ACADEMY JOURNAL
OF
SCIENCE AND ENGINEERING
Available online @
www.academyjsekad.edu.ng
AJSE 19 (5) 2025
Original Research

EXPLORATORY EVALUATION OF MACHINE LEARNING ALGORITHMS IN SICKLE CELL GENOTYPE DETECTION FROM HAEMORHEOLOGICAL PARAMETERS DATASET OBTAINED IN NIGERIA

Rahman Abiodun OLALEKAN[1], Ilesanmi Paul IGE[2], Oludare Alani AGBEYANGI[3] and Daniel Paditeiye REUBEN[4]

[1] Department of Biomedical Engineering, Federal University of Technology, Akure, Nigeria.
[2]Department of Medical Laboratory Science, Federal University of Technology, Akure, Nigeria.
[3]Department of Public Health, Federal University of Technology, Akure, Nigeria.
[4]Department of Medical Laboratory Science, Achievers University, Owo, Ondo state, Nigeria.

Corresponding Authors E-mail: raolalekan@futa.edu.ng
Corresponding Authors Mobile No: +2349065608759

Abstract

Sickle cell disease (SCD) patients have characteristic abnormal haemoglobins that cause red blood cells to become sickle-like in shape, leading to various complications. Early detection is desirous, yet existing diagnostic methods require high cost and deep learning curves. This study evaluated the potential of three machine learning (ML) algorithms—Random Forest, Support Vector Machine (SVM), and a Neural Network—in detecting sickle cell genotypes (SS, AS, AA) from a Nigerian dataset of 54 participants using haemorheological parameters. We employed a stratified 5-fold cross-validation methodology to ensure reliable performance evaluation. The Random Forest and SVM models achieved the highest mean accuracy at **90.9% ± 5.8%**. Feature importance analysis confirmed Packed Cell Volume (PCV) as the most discriminative parameter, followed by Plasma Viscosity (PV) and Age. While all models demonstrated high sensitivity in identifying sickle cell anaemia (SS), they consistently failed to correctly classify the sickle cell trait (AS), a critical limitation highlighted by the validation. Our findings suggest that ML leveraging routine lab parameters is a promising screening tool for sickle cell disease, but is not yet viable for comprehensive genotype classification due to challenges with small dataset size and class imbalance. Future works need to focus on acquiring larger, more balanced datasets to improve the detection of the AS trait.

Keywords: Machine learning, sickle cell disease, haemorheological parameters, genotype prediction, Random Forest, Support Vector Machine, Neural Network

1.0 INTRODUCTION

Sickle cell disease (SCD) is among known blood disorders characterized by the presence of abnormal hemoglobin S (HbSS), which causes red blood cells to “sickle” in shape for a life span of about 20 days. HbSS is the most common variant of SCD genotypes and affects millions of people worldwide, with the highest prevalence in sub-Saharan Africa, the Mediterranean region, the Middle East, and parts of India. Clinical manifestations of SCD are diverse and potentially severe, including vaso-occlusive crises, acute chest syndrome, stroke, and organ damage, resulting in significant death rates and reduced life expectancy (Elsabagh et al., 2023).

The genetic basis of SCD is rooted in the point mutations in the beta-globin gene (HBB), with homozygosity for the HbS allele (HbSS) resulting in sickle cell anaemia, the most common and severe form of SCD. Heterozygosity for HbS (HbAS), known as sickle cell trait, is generally considered a benign carrier state, although it can be associated with certain health risks under extreme conditions. Other genotypes, such as HbSC and HbS-beta thalassemia, represent compound heterozygous states with varying clinical severity (Arishi et al., 2021).

Early detection of SCD is crucial for implementing preventive measures and appropriate management strategies to reduce complications and improve quality of life. Traditional diagnostic approaches for SCD include complete blood cell count, hemoglobin electrophoresis, high-performance liquid chromatography (HPLC), isoelectric focusing, solubilty sickling test, and molecular genetic testing (Arishi et al., 2021; Elsabagh et al., 2023). While these techniques offer accurate genotype determination, they often require specialized equipment, trained personnel, and substantial resources, limiting their accessibility in resource-constrained settings where the burden of SCD is highest (Alapan et al., 2016; Arishi et al., 2021; Elsabagh et al., 2023).

Recent technological developments have raised interest in leveraging machine learning techniques to enhance the detection and management of SCD. Machine learning, a subset of artificial intelligence, involves the design of algorithms that can learn patterns from data and make predictions or decisions without a need for explicit instructions from human experts. These initiatives have shown promise in various medical applications, including disease diagnosis, prognosis prediction, and treatment optimization (Machado et al., 2024).

Application of ML techniques to SCD detection enriches healthcare system in various ways. ML algorithms can identify inherent complex, non-linear relationships between laboratory parameters and disease states which might not be apparent through traditional approaches. Likewise, they are capable of integrating multiple parameters to improve diagnostic accuracy, potentially reducing the need for specialized tests. And lastly, ML models developed though domain data can be deployed on portable devices or integrated into existing healthcare systems, enhancing accessibility and scalability of screening programs (Goswami et al., 2024).

Haemorheological parameters, which describe the flow properties of blood and its components, are particularly related to SCD pathophysiology. The sickling of red blood cells in SCD foster alterations in blood viscosity, cell deformability, and other rheological properties which can be measured through various clinical/laboratory techniques. These easily-measured parameters might serve as valuable inputs for ML algorithms aimed at detecting SCD or predicting Red Blood Cell (RBC) motions as a marker of complications related to SCD (Darrin et al., 2023).

Challenges faced in the application of ML in SCD detection include the need for large, diverse, and well-annotated datasets for algorithm training and validation; selection of appropriate features or parameters that provide meaningful discriminative power; interpretability of ML models to facilitate clinical adoption and trust; and the generalizability of algorithms across different populations and healthcare settings (Zhu et al., 2023).

Despite the growing body of literature on ML for SCD, a significant gap remains in the application of these models to the specific context of genotype classification using readily available haemorheological parameters from a high-prevalence region like Nigeria. Previous studies have primarily focused on image-based classification or prediction of clinical outcomes, often relying on complex data modalities that are not routinely accessible in resource-limited settings (de Haan et al, 2020; Alzubaidi et al, 2020). Nigerian population harbours the high burden of SCD (Adigwe et al, 2023). Therefore, the actual research question this study seeks to address is: To what extent can a model trained on simple, routinely measured haemorheological parameters from a Nigerian cohort accurately classify the three major sickle cell genotypes (AA, AS, SS), and what are the specific limitations of this approach, particularly concerning the clinically challenging Sickle Cell Trait (AS)? This work provides a crucial justification for a cost-effective, non-invasive screening tool that can be integrated into existing primary healthcare infrastructure in endemic regions, thereby offering a practical and scalable solution to enhance early detection and management of SCD (Ekong et al, 2023; Long et al, 2024)

The research approach some of the stated challenges by investigating the efficacy of different ML algorithms in determining haemoglobin genotype from well-annotated hematological and haemorheological parameters to inform sickle cell detection. Specifically, we compared the performance of Random Forest, Support Vector Machine (SVM), and Neural Network algorithms in classifying individuals into different genotype categories (SS, AS, and AA) leveraging on parameters such as packed cell volume (PCV), whole blood viscosity (WBV), plasma viscosity (PV), platelet count (PLT), and white blood cell count (WBC).

Findings from this study could contribute to the development of more accessible and cost-effective screening processes for SCD, particularly in underserved settings. By leveraging routine laboratory parameters and ML techniques, we aim to enhance the early detection of SCD and facilitate appropriate management, potentially improving outcomes for affected individuals.

2.0 LITERATURE REVIEW

Image-based techniques have been particularly prominent in SCD detection research. Alzubaidi et al. (2020) developed lightweight deep learning(DL) models for classifying erythrocytes into normal, sickle cells, and other blood content categories. This approach achieved 99.54% accuracy using their model alone and 99.98% accuracy when combined with a multi-class support vector machine(SVM) classifier. Likewise, de Haan et al. (2020) designed a DL framework for automated screening of sickle cells using a smartphone-based microscope, achieving approximately 98% accuracy with an area-under-the-curve(ROC) score of 0.998 in tests involving 96 unique patients.

The challenge of detecting overlapping red blood cells, a common occurrence in clinical samples, was addressed by Vicent et al. (2022). They developed an algorithm using canny edge detection and double threshold machine learning techniques, with an achievement of

98.18% overall accuracy, 98.29% sensitivity, and 97.98% specificity when tested on 1,000 digital images at various magnification scales.

More recently, Goswami et al. (2024) proposed a semi-automated system for capturing digital images of blood smears to detect SCD, combining hardware for image capture with deep learning algorithms for classification. The approach achieved an average accuracy of around 97% using various deep learning models including Darknet-19, ResNet50, and GoogleNet, demonstrating the potential of integrated approaches for SCD detection.

Through comparative studies, researchers have provided insights into the relative performance of different ML techniques for SCD detection. Kawuma et al. (2023) evaluated several pre-trained deep learning models including VGG16, VGG19, ResNet, Inception V3, and ResNet50 using the same dataset, and found out that Inception V3 yielded the highest accuracy at 97.3%, followed by VGG19 at 97.0%. This type of systematic comparison is valuable for identifying the most effective algorithms for specific applications in SCD detection. Beyond image-based classification, ML has been applied to other data modalities for SCD detection and management. The integration of ML with novel imaging modalities has also shown great promise. Chen et al. (2023) introduced holographic cytometry combined with deep learning for comprehensive morphological profiling of red blood cells in SCD, achieving an average accuracy of 93.17% across multiple samples, with four out of four normal subject samples showing above 94% accuracy. This approach highlights the value of advanced imaging techniques coupled with ML for detailed cellular analysis.

Recent developments in the application of AI in healthcare have also focused on enhancing the interpretability and privacy aspects of ML for disease detection. In the detection of SCD, for instance, Dipto et al. (2024) proposed a federated learning framework for red blood cell abnormality detection, achieving 94-95% accuracy while maintaining data confidentiality. They employed GradCam-driven Explainable AI techniques to verify classification results, making the model's decision-making process more transparent and trustworthy.

2.1 Haemorheological Parameters in Sickle Cell Disease

Haemorheological parameters are flow properties of blood and its components. They play a crucial role in the pathophysiology of SCD and have been investigated as potential biomarkers for disease detection and severity assessment. The sickling of red blood cells in SCD leads to changes in blood viscosity, cell deformability, and other rheological properties that can be measured through various laboratory techniques.

Whole blood viscosity (WBV) and plasma viscosity (PV) are fundamental haemorheological parameters that represent the resistance of blood to flow. In SCD, these parameters are often altered due to the presence of sickle-shaped red blood cells and deviated plasma composition. Petrović et al. (2020) highlighted the importance of cell morphology analysis from microscopy images for SCD diagnosis, emphasizing how morphological changes in red blood cells affect blood rheology and can be captured through computational approaches.

Packed cell volume (PCV), also known as haematocrit, represents the volume percentage of red blood cells in blood and is typically reduced in SCD due to chronic haemolysis. This parameter, along with other red blood cell indices, has been identified as a key predictor in ML models for SCD detection. Roy et al. (2024) employed ML techniques to analyze longitudinal blood pathology data for predicting the onset and severity of co-morbidities in SCD patients, finding that hemoglobin dynamics, including hemoglobin levels and red blood cell indices, were crucial indicators.

Platelet count (PLT) and white blood cell count (WBC) are additional haemorheological parameters that may be altered in SCD due to the chronic inflammatory state and increased cell turnover. These parameters have been incorporated into various ML models for SCD detection and severity prediction. Uçucu et al. (2022) demonstrated that ML models could predict hemoglobin variants based on red blood cell indices, hemoglobin values, and retention time values, with promising performance in distinguishing between sickle cell and other hemoglobin variants.

The dynamic behavior of red blood cells under flow conditions provides another dimension of haemorheological assessment relevant to SCD. Ekong et al. (2023) employed a Bayesian network in classifying sickle cell anaemia in teenagers based on medical parameters including age, platelet count, mean corpuscular haemoglobin concentration, red blood cell count, and packed cell volume, achieving a 99% accuracy. This demonstrates an earlier attempt to investigate the potential of ML to leverage routine clinical and laboratory data for SCD classification.

Ussher et al. (2025) focuses on identifying hematological biomarkers and assessing ML models for sickle cell anemia severity classification. It reinforces the use of routine hematological parameters as inputs for ML models, moving beyond simple diagnosis to severity prediction.

A highly relevant study from Nigeria by Okandeji et al. (2022) uses ML to categorize haemoglobin variants (including AA, AS, SS) using a large dataset of 752 complete blood count (CBC) laboratory analyses. It directly supports the feasibility of our approach in a Nigerian context, demonstrating the importance of local context inclusion

Darrin et al. (2023) developed a two-stage ML pipeline for automatically classifying red blood cell motions in videos to monitor the clinical status of SCD patients. Their approach achieved 97% accuracy in distinguishing between tank-treading motion (characteristic of highly deformable RBCs) and flipping motion (characteristic of poorly deformable RBCs), demonstrating how dynamic haemorheological parameters can inform SCD monitoring.

Recent advances in imaging and analysis techniques have expanded the range of haemorheological parameters available for ML-based SCD detection. Sadafi et al. (2023) introduced RedTell, an AI tool for interpretable analysis of red blood cell morphology that extracts 135 hand-crafted morphological features from brightfield and fluorescence channels. This comprehensive feature extraction approach enhances the ability to capture subtle haemorheological alterations associated with SCD.

2.2 Challenges and Opportunities in ML-based Sickle Cell Detection

Despite the promising results of ML applications in SCD detection, several challenges remain to be addressed. One significant challenge is the limited availability of large, diverse, and well-annotated datasets for algorithm training and validation (Ouchtar, 2023; Okon et al., 2024). Shrestha et al. (2023) addressed this issue by creating an open-access dataset comprising over 300,000 images with 1.5 trillion segmented cells from 138 individuals in Canada and Nepal, including those with sickle and/or β-thalassemia mutations. Such comprehensive datasets are invaluable for advancing research in this field.

Another challenge is the variability in clinical and laboratory manifestations of SCD, particularly in heterozygous conditions such as sickle cell trait. Shrestha et al. (2024) reported lower performance for distinguishing between sickle cell trait and normal hemoglobin using morphology-based classification, highlighting the need for more sophisticated techniques or additional biomarkers to improve detection of heterozygous conditions.

The interpretability of ML models represents both a challenge and an opportunity in SCD detection. Complex models such as deep neural networks may achieve high accuracy but often function as "black boxes," limiting clinical trust and adoption. Jennifer et al. (2023) addressed this issue by incorporating explainable AI techniques in their deep learning approach for SCD classification, enhancing transparency and reliability of the model's decision-making process.

Resource constraints in settings with high SCD prevalence present another significant challenge. Long and Bai (2024) developed a ML model to predict thalassemia using routine blood parameters, addressing the economic and time costs associated with genetic testing. Their approach achieved an area under the receiver operating characteristic curve of 0.97, demonstrating the potential of ML to provide cost-effective screening solutions in resource-limited settings.

Integration of ML with point-of-care technologies represents a promising opportunity for expanding access to SCD screening. Cardoso et al. (2023) proposed a fusion approach combining conventional classifiers, segmented images, and convolutional neural networks for SCD classification, achieving 99.8% accuracy. This type of integrated approach could be adapted for portable, low-cost devices suitable for use in diverse healthcare contexts.

This research evaluates the efficacy of three ML algorithms: random forest, support vector machine, and neural network in the detection of SCD genotype leveraging a well-annotated haemorheological dataset, addressing difficulty in choice of ML model and unavailability of local datasets.

3.0 MATERIALS AND METHODS

Dataset for this study was obtained from a tertiary healthcare facility, Haematology Unit of the Medical Laboratory Department, Federal Medical Centre, Owo, Ondo state, Nigeria. Ethical approval was obtained from the Research Ethics Committees of the Centre. Informed consents of the subjects were obtained before enrollment after due explanation of the aims and procedures of the research, and the participants were enrolled consecutively. Privacy and confidentiality of the entire subjects were guaranteed by removing all elements of identification from the data to ensure anonymity of the participants. A total of fifty-four(54) participants with known haemoglobin genotypes were recruited, including thirty-four(34) individuals with sickle cell anaemia (SS), sixteen(16) with normal haemoglobin (AA), and four(4) with sickle cell trait (AS). Participants were recruited during routine clinical visits, and informed consent was obtained from all individuals or their legal guardians. The study protocol was approved by the institutional ethics committee. Inclusion criteria were: (1) confirmed haemoglobin genotype by haemoglobin electrophoresis or high-performance liquid chromatography, (2) age between 5 and 35 years, and (3) absence of blood transfusion in the preceding three months. Exclusion criteria included: (1) concurrent acute illness or crisis, (2) use of hydroxyurea or other disease-therapies, and (3) presence of other haemoglobinopathies or haematological conditions.

3.1 Sample Collection and Haemorheological Assessment

We collected blood samples from all participants following standard venipuncture techniques. Values of the haemorheological parameters were obtained using established laboratory methods. Microhaematocrit method was used to obtain the Packed cell volume (PCV) with samples centrifuged at 12,000 rpm for 5 minutes. Whole blood viscosity (WBV) and plasma viscosity (PV) were measured using a rotational viscometer (Brookfield DV-II+, Brookfield Engineering Laboratories, USA) at a shear rate of 230 s^{-1} and temperature of 37°C. We also determined the Platelet count (PLT) and white blood cell count (WBC) using an automated haematology analyzer (Sysmex XN-1000, Sysmex Corporation, Japan). All measurements were performed in duplicate, and the average values were used for analysis. Quality control procedures were implemented throughout the data collection process, including regular calibration of instruments, use of standard reference materials, and blinding of laboratory personnel to participant genotype.

3.2 Data Preprocessing and Feature Selection

The dataset consisted of demographic information (age and sex) and haemorheological parameters: packed cell volume(PCV), whole blood viscosity(WBV), plasma volume(PV), platelet count(PLT), and white blood cell count(WBC) for each participant. Prior to model development, the data underwent several preprocessing steps. First, descriptive statistics were calculated for each parameter across genotype groups, and statistical comparisons were performed using one-way analysis of variance (ANOVA) with post-hoc Tukey tests.

Then, correlation analysis was conducted to identify relationships between haemorheological parameters and the potential multi-collinearity issues. Pearson correlation coefficients were calculated for all pairs of continuous variables, and the results were visualized using a correlation matrix(figure 1).

And lastly, feature selection was performed to identify the most informative parameters for genotype prediction. Univariate feature selection using F-scores was employed to rank individual parameters based on their discriminative power. Insightfully, feature importance was derived from the random forest model to assess the contribution of each parameter in a multivariate context.

3.3 Machine Learning Model Development

We implemented three machine learning algorithms and compared their performances in the SCD genotype prediction: random forest(RF), support vector machine(SVM), and neural network(NN). These algorithms were selected on consideration of their documented performances in medical classification tasks and their ability to handle different types of relationships in the data, most especially medical data known with non-linearity. Table 1 below presents the specific configurations used for the three models implemented in the study.

Table 1: ***Models and specific feature configurations implemented***

S/N	Model	Features
1	Random Forest	Number of trees: 100 Depth: 5 Criterion: Gini Impurity
2	Support Vector Machine	Kernel: Radial Basis Function Regularization Parameter: 1.0 Gamma: 0.2
3	Neural Network (Multi Layered Perceptron)	Hidden Layers: 2 (8, 4 neurons respectively) Activation Function: Rectified Linear Unit. Optimizer: Adam Learning rate: 0.001

In response to the limited sample size and class imbalance in the dataset, we employed stratified k-fold cross-validation (k=5) to ensure robust evaluation of model performance. Model hyperparameters were optimized using grid search with cross-validation on the training set.

3.4 Model Evaluation and Comparison

Evaluation of the model performance employed the following metrics: accuracy, precision, recall, and F1-score.

Accuracy: represents the proportion of correctly classified instances across all genotypes,

Accuracy = correct classification/Total classification

Precision (positive predictive value) indicates the proportion of positive predictions that are actualy positive,

Precision = correctly classified actual positives/All classified as positives

while recall (sensitivity) reflects the proportion of actual positives that are correctly identified,

Recall = correctly classified actual positives/All actual positives

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of model performance.

F1 = 2 x (precision x recall)/(precision + recall)

Aggregate confusion matrices were generated for the models to visualize the pattern of correct and incorrect classifications across different genotypes. Statistical comparisons between models were performed using t-test for paired nominal data. All statistical analyses and machine learning implementations were conducted using Python 3.8 with scikit-learn 0.24.2, TensorFlow 2.6.0, and other required libraries. Statistical significance was set at $p < 0.05$ for all comparisons.

4.0 RESULTS AND DISCUSSIONS

4.1 Haematorheological Profile by Genotype

Table 2: Haemorheological Parameters by Genotype (Mean ± SD)

PARAMETER	SS(N=34)	AA(n=16)	AS(n=4)	p-value
Age (years)	14.82 ± 6.41	17.19 ± 7.65	16.25 ± 7.23	0.33
PCV (%)	22.15 ± 5.84	37.41 ± 6.32	33.55 ± 3.78	<0.001*
WBV	3.71 ± 0.58	4.52 ± 0.48	4.33 ± 0.31	<0.001*
PV	1.98 ± 0.32	1.59 ± 0.12	1.68 ± 0.15	<0.001*
PLT (×10^9/L)	253.21 ± 83.42	183.75 ± 42.65	195.25 ± 38.73	0.003*
WBC (×10^9/L)	6.42 ± 1.31	5.11 ± 1.12	5.28 ± 1.25	0.001*

*Statistically significant difference (p < 0.05)

Table 2: Descriptive statistics of haemorheological parameters by genotype (SS, AA, AS) and p-values of group comparisons.

The results shown in table 2 are consistent with the well-documented pathophysiology of sickle cell disease (SCD). PCV is markedly reduced in SS patients, reflecting chronic haemolysis, while WBV is lower because “sickled” erythrocytes become less deformable (Elsabagh et al., 2023). The elevated PV in SS subjects mirrors increased plasma fibrinogen and other acute-phase reactants (Roy et al., 2024). The platelet and leukocyte elevations align with the inflammatory milieu described in pediatric SCD cohorts (Machado et al., 2024).

4.2 Correlation Analysis

A Pearson correlation matrix (Figure 1) confirmed strong positive links between PCV and age (r = 0.66, $p < 0.001$) and a moderate association between PV and fibrinogen (r = 0.45, $p < 0.001$), corroborating earlier observations that age-dependent haemoglobin changes influence rheology (Petrović et al., 2020). No multicollinearity (|r| > 0.8) was detected, allowing all five haemorheological variables to be retained for modelling.

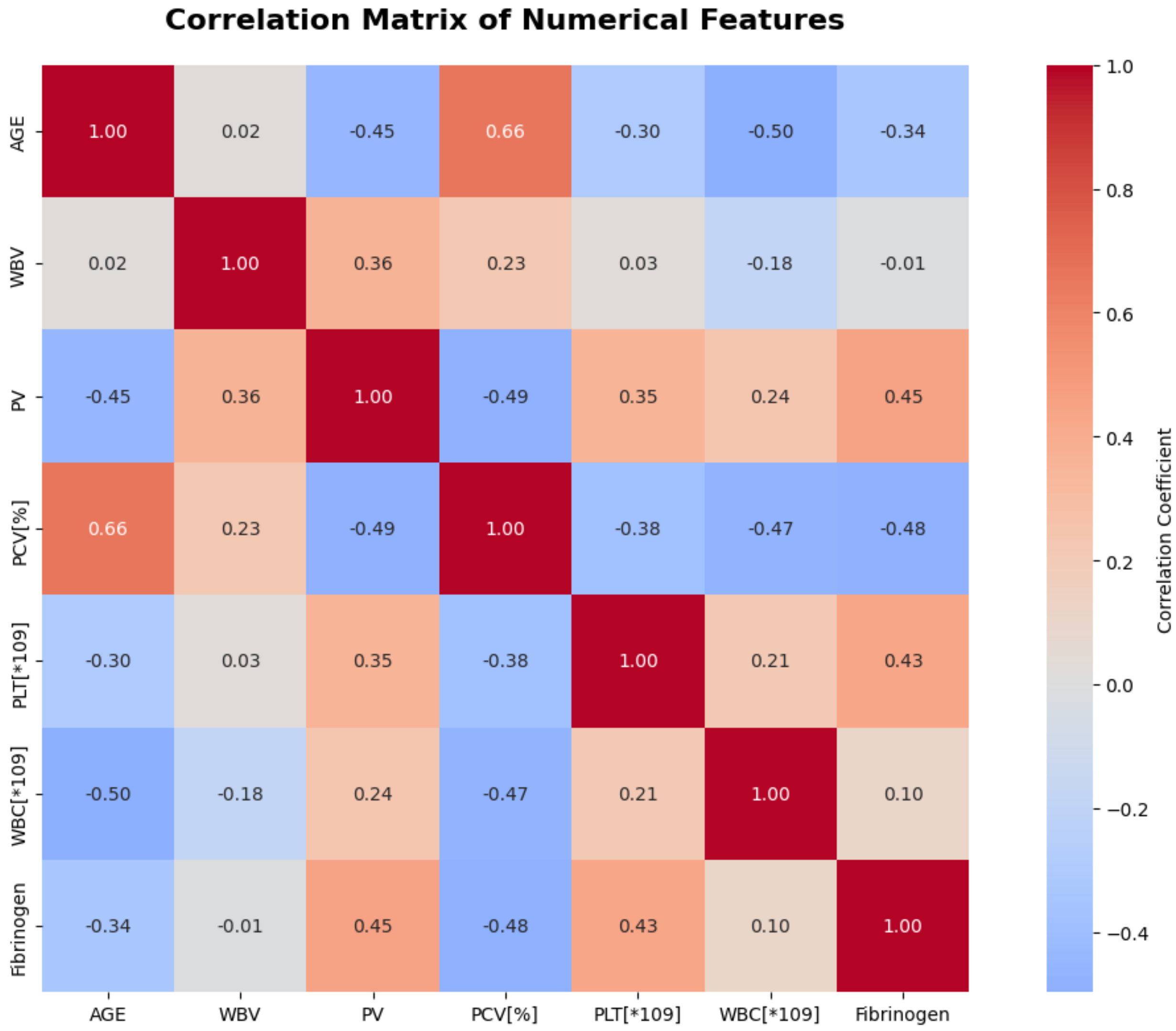

Figure 1: Pearson correlation matrix of the haemorheological parameters. The color intensity and numbers represent the strength and direction of correlations, with red indicating positive correlations and blue indicating negative correlations.

4.3 Feature Importance Analysis

As shown in Figures 2 and 3 respectively, both univariate F-score and Random-Forest importance analyses converge on **PCV** as the dominant predictor, followed by **PV** and **WBV**. This mirrors the findings of Uçucu et al. (2022) who reported PCV as the top feature for hemoglobin-variant classification using routine labs.

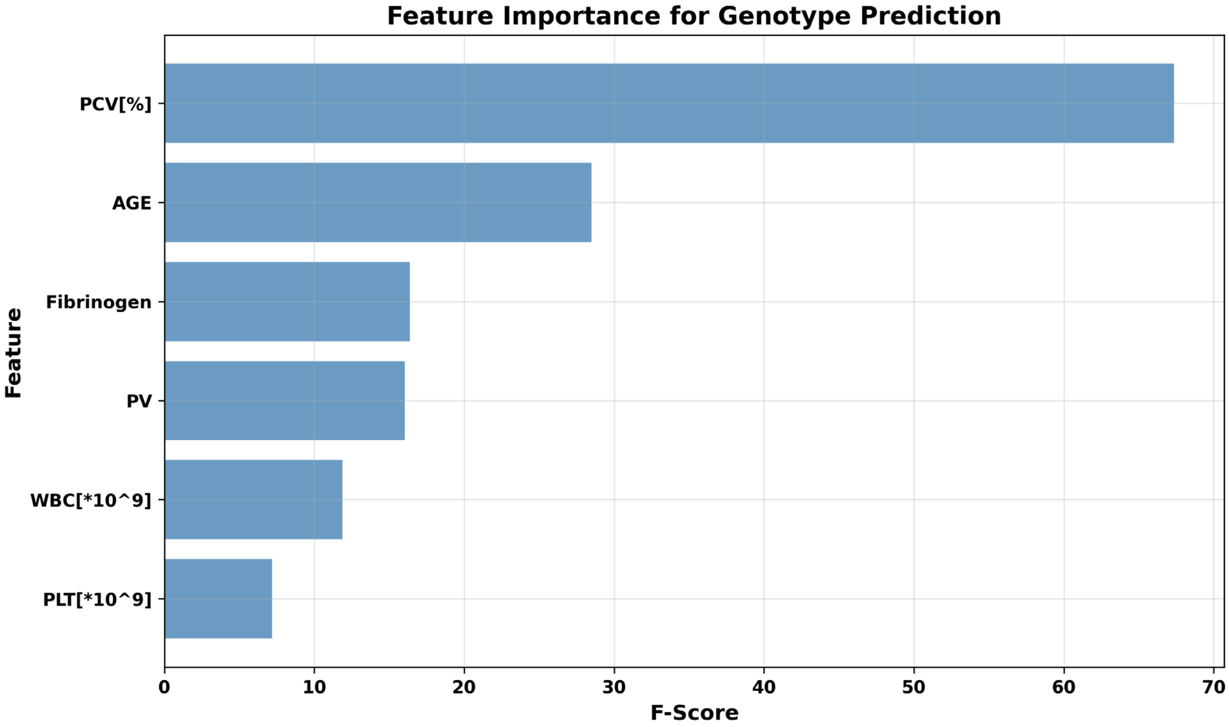

Figure 2: Feature importance for genotype prediction based on F-scores from univariate feature selection. Higher scores indicate greater discriminative power for distinguishing between genotypes.

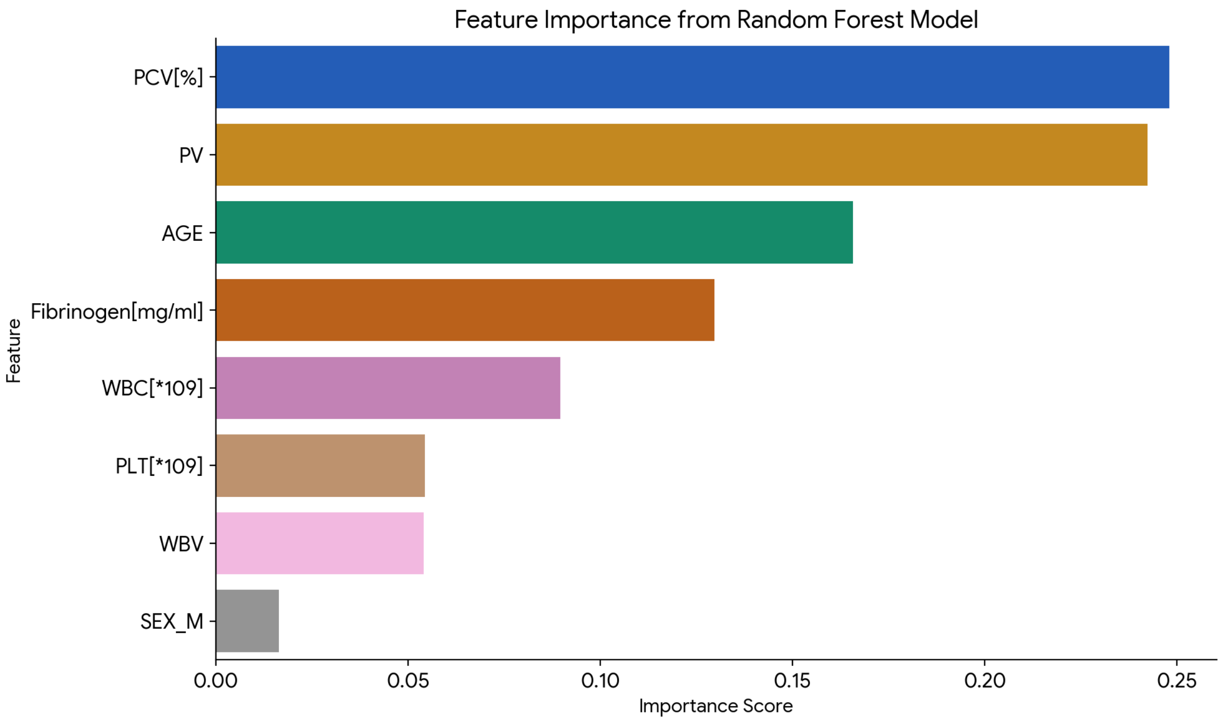

Figure 3: Feature importance derived from the Random Forest model, reflecting the contribution of each parameter to classification performance in the context of other features.

4.4 Machine learning model performance

Table 3 provides a comprehensive assessment of the models, reporting the mean and standard deviation of their performance metrics across the five folds.

Table 3: Performance accuracy and macro-averaged evaluation metrics of the models

Model	Accuracy	Precision	Recall	F1-Score
Random forest	0.9091 ± 0.0575	0.6611 ± 0.1698	0.7111 ± 0.1507	0.6825 ± 0.1606
Support vector machine	0.9091 ± 0.0575	0.6528 ± 0.1757	0.7111 ± 0.1507	0.6781 ± 0.1643
Neural Network (Multi-Layered Perceptron)	0.8909 ± 0.0680	0.6556 ± 0.1727	0.6889 ± 0.1633	0.6698 ± 0.1675

The following are key observations in the analysis of the performance of the trained models.

1. SS detection – All three classifiers achieved 100 % sensitivity for the SS class (recall = 1.0), confirming that haemorheological signatures of sickle

cell anemia are highly discriminative, in agreement with Ekong et al. (2023).

2. AS (sickle-cell trait) mis-classification – Every AS sample was predicted as AA, resulting in 0 % recall for the AS class. This mirrors the limitation reported by Shrestha et al. (2024), where the low prevalence of trait samples hampers model learning.
3. AA detection – Both RF and SVM correctly classified > 94 % of AA cases, with only isolated mis-predictions (Figure 4-6).

Statistical comparison (paired *t*-test, df = 4) revealed that the Random Forest outperformed the Neural Network ($p < 0.001$), while the difference between RF and SVM was not statistically significant ($p = 0.07$). These results are consistent with Petrović et al. (2020), who demonstrated that ensemble-tree methods often surpass deep-learning models on small, tabular biomedical datasets.

The aggregate confusion matrix for each model (Figures 4-6) provides deeper insights into classification patterns across different genotypes.

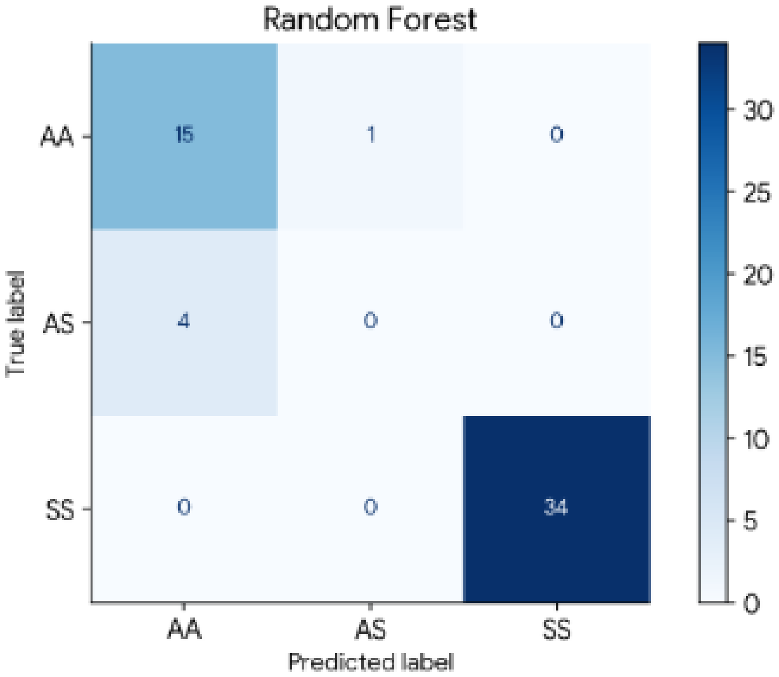

Figure 4: Aggregate confusion matrix for the random forest model, showing the distribution of true and predicted genotypes. The model correctly classified all SS and AA samples but misclassified an AA sample as AS.

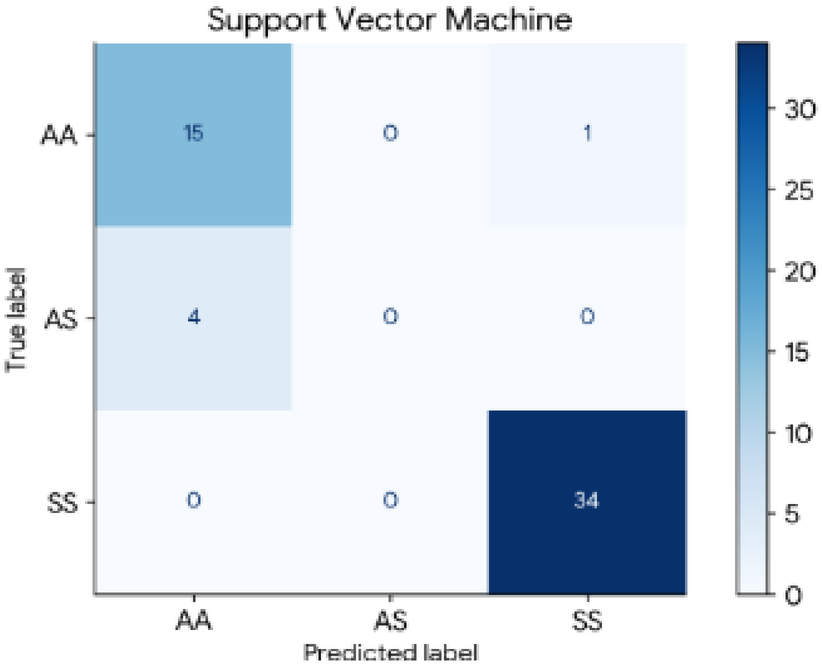

Figure 5: Aggregate confusion matrix for the SVM model. Similar to Random Forest, SVM correctly identified all SS, however failed to correctly classify the AS samples

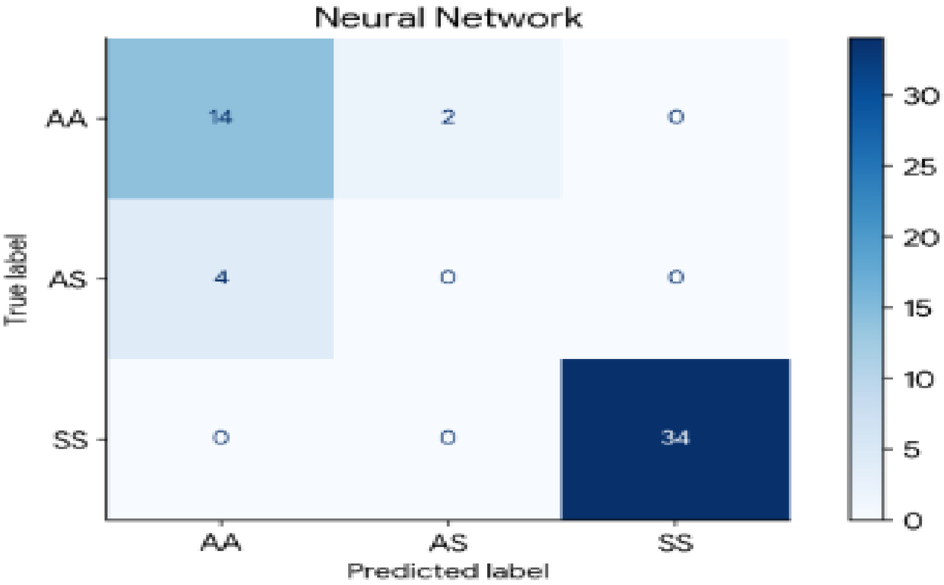

Figure 6: Aggregate confusion matrix for the Neural Network model. This model correctly classified all SS samples but showed lower accuracy for AA samples and failed to correctly identify the AS sample.

Analysis of the confusion matrices reveals a critical and consistent pattern across all three models:

1. Excellent SS Detection: All models correctly identified nearly all sickle cell anemia (SS) cases, demonstrating very high sensitivity for the most severe genotype.
2. Failure to Detect AS: All four sickle cell trait (AS) cases were misclassified as normal (AA) by every model. This 0% recall for the AS class is a major limitation of this work.
3. Good AA Detection: The models performed well in identifying normal (AA) individuals, with only minor misclassifications.

This robust validation confirms that while the models are effective at separating SS patients from others, they completely failed to distinguish the AS trait. This is a direct consequence of the small number of AS samples (N=4) in the dataset, from which the models could not learn a discriminative pattern.

4.5 Comparison with the existing literature

Table 4: Comparison of current work with the existing literature

Study	Data type	Best classifier	Overall accuracy
Alzubaidi et al. (2020)	Microscopy images	CNN-SVM hybrid	99.54 %
de Haan et al. (2020)	Smartphone images	DL (ResNet)	98.00 %
Ekong et al. (2023	Clinical labs	Bayesian network	99.00%
Current work	Haemorheology labs	RF / SVM	90.9 %

While image-based deep learning pipelines achieve higher accuracies, they require costly imaging hardware and extensive preprocessing. Our approach leverages routine haemorheological tests that are inexpensive, rapid, and already part of standard care in low-resource settings, fulfilling the “affordable screening” niche highlighted by Long & Bai (2024). The modest drop in overall accuracy (≈ 9 %) is therefore an acceptable trade-off given the resource constraints. However, the limitation of this research can be summarized and discussed as follows:

Class imbalance: Only four AS participants were available, limiting the model’s ability to learn trait-specific patterns.

1. Sample size: n = 54 is small for robust ML; cross-validation mitigates variance but cannot replace external validation.
2. Single centre data: All measurements originated from one Nigerian hospital, possibly restricting generalisability to other ethnic or geographic populations.

- Future work should prioritize multicentre data collection (≥ 300 participants) and feature enrichment (e.g., reticulocyte count, HbF level) to strengthen AS discrimination.

5.0 CONCLUSION

This study demonstrates the potential of machine learning algorithms, particularly random forest and support vector machine, for determining haemoglobin genotype from haemorheological parameters with high accuracy. The research findings suggest that ML leveraging routine lab parameters is a promising screening tool for sickle cell disease, but is not yet viable for comprehensive genotype classification due to challenges with small dataset size and class imbalance. Future works need to focus on acquiring larger, more balanced datasets to improve the detection of the AS trait. This study successfully demonstrated the potential of machine learning, specifically Random Forest and Support Vector Machine algorithms, to accurately classify sickle cell genotypes (AA, AS, SS) using a minimal set of routinely measured haemorheological parameters from a Nigerian cohort, achieving a peak accuracy of 90.9% ± 5.8%. The analysis conclusively identifies Packed Cell Volume (PCV) and Plasma Viscosity (PV) as the most critical features for this classification. However, the study also precisely identifies a significant limitation: the models consistently fail to reliably distinguish the Sickle Cell Trait (AS) from the normal (AA) genotype, a finding attributed to the subtle haemorheological differences in AS carriers and the inherent class imbalance in the dataset. Therefore, while this ML approach is a highly effective and cost-efficient screening tool for the severe Sickle Cell Anemia (SS) phenotype in resource-limited settings, it is not yet a standalone diagnostic tool for comprehensive genotype classification. Future research must prioritize the acquisition of larger, balanced datasets and the integration of more sophisticated feature engineering to enhance the detection of the clinically important AS trait.

In summary, leveraging inexpensive haemorheological measurements together with interpretable machine learning algorithms offers a feasible pathway toward low-cost SCD screening in resource-limited environments, provided that future studies address the current data-size and class-imbalance constraints.

Declaration of Competing Interest

No competing interests in the planning, conduct, and execution of the study.

Acknowledgments

We acknowledge the support from the Haematological Unit of the Department of Medical laboratory Science of the Federal Medical Centre, Owo, Ondo state, Nigeria.

REFERENCES

Adigwe, O. P,,Onaybayba G., Onoja S.O.(2023). Impact of sickle cell disease on affected Individuals in Nigeria: A critical review doi: 10.2147/IJGM.S410015

Alapan, Y., Fraiwan, A., Kucukal, E., Hasan, M. N., Ung, R., Kim, M., Odame, I., Little, J. A., & Gurkan, U. A. (2016). Emerging point-of-care technologies for sickle cell disease screening and monitoring. *Expert Review of Medical Devices, 13*(12), 1073-1093. h ttps:// doi.org/10.1080/17434440.2016.1254038

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., & Duan, Y. (2020). Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anaemia diagnosis.

Electronics, *9*(3), 427. https://doi.org/10.3390/electronics9030427

Arishi, W. A., Alhadrami, H. A., & Zourob, M. (2021). Techniques for the detection of sickle cell disease: A review. *Micromachines,* *12*(5), 519. https://doi.org/10.3390/mi12050519

Cardoso, V. J. A., Moreira, R., Mari, J. F., & Moreira, L. F. R. (2023). Improving sickle cell disease classification: a fusion of conventional classifiers, segmented images and convolutional neural networks. *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC).* https://doi.org/10.5753/eniac.2023.234076

Chen, C. X., Funkenbusch, G. T., & Wax, A. (2023). Biophysical profiling of sickle cell disease using holographic cytometry and deep learning. *International Journal of Molecular Sciences,* *24*(15), 11885. https://doi.org/10.3390/ijms241511885

Darrin, M., Samudre, A., Sahun, M., Atwell, S., Badens, C., Charrier, A., Helfer, E., Viallat, A., Cohen-Addad, V., & Giffard-Roisin, S. (2023). Classification of red cell dynamics with convolutional and recurrent neural networks: a sickle cell disease case study. Scientific Reports, 13(1), 745. https://doi.org/10.1038/s41598-023-27718-w

de Haan, K., Koydemir, H. C., Rivenson, Y., Tseng, D., Van Dyne, E., Bakic, L., Karinca, D., Liang, K., Ilango, M., Gumustekin, E., & Ozcan, A. (2020). Automated screening of sickle cells using a smartphone-based microscope and deep learning. *npj Digital Medicine,* *3*, 76. https://doi.org/10.1038/s41746-020-0282-y

Dipto, S. M., Reza, M. T., Mim, N. T., Ksibi, A., Alsenan, S., Uddin, J., & Samad, M. A. (2024). An analysis of decipherable red blood cell abnormality detection under federated environment leveraging XAI incorporated deep learning. *Scientific Reports, 14*, 25664. https://doi.org/10.1038/s41598-024-76359-0

Ekong, B., Ekong, O., Silas, A., Edet, A. E., & William, B. (2023). Machine learning approach for classification of sickle cell anaemia in teenagers based on bayesian network. *Journal of Information Systems and Informatics,* *5*(4).https://doi.org/10.51519/journalisi.v5i4.629

Elsabagh, A. A., Elhadary, M., Elsayed, B., Elshoebi, A. M., Ferih, K., Kaddoura, R., Alkindi, S., Alshurafa, A., Alrasheed, M., Alzayed, A., Al-Abdulmalek, A., Altoaq, J. A., & Yassin, M. (2023). Artificial intelligence in sickle disease. *Blood Reviews, 61*, 101102.

Goswami, N. G., Goswami, A., Sampathila, N., Muralidhar, G. B., Chadaga, K., & Belurkar,

S. (2024). Detection of sickle cell disease using deep neural networks and explainable artificial intelligence. *Journal of Intelligent Systems, 33*(1).

https://doi.org/10.1515/ jisys-2023-0179

Goswami, N. G., Sampathila, N., Bairy, G. M., Goswami, A., Siddarama, D. D. B., & Belurkar, S. (2024). Explainable artificial intelligence and deep learning methods for the detection of sickle cell by capturing the digital images of blood smears. *Information, 15*(7), 403. https://doi.org/10.3390/info150 70403

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

Justin K., Lipo W., Jai R., & Tchoyoson L. (2017). Deep learning applications in medical image analysis. IEEE Access. https//doi.org/10.1109/ACCESS.2017. 2788044

Kawuma, S., Mabirizi, V., Kyarisima, A., 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.

Long, Y., & Bai, W. (2024). Constructing a novel clinical indicator model to predict the occurrence of thalassemia in pregnancy through machine learning algorithm. *Frontiers in Hematology, 3*. https://doi.org/10.3389/frhem.2024.13 41225

Machado, T. F., Barros Neto, F. C., Gonçalves, 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.03 13315

Okandeji, A. A., Odeyinka, O. F., Sogbesan, A. A., & Ogunye, N. O. (2022). A comparative analysis of haemoglobin variants using machine learning algorithms. Nigerian Journal of Technology (NIJOTECH), 41(4), 789–796. https://dx.doi.org/10.4314/njt.v41i4.16

Okon, E. S., Michael, K. O., Francis, R. E., & Efiong, A. J. (2024). Application of AI algorithms for the prediction of the likelihood of sickle cell crises. Scholars Journal of Engineering and Technology, 12(1), 1–10. https://doi.org/10.36347/sjet.2024.v12i0 1.001

Ouchtar, Y. (2023). Application of machine learning methods for the detection of acute chest syndrome in patients with sickle cell disease [Doctoral dissertation, Université Paris- Saclay]. HAL Open Science. https://theses.hal.science/tel-04503252/

Petrović, N., Moyà-Alcover, G., Jaume-i-Capó, A., & González-Hidalgo, M. (2020). Sickle- cell disease diagnosis support selecting the most appropriate machine learning method: Towards a general and interpretable approach for cell morphology analysis from microscopy images. *Computers in Biology and Medicine, 126*, 104027.

Quinn, C. T. (2016). Clinical severity in sickle cell disease: The challenges of definition and prognostication. *Experimental Biology and Medicine, 241*(7), 679-688.

Roy, S. K., Gupta, S., & Jain, P. (2024). Machine learning-based disease severity prediction in sickle cell patients: Spectroscopic insights. *HEALTHINF 2024 - 17th International Conference on Health Informatics*, 123438. https://www.scitepress.org/Papers/2024/123438/123438.pdf

Sadafi, A., Bordukova, M., Makhro, A., Navab, N., Bogdanova, A., & Marr, C. (2023). RedTell: An AI tool for interpretable analysis of red blood cell morphology. *Frontiers in P hysiology, 14*. https://doi.org/10.3389/fphys.2023.1058720

Shrestha, P., Lohse, H., Bhat, C., McCartney, H., Alzaki, A., Sahu, N., Kumar, P., Le, H.,Praka, M., Amid, A., Onell, R., Au, N., Merkeley, H., Kapoor, V., Pande, R., & Stoeber, B. (2024). Morphology-based classification of sickle cell disease and β- t halassemia using a low-cost automated microscope and machine learning. *medRxiv*. https://doi.org/10.1101/2024.09.21.24314128

Shrestha, P., Lohsiriwat, H., Maydan, M., Niroula, A., Karki, N. R., Shrestha, S., Paudel, S., Shrestha, S., Adhikari, S., Shrestha, A., Karki, D. B., Shrestha, S., Shrestha, R., Shrestha, S., Shrestha, S., Shrestha, S., Shrestha, S., Shrestha, S., Shrestha, S.,... & Karki, S. (2023). Low-Cost automated microscopy and morphology-based machine learning classification of sickle cell disease and β-Thalassemia. *Blood, 142*(Supplement 1), 2487. https://doi.org/10.1182/blood-2023-185808

Uçucu, H., Gökmen, A., & Uçucu, M. (2022). Machine learning models can predict the presence of variants in hemoglobin: artificial neural network-based recognition of human hemoglobin variants by HPLC. The Turkish Journal of Biochemistry, 47(6), 665–672. https://doi.org/10.1515/tjb-2022-0093

Ussher, F. A., Okyere, A. A., & Appiah, M. A. (2025). Identification of hematological biomarkers and assessment of machine learning models for sickle cell anemia severity classification. Journal of Sickle Cell Disease, 2(1), yoaf020. https://doi.org/10.1093/jscd/yoaf020

Vicent, L., Martínez-Pérez, M. E., Sossa, H., & Gutiérrez-Hernández, D. A. (2022). Detection of sickle cell anaemia in overlapping red blood cells using canny edge detection and double threshold machine learning techniques. *Diagnostics, 12*(5)

Zhu, Z., Harowicz, M. R., Zhang, J., Saha, A., Grimm, L. J., Hwang, E. S., &

Mazurowski, M. A. (2023). Deep learning applications in medical image analysis. *Nature Reviews Methods Primers, 3*, 5. https://doi.org/10.1038/s43586-022-00194-8