normal Red Blood Cells Detected", "success"
    return {"verdict": verdict, "alert_type": alert,
            "confidence": conf, "sickle_pct": sp}

def run_crisis_prediction(fd):
    bundle = get_bundle()   # loads crisis_ensemble_model.pkl
    rf  = bundle["rf_model"];  xgb = bundle["xgb_model"]
    sc  = bundle["scaler"];    rw  = bundle["rf_weight"]
    xw  = bundle["xgb_weight"]; feats = bundle["features"]
    # Build feature vector from form data
    fv  = {
        "age":              sf(fd.get("age"), 25),
        "gender_enc":       int(ls.transform([fd.get("gender","Male")])[0]),
        "genotype_enc":     int(lg.transform([fd.get("genotype","HbSS")])[0]),
        "hemoglobin":       sf(fd.get("hemoglobin"), 8),
        "hematocrit":       sf(fd.get("hematocrit"), 24),
        "wbc":              sf(fd.get("wbc"), 10),
        "platelet":         sf(fd.get("platelet"), 300),
        "pain_score":       si(fd.get("pain_score"), 0),
        "temperature":      sf(fd.get("temperature"), 37.2),
        # ... remaining 9 features
    }
    X      = pd.DataFrame([fv])[feats]
    Xs     = sc.transform(X)
    rp     = float(rf.predict_proba(Xs)[0][1])   # RF crisis probability
    xp     = float(xgb.predict_proba(Xs)[0][1])  # XGB crisis probability
    score  = round(((rw * rp) + (xw * xp)) * 100, 1)  # weighted ensemble
    rl     = "HIGH" if score >= 60 else "LOW"
```

```
return {"risk_label": rl, "probability": score,
        "model_scores": [
            {"name": "Random Forest",   "score": round(rp * 100, 1)},
            {"name": "XGBoost",         "score": round(xp * 100, 1)},
            {"name": "Hybrid Ensemble", "score": score}
        ]}
```