EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears

Oluwafisayo Babatope Ayoade[1, *] [0000-0003-2116-8059], Tinuke Omolewa Oladele[2], Agbotiname Lucky Imoize[3], Jerome Adetoye Adeloye[4], Joseph Bamidele Awotunde[5], Segun Omotayo Olorunyomi[6], Olusola Theophilus Faboya[7], and Ayorinde Oladele Idowu[8]

[1,7,8] Department of Computing and Information Science, School of Pure and Applied Sciences, College of Science, Bamidele Olumilua University of Education, Science & Technology, Ikere-Ekiti, Ekiti State. Nigeria
[2,4,5] Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Kwara State. Nigeria
[3] Department of Electrical and Electronics, Faculty of Engineering, University of Lagos, Akoka, Lagos State 100213. Nigeria
[6] Ekiti State Data Center Office, Old Governor's Office, Ado-Ekiti, Ekiti State. Nigeria

Abstract

INTRODUCTION: The human blood as a collection of tissues containing Red Blood Cells (RBCs), circular in shape and acting as an oxygen carrier, are frequently deformed by multiple blood diseases inherited from parents. These hereditary diseases of blood involve abnormal haemoglobin (Hb) or anemia which are major public health issues. Sickle Cell Disease (SCD) is one of the common non-communicable disease and genetic disorder due to changes in hematological conditions of the RBCs which often causes the inheritance of mutant Hb genes by the patient..
OBJECTIVES: The process of manual valuation, predictions and diagnosis of SCD necessitate for a passionate time spending and if not done properly can lead to wrong predictions and diagnosis. Machine Learning (ML), a branch of AI which emphases on building systems that improve performance based on the data they consume is appropriate. Despite previous research efforts in predicting with single ML algorithm, the existing systems still suffer from high false and wrong predictions.
METHODS: Thus, this paper aimed at performing comparative analysis of individual ML algorithms and their ensemble models for effective predictions of SCD (elongated shapes) in erythrocytes blood cells. Three ML algorithms were selected, and ensemble models were developed to perform the predictions and metrics were used to evaluate the performance of the model using accuracy, sensitivity, Receiver Operating Characteristics-Area under Curve (ROC-AUC) and F1 score metrics. The results were compared with existing literature for model(s) with the best prediction metrics performance..
RESULTS: The analysis was carried out using Python programming language. Individual ML algorithms reveals that their accuracies show MLR=87%, XGBoost=90%, and RF=93%, while hybridized RF-MLR=92% and RF-XGBoost=99%. The accuracy of RF-XGBoost of 99% outperformed other individual ML algorithms and Hybrid models.
CONCLUSION: Thus, the study concluded that involving hybridized ML algorithms in medical datasets increased predictions performance as it removed the challenges of high variance, low accuracy and feature noise and biases of medical datasets. The paper concluded that ensemble classifiers should be considered to improve sickle cell disease predictions.

Keywords: Sickle Cell Disease, Erythrocytes, Machine Learning Algorithms, Ensemble Models, Health Information System

Received on June 21 2023, accepted on August 30 2023, published on 19 September 2023

doi: 10.4108/eetpht.9.3913

*Corresponding author. Email: ayoade.oluwafisayo@bouesti.edu.ng

1. Introduction

Malformed red blood cells, which are the primary cause of sickle cell anemia, can obstruct blood flow, resulting in painful symptoms, and ultimately result in death. Aberrant

hemoglobin is another term for it. Hemoglobin is responsible for carrying oxygen throughout the body's blood vessels. Because they are round, flexible, and compact, in which normal red blood cells may readily flow through microscopic capillaries. On the other hand, they get stuck in microscopic capillaries because of the sickle-shaped, stiff, and angular abnormal red blood cells. This may result in pain for the patients, low oxygen levels, and dehydration [1, 2]. SCD is one of the world's most dangerous illnesses, according to [11, 12, 17], with a wide range of symptoms ranging from moderate to deadly. Sickle Cell Disease (SCD) is one of Nigeria's most frequent hereditary RBC diseases, affecting persons of all ages and resulting in hemolysis and vaso-occlusive crises. According to research, Nigeria has the biggest number of people living with SC illness, with over 150,000 parturitions (births) every year. SCD patients showed hemoglobin anomalies in their RBCs, with hemoglobin S or sickle hemoglobin showing up, according to the survey [12-13]. The basic medical indicators of SCD in children and adults, as well as the numerous anticipated symptoms, are depicted in Table 1.

There are millions of red blood cells in a single smear, hence, the manual evaluation, categorization, and counting of biological cells requires a significant investment of time and runs the risk of incorrect classification and counting. Additionally, the varied and complicated forms, overlapping cells, and range of colors make it difficult to classify cells [6]. The "age of data," which has been endowed with better computing power and higher storage capacity, yet the difficulties of making sense of these vast data sets have grown substantially. To solve these concerns, several sectors in education, health, business, and organizations are building intelligent systems that leverage applicable ideas and approaches such as data mining, data science, and even machine learning (ML). ML is the most active of all the Computer Science disciplines. As a consequence, it is referred to as algorithmic application and science, and it gives data intelligence and significance [7, 8].

According to [8], new data is continually developing across numerous gathering techniques due to the increasing velocity of data and digital technologies on a daily basis. ML is an area of Computer Science in which computers help people understand data in the same way that humans do. It is a subject in Artificial Intelligence that comprises pattern detection and computational learning. The progress of human knowledge technology has an impact on ML, which is an AI subject. This refers to changes in systems that perform various AI-related tasks such as recognition, analysis, planning, prediction, machinelike controls (robotics), and so on [8, 9].

Our society's ability to generate and retain data and information is increasing at an exponential pace every day, but our ability to absorb such an enormous amount of data and knowledge is not keeping up [10]. To address these challenges, it is critical to use first-hand technology to reconcile the seemingly opposing goals of scalability and usability in data and information interpretation. [10] asserted that a plethora of data and information analysis techniques have been created and applied all over the world, all with the goal of increasing their complexity. Humans must be involved in the early phases of data analysis and information production, particularly in a number of organizational sectors, the most significant of which is the healthcare sector, which includes multiple subsections such as genetics, human genetics, and medical genetics. Many governmental and private sector organizations consider improving data analysis procedures to be a significant goal and priority.

Medical data analysis enlightenment is influenced by the goal to deliver innovative and suitable services, as well as the need to increase efficiency and comply with regulations. Healthcare medical systems are flooded with data in today's digital era, but having access to this data and knowledge is not enough to properly capitalize on improving patient health. As a result, both commercial and public healthcare companies must be able to aggregate, align, or anonymize data from many sources, such as electronic health records (EHRs), health surveys, administrative data, physician notes, and consultant reports, to name a few. This would provide the patient a full medical picture of their situation [11]. A wide range of patient data is combined, analyzed, and reported on using healthcare data analysis tools. Additionally, healthcare companies make better administrative, clinical, and financial decisions that increase patient involvement and care. Ence, leveraging Information and Communication Technology (ICT). to modify patient medical data in terms of their health is crucial [12].

Furthermore, as the use of IT in data analysis in the health sector expands, the best framework for patient medical data analysis must be properly handled in order to improve healthcare diagnosis and management. Medical consultants, hematologists, geneticists, and others may find themselves in danger of failing to respect patient decisions when the healthcare business shifts to a value-based paradigm. Therefore, examining medical data in order to make informed decisions is no longer a good idea, but a need [7–9].

Table 1. Medical Indicators of Children and Adult SCD (adapted from [4, 16])

Children with SCD		
Indications/Symptoms		
	Infants	**Children**
	Pain in chest, abdomen, and limbs/joints	Pain (acute or chronic)
	Dactylitis	Acute anemia
	Anemia	Infections
	Mild jaundice	Jaundice
	Enlarged spleen	Poor nutritional status
	Fever	and growth
	Frequent upper respiratory infections	Academic failure
		Delayed puberty
Complications		
CNS	Stroke	
Eye	Retinal artery occlusion/retinopathy	
Lung	ACS	
	Asthma	
Heart	Left ventricular hypertrophy	
	Cardiomyopathy	
Spleen	Acute splenic sequestration	
	Impaired immunity (e.g. bacterial	
Liver	infection, sepsis)	
Kidney		
	Hyposthenia	
	Proteinuria--	
Gall bladder	Renal impairment/failure	
Genitals	Cholelithiasis	
Bones/joints	Priapism	
	Avascular necrosis	
Skin	Aplastic crisis	
	Chronic ulcers, typically on the ankles	
	Adult with SCD showing additional indications and problems	
Indications/Symptoms		
	Severe joint pain	
	Chronic leg ulcers	
	Retinopathy	
	Thromboembolic complications	
	Neurocognitive impairments	
	Narcotic dependence/tolerance	
Complications		
CNS	Recurrent ischemic stroke, hemorrhagic stroke Progressive	
Eye	retinopathy	
Lung	Recurrent ACS	
	Pulmonary hypertension	
	Chronic lung disease	
	Premature coronary artery disease	
Heart	Heart failure	
	Auto-infarction	
Spleen	Functional asplaenia	
Liver	Hepatic sequestration	
	Liver failure secondary to transfusional iron overload	
Kidney	Nephropathy	
	Frequent urinary tract infections	
Gall bladder	Cholelithiasis	
Genitals	Priapism	
Bones/joints	Avascular necrosis	
	Early loss of bone density	
Skin	Chronic leg ulcers	

Keys: CNS (Central Nervous System), ACS (Acute Chest Syndrome) Adapted from [5]

A SCD is an abnormality of the Red Blood Cells (RBCs) caused by a lack of oxygen within the RBCs. The result is a category of disorders known as haemoglobinopathies, with thalassaemia and sickle cell anemia (SCA) being the most common. SCD is a congenital disorder of the hemoglobin structure that can only be cured in children by bone marrow or cord blood transplantation, but with no other therapeutic options for adults [16]. According to [16] and [17], Haemoglobinopathies (abnormal Hb or anemia) are caused by genes occurring in around 5% of the world's population. Because more healthy adults are thalassaemia carriers than sickle cell anemia carriers (who receive just one mutant gene from their parents), there is a high frequency of sickle cell gene carriers compared to the high incidence of afflicted neonates. [3] opined that malaria, cancer, and SCD are among the illnesses for which ML algorithms have been used to model, forecast, and diagnose. SCD is one of the most significant health conditions in the world, with a wide range of symptoms ranging from minor to catastrophic.

In Nigeria, SCD is one of the most common human genetic RBC diseases, causing hemolysis and vaso-occlusive crises in people of all ages. Nigeria has the highest number of persons living with SC disease, according to studies, with over 150,000 parturitions (births) every year. Hemoglobin abnormalities, such as *hemoglobin S* or sickle hemoglobin, are seen in the RBCs of people with SCD [12,13]. Human RBCs include hemoglobin, the protein, that carries oxygen throughout the body. SCD is a non-communicable disease caused by genes that are handed down from generation to generation [19]. It is not infectious like Ebola, Zika, Tuberculosis, or common colds and flu. Years of acute pain (vaso-occlusive episodes (VOEs)), long-term suffering, different organ damage (such as kidney failure, heart failure, and so on), a short lifespan, bone infections, and stroke, to mention a few symptoms and difficulties are all symptoms and issues that patients with SCD face [15,16].

According to [3], [19], and [21] SCD is produced by a single-point mutation (glutamic acid substitution) of valine at position 6 on the beta globin (-globin) subunit of hemoglobin, which results in sickle hemoglobin (*HbS*) deformation or transmutation. [19] submitted that patients who inherit two copies of the *HbS* transmuted or mutations are homozygous (*HbSS*) and have the sickle cell phenotypic form of the disease, whereas those who inherit just one copy of the *HbS* mutation are heterozygous carriers (*HbAS*) and do not have the sickle cell phenotypic form of the disease.

Providing a platform for medical officers, hematologists, and health care managers to strategize on how to make life meaningful and fulfilled for these patients by incorporating ML algorithms into the analysis of SCD medical datasets is critical in order to reduce the rate of mortality and other problems that patients with SCD face [22].

Therefore, ensemble models, a machine learning approach that combines several other models as base estimators in the prediction process, were created, addressing practical difficulties such as considerable variation, low accuracies, and the existence of feature noises and biases. By combining the output from each model and lowering the model's error rate while retaining the model's generalization, building and merging numerous models improves the accuracies of some models [23].

Therefore, the availability of vast databases of relevant genomic data will let researchers to concentrate their efforts on improving various strategies for identifying and treating human genetic illnesses such as SCD. In this context, the utilization of genomic databases for SCD may be difficult to achieve without the right development of sophisticated data analysis tools, which is a major task in and of itself because to the magnitude and variety of the data [24].

As a consequence, this paper presents a health information system that uses ensemble models of machine learning algorithms to predict Sickle Cell Disease from erythrocyte smears. This study aimed at performing a comparative analysis of individual and hybridized ML algorohtms for appropriate predictions of SCD from erythrocytesUDB datasets. The performance of these chosen ML algorithms was examined individually and jointly in order to develop predictions based on a variety of performance parameters for accurate diagnosis and prediction of SCD attribute selection for medical decision-making.

1.1 Motivation

Correctly diagnosing SCD and predicting the patient's chances of survival and life expectancy over time has proven challenging and time demanding for doctors. Traditional or clinical procedures for detecting the presence of sickle cell disease or its mutations, according to research, require a long time. Therefore, ML technologies are increasingly being used to establish a non-clinical diagnosis for SCD patients. SCD is thought to be preventable in 30-50 percent of cases [2,10]. Advanced computer methodologies are required to create precise, objective, and systematic human blood cell predictions.

The value of medical data analysis in the economy of any country cannot be overstated [6]. Data analysis concerns are one of the obstacles that the healthcare industry must overcome, particularly in the context of non-communicable disorder like SCD. The incidence of births in Nigeria is relatively high and at a critical stage when compared to the low acceptable medical infrastructure on the ground for this human genetic concern in medical data evaluation [25]. Missing parameter values (incompleteness), systematic or data random noise, and improper parameter selection have all been issues when learning from large medical datasets like SCD. Using an ensemble model of machine learning techniques, the capacity to deal with the challenges that medical datasets like SCD present will be offered [26].

The study of SCD condition has been the subject of several research studies, mainly in affluent nations and a few emerging nations, particularly in Africa. Others have also worked on various classifications, forecasts, feature extractions, and selections of properties from various blood cell samples.

In spite of earlier research efforts in forecasting using a single ML algorithm, present systems still have a high rate of

erroneous and incorrect predictions. Hence, the goal of this work was to provide an ensemble model of ML algorithms for efficient SCD prediction in erythrocyte smears. The goals were to: execute feature selection on the erythrocytes smears; build an ensemble model using ML techniques for the erythrocytes dataset utilizing the University of the Balearic Islands-UIB and Universidad de Oriente, Cuba-UOC erythrocyteIDB datasets [27]; apply the prediction process; The accuracy, sensitivity, Receiver Operating Characteristics-Area Under Curve (ROC-AUC), and F1 score metrics are used to evaluate the ensemble model's performance, and the ensemble models are then compared to literature for the best prediction metrics performance.
The accuracy of the prediction model is improved by using an ensemble model for choosing and predicting SCD datasets. For proper prediction, the ensemble model would be guided through the training and testing stages (where prediction processes are made from training datasets), and the performance of the chosen ML algorithms would be compared to other literature on SCD morbidity and mortality rates using various model performance metrics.

2 Related Work

While developing mathematical classifiers, ML's method emphasizes the notion of statistics, as evidenced by the classification process. Because the primary purpose is to analyze data from samples, this is valid. For example, during the training phase, a well-organized and competent classifier is needed to store and manage a large amount of data while also resolving any optimization issues that may arise. Once the classifier has been taught, the portrayal and algorithmic solution for interpretations are supposed to be competent. Several machine learning classifiers can be employed to solve the same problems in supervised learning categories including classification and regression [22-23].
In terms of the hybridization of ML algorithms for predictions and feature classification, [30] are alternate and most pertinent works to the topic. Their efforts show how several ML algorithms may be used to anticipate the optimum performance evaluation. In their study, [30] introduced a hybrid method that used Random Forest and Logistic Regression to predict breast cancer survivorship. They employed the RF algorithm to carry out the first variable screening and to get significant rankings, and then fed the extracted data into the LR algorithm, which was in charge of creating understandable models for forecasting breast cancer survival. In their research, they used a 10-fold cross validation and fundamental performance measures to assess the hybridization model. With a basic model, their results yielded a positive better accuracy while they based only on cancer ailment.
The research of [31] was the most pertinent and directly connected to the study of RBCs smear with regard to individual ML algorithms and hybridization. They worked on developing strategies for predicting outcomes based on patients' physical examination markers of hypertension. They divided the patients' physical predictions into two categories and then conducted their study. Key feature extraction from the patients' physical indicators was done in the first phase, and the patients' outcomes were predicted using the characteristics that were extracted. They put forth a model that combines cross-validation, classification techniques, and recursive feature removal. Support Vector Machine (SVM), C4.5 Decision Tree, Extreme Gradient Boosting (XGBoost), and Random Forest were the four classification techniques they took into consideration. Various performance assessment measures were used to examine the chosen ML algorithms. Even though three of the four ML algorithms under consideration produced improved results, XGBoost showed the greatest performance for patient hypertension prediction. Their research showed an advantage that different ML algorithms can yield a level of accurate prediction results while its limitation rest on focusing on physical hypertension markers.
[32] conducted research on the image processing and categorization of blood cells. Their study illustrates several methods for handling blood cell pictures. Through the use of microscopic pictures of blood smears, they proposed techniques for the automated identification and categorization of red blood cells into several groups relevant to diagnosis. The study employs many image processing methods, including binarization, contrast enhancement, noise eradication, morphological operations, labeling, and extraction of some of the blood smear's key properties. In the categorization of the blood cells, which was done in two stages, they were also able to compute several parameters. Blood cells are divided into normal and abnormal cells in the first phase, and the classification of the aberrant cells into other groups of characteristics is the focus of the second phase. Their efforts are directed towards speeding up categorization analysis. They used MATLAB to carry out their study and analysis.
Another research on the application of data mining techniques in illness prediction was conducted by [33]. They used the WEKA analysis tool to examine several categorization algorithms in their study effort to find the optimum end-user functionality for hematological data. Random Forest and neural networks were the machine learning methods that were taken into consideration (multilayer perceptron classier). RF fared exceptionally well in their job, achieving an accuracy of 96.47% with a processing time of 0.16 seconds, a Bayesian Network accuracy of 84.70%, and a Neural Network accuracy of 75.29% with a processing time of 1.92 seconds. In order to find the best, the study conducted two experiments. After the second phase, RF continues to stand with a better accuracy of 86.44%, Bayesian network, 74.57%, and Neural network, 52.54%. They finished their study by affirming that researchers can get informative results from a specific illness dataset and that their findings can help researchers evaluate disease datasets more quickly.
Additionally, [34]'s study is another example of how ML algorithms may be used to predict and categorize useable information. To categorize data of varied amount and diversity, they merged Decision Tree and Neural Networks in their research. The model's structure is a decision tree with

each node representing a neural network that has been trained to categorize a certain output category using binary classification. The Federal Aviation Administration's (FAA) Boeing 737 maintenance dataset, which consists of a sizable number of records composed of 72 variables apiece, was the dataset that they took into consideration. To their advantage, their work is around figuring out the likelihood of incidents, whether or not they happen during scheduled maintenance procedures. Their hybrid was able to accurately and precisely categorize incidence while also being able to pinpoint the most important inputs for classification, enabling better performance and more optimization. Their work also used real-world datasets to show that a hybrid ML algorithm outperformed a single ML method in terms of performance evaluation.

[35] used a modified version of Adaptive Neuro-Fuzzy Inference System (ANFIS) to categorize diabetes diagnoses in Pima Indians datasets. They used this modified version to research the effects of patient features, measurements of anomalies, and decisions regarding the presence or absence of disease on several diseases, such as Pima Indian Diabetes. The effectiveness of ANFIS was assessed using training and testing datasets for Pima Indians, where it evaluates the system's total accuracy, sensitivity, and specificity. Their research's findings demonstrate that modified ANFIS outperforms traditional ANFIS. For training and testing the datasets, they used MATLAB R2014a to accomplish their work. Their research is only confined to datasets related to diabetes, and it has not been tested on datasets related to other diseases.

In terms of individual ML algorithms and hybridization, the work of [31] was the most significant and directly related to RBCs smear study. Their study looked into methods for predicting hypertension patient outcomes based on physical examination parameters. For the sake of analysis, they separated the patients' physical forecasts into two groups. In the first step, key elements from the patients' physical indicators were extracted and used to predict their results. They devised a model that incorporates recursive feature reduction, cross-validation, and classification algorithms. Support Vector Machine (SVM), C4.5 Decision Tree, Extreme Gradient Boosting (XGBoost), and Random Forest were the four classification techniques studied. Various performance evaluation measures were used to evaluate the selected ML algorithms. Despite the fact that three of the four ML algorithms evaluated performed better, XGBoost provided the best performance prediction of hypertension in patients. Their research showed that individual machine learning algorithms can give better prediction results but hybridizing models can produce much better results.

[41–43] are all alternate and related works in terms of hybridization of ML algorithms for feature predictions and classification. Their research shows how several machine learning algorithms can be combined to anticipate the optimum performance evaluation. [30] developed a breast cancer survivorship prediction model using a hybrid model of Random Forest and Logistic Regression methods. They employed the RF algorithm to do preliminary variable screening as well as get important ranks, and then fed the extracted data into the LR algorithm, which is in charge of creating interpretable models for predicting breast cancer survivorship prediction model. They used the RF algorithm to do preliminary variable screening and get important rankings, and then passed the extracted data to the LR algorithm, which was in charge of constructing interpretable models for predicting breast cancer survival. They also tested the hybridization model utilizing basic performance metrics using a 10-fold cross validation. Their results were more accurate even with a basic model.

Another study [34] illustrates how ML approaches may be used to predict and discover useful traits. In their research, they employed a mix of Decision Trees and Neural Networks to categorize data of varying amounts and variety. The model is organized as a decision tree, with each node representing a neural network trained to identify a specific output category using binary classification. They employed the Federal Aviation Administration's (FAA) Boeing 737 maintenance dataset, which includes a large number of records with 72 variables apiece. Their responsibilities include evaluating the likelihood of event occurrences, whether or not they occurred during routine maintenance. Their hybrid was able to describe incidence with great accuracy and precision, as well as identify the most important classification inputs, resulting in improved performance and optimization. Their study looked at real-world data and found that a hybrid strategy outperformed individual machine learning techniques.

Another study based on ML hybridization is that of [36]. Their research is based on a hybrid model that uses XGBoost and Genetic Algorithm to estimate mangrove above-ground biomass, which includes all small and shrub mangrove in patches along Vietnam's northern coast. They compared the XGB-GA model to various machine learning techniques as CatBoost Regression, Gradient Boosting, Support Vector Machine (SVM), and RF. The results reveal that the XGB-GA model outperformed individual ML algorithms in terms of performance measures. As a result, their research showed that ML hybridization provides a more accurate estimate of mangrove above-ground biomass in Northern Vietnam.

To categorize medical diagnoses of Pima Indians with diabetes, [35] used a modified ANFIS. They used this modified version to look at how patient characteristics and measurements of anomalies, as well as a judgment regarding the existence or absence of disease, impact various diseases like Pima Indian Diabetes. The accuracy, sensitivity, and specificity of ANFIS were measured on the Pima Indians datasets during training and testing. The results of their research suggest that modified ANFIS outperforms standard ANFIS. For the training and testing of the datasets, they used MATLAB R2014a. Their work's sole flaw is that it's restricted to diabetes datasets and hasn't been tried on other diseases.

However, [37] was another literature that performed classification of RBCs using Boosting technique. In their work, they performed binary classification of cells and also multiclass classification of abnormal cells (sickled) and their analysis yields an accuracy of 98% while using XGBoost model. From their work, using the individual XGBoost algorithm produced a better performance but when compared

with our model when we introduced RF, we obtained a better performance of 99%. Hence hybridizing the models improves the performance of the analysis.
In addition, [27] was one of our closest literature who worked on erythrocytesIDB datasets and in their work they performed their analysis using ellipse adjustments in peripheral blood smear samples of the erythrocytes and new algorithms for detecting new notable points. Their work did not make use of XGBoost algorithms or hybridizing as part of the measure for calculating performance. They used three image types to validate their method which as artificial images (automatically generated using computer codes), real images (obtained from peripheral blood smear samples) and synthetic images (generated from real isolated cells) Their work yields an accuracy 99%, 98% and 99.35% respectively where 98% accuracy focused on the elongated shape (sickled).
The work of [1] was another useful literature that worked on SCD classification. Their work also makes use of the erythrocytesIDB datasets but in their own case they implement their method using deep learning approach with emphasis on SVM algorithm. Their work addresses the lack of training data using transfer learning techniques. Their work also yields a good performance result with an accuracy of 99.98%. However, their work did not implement hybridization of ML algorithms which can stand as a contrast to our own model using RF-XGBoost hybrid models.
Furthermore, authors in [17] performed numerical investigation of the hemodynamics of RBCs in microvessels with stenosis. In their work, they employed spring model approach for representing deformity of RBC membrane. Their model was able to simulate successfully complex motion and deformity of RBCs in microscale flow manner. In the same vein, [21] also performed classification of data on genetic diseases for microarray analysis. They implemented enhanced backpropagation algorithm as a useful ML algorithm to perform their data classification analysis. Their work shows that enhancing the backpropagation algorithm produces a high classification accuracy with the ability of identifying the most significant gene in classifying genetic diseases from RBCs.
However, current literature on classification of RBCs and comparison of ML algorithms on Erythrocyte smears for predicting SCDs has shown trends of aritificial intelligence involvement (AI) and other state of the art analysis tools of RBCs classification [38]–[42]. The literature demonstrates the comparison between between normal RBCs and sickled cells while also performing the analysis of these RBCs cell morphology.
Therefore, with respect to the literature, this paper has been able to demonstrate that the RF demonstrates a better predicting power when it comes to individual ML algorithms under study while the RF-XGBoost Hybrid model is capable of increasing the rate of predicting the SCD from the erythrocytesIDB database in order to determine the likelihood of SC disease features in human blood samples.

3 Materials and Methods

The purpose of this study was to explore the prediction of SCD in human erythrocytes smears using different machine learning methodologies and hybrid models. Human blood cells are meant to be normal and pure in nature, however their patterns are altered by morphological conditions (such as normal, aberrant, or mutant) as well as external variables like parents' blood genotypes and environments. The most essential concern, however, is establishing a patient's precise prospective Hemoglobin (Hb) condition in order to keep them in a healthy state, long-term viability, and safety. A range of inherited human qualities can be attributed to genes. Human genetics study may help answer concerns about human nature, better understand illnesses, and develop efficient disease therapies, as well as learn about the genetics of human existence [43].
ML algorithms have been used to optimize performance measures and benchmarks in a variety of situations, based on provided example data or prior experience [44]. Different types of classifiers are chosen to be utilized in the classification process in this study based on their capacity to categorize both simulated and real-world medical data (such as Erythrocytes datasets) to see which classifier performs best on test data. The following are the classifiers that were investigated in this paper.

3.1 Multinomial Logistic Regression (MLR)

The statistical approach of LR is used to forecast or predict binary classes. With only two possible output classes, the outcome, output, or goal variable is dichotomous in nature [45]. The chance of an event occurring is calculated using LR. It's one of the most used and fundamental ML methods for two-class categorization. LR is a simple binary classification program that may be used to solve any binary classification issue. In its most basic form, LR accomplishes description and assessments by estimating the connection between a single dependent binary variable and an independent variable. LR is a basic classification approach that belongs to the linear classifier family but is related to polynomial and linear regression in several ways. LR is quick and simple to use, and the results are easy to understand [25, 26].
The linear function f(x)= b0+b1x1+...+ brxn, also known as logit, is used by LR, which is a linear classifier. It use the logit function to anticipate the chance of a binary event occurring, with the log of odds serving as the dependent variable [47]. The sigmoid function is represented in Eq. 1, according to [46], where the e is the base of the natural logarithms (euler's number or the exponential EXP() function) and the x is the actual number to be transformed as shown in Eq. 1.

$$f(x) = \frac{1}{1+e^{-(x)}} \quad (1)$$

When the objective variable contains three or more nominal categories, such as predicting the values of different brands of products, Multinomial Logistic Regression (MLR) is used.

This paper employs the MLR because the predicting values comes as output of three products: normal, abnormal and mutants.

3.2 Extreme Gradient Boosting Algorithm (XGBoost)

DThe eXtreme Gradient Boosting (XGBoost) technique is a type of boosting algorithm that uses gradient boosted decision trees to improve performance and speed. Because of the sequential model training approach, gradient boosting technologies are quite slow to apply. Gradient Boosting is characterized in this scenario as not being highly scalable in terms of implementation. As a result, the XGBoost method aims to overcome this disadvantage by concentrating on computing speed and model performance [30,31].
XGBoost is an ensemble method that uses boosting trees to operate. XGBoost is known as Gradient Boosting because it employs a gradient descent technique. In the following stage, it corrects the model's prior error, learns from it, and increases efficiency. The prior findings have been adjusted, and the consistency of the results has been enhanced. This cycle is continued until there is no more space for improvement. Regularization is the most significant function in this type of prediction algorithm. It's simple to operate and has good accuracy. This technique is commonly used because of its ability to handle missing values and minimize over-fitting. In this type of algorithm, a booster, learning rate, objective, and other hyper-parameters are employed.
The algorithm implementing the XGBoost algorithm is summarized in the following equations according to [30] as:

Step 1: To predict the target variable y, an initial model $F_0(x)$ is specified.
A residual $y - F_0(x)$ will be correlated with this model. Fit a model to the data
Step 2: Fit a model to the residuals (remaining errors),

$$h_1(x) = y - F_0(x) \quad (2)$$

The residual from the previous stage is fitted to a new standard $h_1(x)$.
Ste3: Create a new model,

$$F_1(x) = F_0(x) + h_1(x) \quad (3)$$

$F_0(x)$ and $h_1(x)$ are merged here to produce $F_1(x)$, a boosted version of $F_0(x)$
As a result, $F_1(x)$ mean squared error would be smaller than $F_0(x)$
To boost the efficiency of $F_1(x)$, we might build a new model $F_2(x)$ based on $F_1(x)$ residuals to obtain:

$$F_2(x) = F_1(x) + h_2(x) \quad (4)$$

Therefore, this can be repeated for n-th iteration until the residuals as minimal as possible to obtain

$$F_n(x) = F_{n-1}(x) + h_n(x) \quad (5)$$

Step 4: In this respect, combining weak learner after weak learner, the final model is able to make up for lots of error from the original model and thereby reduces the error over time. The additive learners do not interfere with the functions generated in the previous steps in this process. Instead, they use their own data to predict how many mistakes will occur.

3.3 Random Forest Algorithm (RF)

Random Forest (RF) is a set of classification and regression trees (CART) designed to be trained on datasets of the same size as the training set, known as bootstraps, which are created by random resampling on the training set [50]. In another type, RF is just an ensemble of unpruned classification trees, which is preferred for its greater performance on real-world problems, as well as its lack of sensitivity to input noise and resistance to overfitting [51].
RF is a supervised learning approach used for both classification and prediction regression issues, according to [4, 33–35]. According to their findings, a typical forest consists of a variety of trees, and the more trees in a forest, the more resilient the forest. Hence, the RF algorithm generates decision trees on a data sample, then derives predictions from each of them, and finally selects the best answer through a vote procedure. RF is a more robust ensemble approach than a single decision tree since it avoids overfitting by averaging the results [54].
RF is made up of hundreds of individual decision trees that work together to form an ensemble. Here, each tree in the RF splits off a class prediction, with the class with the most votes being the model's forecast. Each individual tree is created by RF using bagging and feature randomization, leading in an uncorrelated forest of trees whose prediction is more accurate than any other individual tree. The forest's creation is based on the need that the forest's trees and their projections be uncorrelated, as well as the necessity for at least one predictive power [55].
The technique of how the RF algorithm carried out the phases is illustrated in an illustrative mode in Fig. 1, according to [4].
The Gini Index is a non-parametric measure of prediction capacity of variables in regression or classification, based on the idea of impurity reduction. It is likewise non-parametric and hence does not rely on data belonging to a certain type of distribution. The Gini Index of a node $\boldsymbol{n}$ is determined using the method in Eq. 6 for binary splits, according to [56], such as binary data types in erythrocyte datasets.

$$Gini\ (n) = 1 - \sum_{j=1}^{2}(p_j)^2 \quad (6)$$

where $\boldsymbol{p_j}$ is the relative frequency of class $\boldsymbol{j}$ in the node $\boldsymbol{n}$.

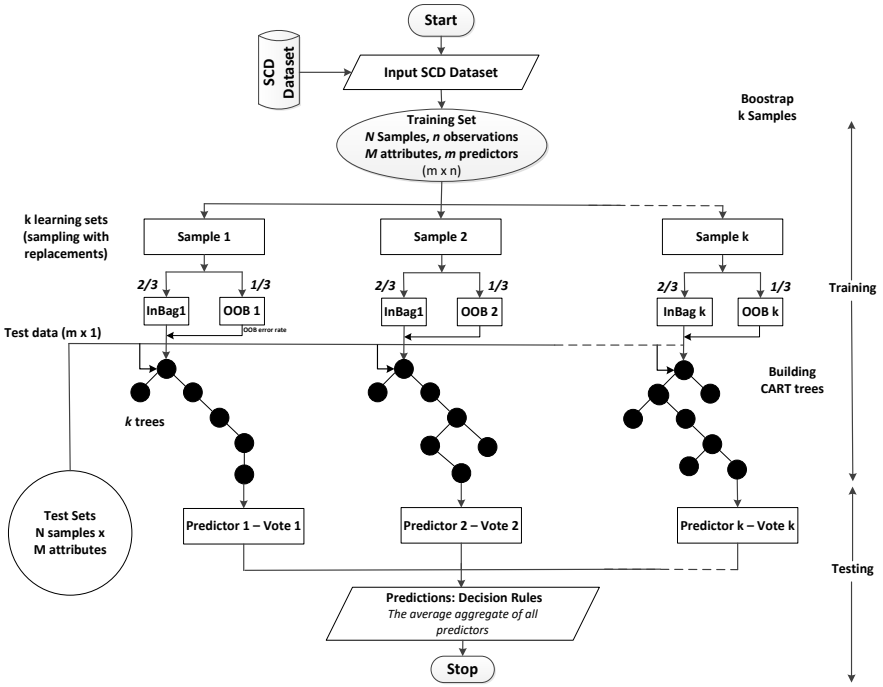

Fig. 1. Random Forest (RF) Algorithm illustrated [4]

3.3 Ensemble Models

In a hybridization process, the RF algorithm, which is a mixture of trees, is utilized to improve on other ML algorithms to gain higher performance. There are various derivable objectives for RF-MLR Classifier model analysis, and these objectives may be satisfied using a variety of parameter estimation techniques. Unordered categorical response variables are modeled using the RF-MLR Classifier's generalized linear modeling strategy. MLR in this case merely functions as a more complex kind of logistic regression. By comparing each category of the unordered answer variable to a random reference category, this method creates a series of logistic regression models. This method generates a set of logistic regression models that reflect specific comparisons between answer categories. The response variable contains j categories, but the RF-MLR Classifier model has j – 1 logic equations. Hence, the algorithm for RF-MLR is given as follows:

Algorithm for RF-MLR Hybrid Model
Aim: Sickle Cell Disease Dataset
Step 1: Start
Step 2: for every attributes in SCD dataset do
Step 3: Construct decision tree
Step 4: for RF (every decision tree constructed) then
Step 5: draw n-tree bootstrap samples for each bootstrap sample, grow un-pruned tree by choosing best split based of random sample of m-try prediction at each node perform features selection randomly from the training sample pool
t: optimize m-try to reduce OOB error
Step 6: Parent node is partitioned into child nodes according to the Gini index using Step (4) and Step (5)
Step 7: if (N number of trees are constructed) then
Step 8: RF classifiers combines the results of all weak classifier to a strong classifier for SCD disease prediction using Step (6) and Step (7)
Step 9: RF-MLR Classifier model is applied for classifying multiple levels of SCD disease as normal, abnormal, and mutant (carriers) using Step (8) and Step (9)
Step 10: end if
Step 11: end for
Step 12: end for
Result: Improved Classification and Prediction Accuracy
Step 13: End

Furthermore, the simpler error checking is in XGBoost, but because of the slower learning rate, more trees must be built. Therefore, RF and XGBoost collaborate to improve performance accuracy. XGBoost is used to support poor learners who are struggling and have strong biases and low variance. Little decision stumps or shallow trees are weak learners (trees with two leaves). The XGBoost would continually change the weights of the training collection based on earlier poorer learners in order to increase the worth

of data that has been incorrectly appraised. The RF algorithms with the XGBoost functions n-rounds, eta, a (alpha), and generate the RF-XGBoost hybrid technique (lambda).
These functions were chosen because they make it possible to keep track of how many times a tree is built, which leads to a more precise tree-based model. When employing bagging procedures in RF, these XGBoost routines are meant to aid with the development of individual trees and raise the weight of the majority vote. Therefore, by lowering the Out-of-bag (OOB) error value and increasing model accuracy, RF-XGBoost hybridization improves the RF model. This is because RF overfits the model because it is unable to calculate the ideal number of trees to utilize. Hence, the algorithm for the RF-XGBoost model is given as follows:

Algorithm for Random Forest-XGBoost Hybrid Model
1: procedure RF-XGBoost (n-tree, m-try, n-rounds, c, eta, α, λ)
2: for each class, $Ci \epsilon$ SCt do
3: Specify the training-control with 5 fold of cross-validation and grid search
4: end for
5: for RF functions do
6: draw n-tree bootstrap samples
7: for each bootstrap sample,
grow un-pruned tree by choosing best split based of random sample of m-try prediction at each node
8: t: optimize m-try to reduce OOB error
9: end for
10: for t, use (SCt), do
11: specify the supporting parameter of model t and n-tree
12: determine best m-try value
13: end for
14: for XGBoost Functions do
15: t: optimize n-rounds, c, eta, λ, α to reduce OOB
16: create new model t1 combine the function from Step 5 and Step 6
17: end for
18: for Function t1 do
19: Predict the result using testing data based on final model t1
20: end for
21: Print the result \
22: return
23: end procedure

In this study, the XGBoost algorithm was chosen to be included into the RF algorithms. XGBoost is well-known for assisting slow learners in becoming powerful learners. The XGBoost technique for hybridizing trees has the advantage of progressively increasing the weight of each branch, which strengthens the trees. It employs a gradient boosting architecture to provide a machine learning algorithm that is high-performing, versatile, and portable. In XGBoost, the lower the learning rate, the easier it is to check for errors, but this means more trees are created. As a result, the use of RF in conjunction with XGBoost improves performance precision. XGBoost is used to help struggling weak learners, who are characterized by large biases and low variance.
Weak learners are like little trees with shallow roots, often as small as mere stumps (trees with two leaves). In order to boost the value of data that has been improperly graded, the XGBoost would continually changing the weights of the training collection based on past poorer learners. The RF-XGBoost hybrid method combines the RF algorithms with XGBoost functions such as n-rounds, eta, (alpha), and (lambda). These functions were chosen because they allow for the tracking of the number of iterations required to build the trees, resulting in a more accurate tree-based model.

3.4 The RBCs smears (erythrocytes) databases samples from the University of the Balearic Islands (UIB) in Spain and the Universidad de Oriente (UOC) in Cuba were combined to create the erythrocytes The Erythrocytes Smears Dataset

dataset utilized in this study for SCD predictions. The *erythrocytesIDB* repository database provided the dataset databases which is available at http://erythrocytesidb.uib.es/. Permission to utilize the datasets with specific instructions was requested, and it was given. The erythrocytesIDB contains three RBC sample databases. The gene combinations in these erythrocytes are normal (A), aberrant (S), and mutant (C) values. The alleles produce genotypes like *AA, AS, AC, SS, SC, CC*, and others, which may be obtained from RBCs.
Erythrocytes images with 624 rows and 6400 columns make up the original data in the erythrocyte databases from UIB and UOC. The number of features is large (6400), thus it was reduced to 624 rows and 7 columns (features) using the traditional scalar technique i.e. Z-score for better performance analysis [58]. The Z-score model was used to reduce the original data in the erythrocytes database using the mathematical formula given in Eq. 7. The model conserves its ranges as maximum and minimum and this introduce the dispersion of the series which are measured in standard deviation or variance. Here, the datasets are converted into a distribution of M (0,1) and comparisons between the series are computed. The values of the models indicates the rate of standard deviations (SDs) of the raw score if it is higher or lower from the mean score. If the Z-score is 0, then it indicates the raw score is on the mean average, if positive (+1), it indicates how many SD the raw score is above the population mean average, while if it is negative, then it indicates that the raw score is not on the mean average [4,40].

$$Z_{score} = \frac{x - \mu}{\sigma} \quad (7)$$

where $\boldsymbol{x}$ is the population raw score (i.e. 6400) $\boldsymbol{\mu}$ is the mean average and $\boldsymbol{\sigma}$ is the SD.
The 624 rows were separated into two groups: 80 percent (499) were used for training, while 20 percent (125) were

used for testing. The samples are gathered from databases and categorized as round (normal), elongated (sickle), and other (mutants). The resulting copies of the original and extracted blood smears are shown in Figs. 2 and 3. The obtained RBCs were used to conduct the analysis in this study. Table 2 shows a small representation of the scaled dataset version utilized for the analysis (training and testing).

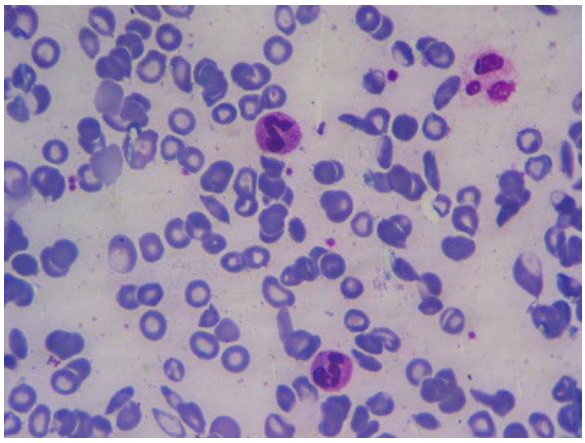

Fig. 2. RBCs (Erythrocytes) Smears

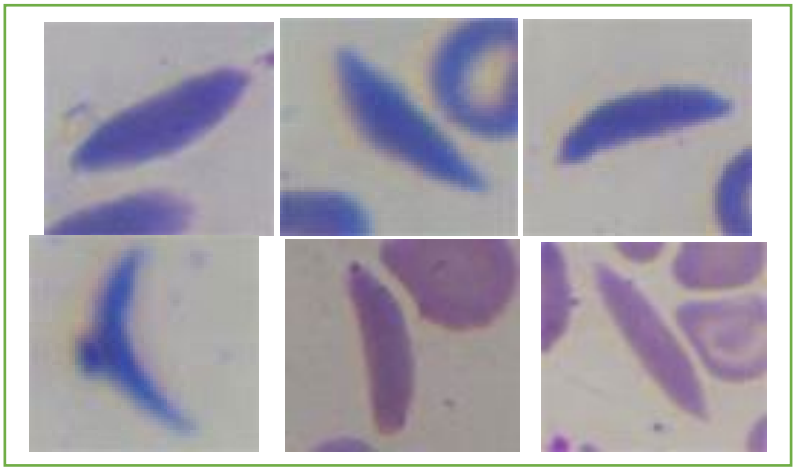

Fig. 3. Elongated or Sickled Erythrocytes

Table 2. ErythrocytesIDB extracted dataset

S/N	AREA	PERMR	APPR_EDGE	CONV_HULL	A_CIRC	A0_ELLIPSE	SHAPE
1	2436.5	185.5807346	4	31	2642.079422	2434.821561	0
2	2760	196.9949472	4	33	2827.433388	2762.546111	0
3	2116	173.1959589	4	27	2290.221044	2145.12767	0
4	1815.5	161.7817446	4	29	1963.495408	1811.94079	0
5	1571.5	148.953317	4	30	1661.902514	1572.698325	0

S/N	AREA	PERMR	APPR_EDGE	CONV_HULL	A_CIRC	A0_ELLIPSE	SHAPE
6	1223	131.3969687	4	24	1256.637061	1236.02099	0
7	1418.5	141.2964631	4	27	1520.530844	1422.48802	0
8	2124.5	173.4385991	4	27	2463.00864	2142.380166	0
…	…	…	…	…	…	…	…
202	2612	192.1665	4	33	2827.433	2601.9	0
203	1351	159.196	3	24	3019.071	1399.118	1
204	1392.5	165.7817	3	20	2827.433	1513.42	1
205	1449	170.5097	3	23	3631.681	1475.27	1
206	1335.5	141.3553	4	16	1809.557	1694.11	1
207	724.5	120.468	2	18	1963.495	722.2212	1
208	1392	159.196	3	27	2827.433	1403.319	1
209	1099	138.0833	3	20	2290.221	1145.39	1
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
620	1437.5	155.4385984	3	23	2642.079422	1532.408004	2
621	1209.5	157.1543279	3	18	2827.433388	1258.859689	2
622	1329	146.8528122	4	23	1809.557368	1433.288255	2
623	1164.5	148.2670263	3	19	2463.00864	1201.96302	2
624	1837.5	173.4385984	3	23	3019.07054	1973.704891	2

Normalization, which is part of the pre-processing stage, is mapping or scaling techniques which is used to predict closer to larger variation. There are numerous techniques of data normalization which ranges from Min-Max normalization, Decimal Scaling normalization, Integer Scaling normalization and Z-score normalization. This study as part of its pre-processing stage employed Z-score algorithm. This method preserves ranges (maximum and minimum) and introduce the dispersion of the series (standard deviation / variance). In this method, the datasets are converted into a M(0,1) distribution and the comparison between series were computed. The Z-score value indicates how many standard deviations the raw score is higher (away from) or lesser than the mean average. If the Z-score is 0, it indicates that the raw score is on the mean average; if it is positive (such as +1, it indicates how many standard deviations the raw score is above the population mean average; and if it is negative, it indicates that the raw score is not on the mean average. Mathematically, the Z-score is represented as shown in Eq. 7. ML algorithms such as the selected ones in this paper performs better when values in the dataset lies between -1 and +1. With respect to our dataset which has very large values, scaling method using the Z-score algorithms reduces the dataset to the desired range of desired.

The training datasets were created from 624 photographs of individual RBCs and were divided into three categories: circular – normal (202), elongated – sickle (210), and others (212). To produce a satisfactory output for the analysis, the individual RBCs are treated to several processes.

Among the processes on RBC photos are grayscale conversion to Gaussian blurring, adaptive threshold, identifying contours to detect erythrocytes position in picture, and filtering contours to get the exact one that is the normal RBC and those that are not. During the contour extraction procedure, each contour feature was assessed to be sufficient for the selected ML algorithms to estimate the form of the RBCs. As a consequence, they were chosen above other pixel values that would have resulted in 6400 features. As a consequence, as stated in Table 2, the following attributes were chosen: *Area, Perimeter, Approximate Edge, Convex Hull, Area Circle, Area Eclipse,* and the erythrocyte *Shape* which are measure in meters and square meters (m^2)

3.5 Performance Models

Accuracy, Sensitivity (Recall), Specificity, F_1 Score and Area under Curve ROC are the most typical metrics of system performance. As a result, these metrics are used to summarize the findings of the ensemble models of ML algorithms that have been evaluated in this paper. According to [31] as reported by [4], these metrics are defined with respect to Eq. 8, 9, 10 and 11. Very positives, false positives, false negatives, and very negatives are abbreviated as VP, FP, FN, and VN.

$$Acc = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} Z(x) \quad (8)$$

where $\boldsymbol{Z(x)}$ is the indicator function where $\boldsymbol{x} = (\boldsymbol{c_i} - \boldsymbol{\hat{c}_i})$. If $\boldsymbol{x}$ holds, then $\boldsymbol{Z(x)} = \boldsymbol{1}$ if otherwise $\boldsymbol{Z(x)} = \boldsymbol{0}$.

$$Sensitivity\ (Recall) = \frac{VP}{VP+FN} \quad (9)$$

$$Specificity = \frac{VN}{VN+FP} \quad (10)$$

APrecision close to one indicates that all the recognized erythrocytes states are in fact correctly predicted.
A recall close to one indicates that all the eyrthrocytes states, in a given sample were correctly recognized by the proposed model.

$$F_1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (11)$$

For Area under curve ROC, it has the following submissions

1. If $AUC = 1$, the classifier is perfect, and there is at least one threshold in this predictive model that produces a perfect prediction.
2. If the AUC is 0.5 or more (that is $0.5 < AUC < 1$), the classifier outperforms random guessing. It is possible for the classifier to have predictive value if the threshold is appropriately set.
3. If the $AUC = 0.5$, the classifier is based on a random guess and so has no predictive value.
4. However, the classifier is poorer than a random guess if the AUC is less than 0.5.

4 Experimental Results and Analysis

Researchers have advised the use of strong and accurate prediction approaches to foresee the effects of independent factors on dependent parameters since empirical equations are insufficient for modeling the links between human blood diseases and other variables. Indeed, these trustworthy individual and hybrid predictive models might help with medical data analysis. As a result, the erythrocyte dataset was used in this study to evaluate a number of ML algorithms and hybrids in order to enhance the detection of SCD in humans. The results of each of the selected ML algorithms' study on erythrocyte datasets are highlighted in the next section, along with comparisons to other hybrids of ML algorithms utilizing Python programming language, and other performance measurement metrics. The outcomes of a number of ML algorithms are assessed.
For this work, the Python programming language was utilized to acquire various values for the retrieved characteristics, as well as other codes, with regard to circular (normal), elongated (sickle), and other (mutants). Fig. 4 depicts an example of the operation procedure on erythrocyte smears, with the various attributes retrieved that were employed in the analysis. Individual erythrocyte pictures, on the other hand, were processed, and the retrieved images were rendered in grayscale format for the training and testing processes, respectively. The retrieved grayscale forms of erythrocyte smears and individual morphologies are gathered, with round (normal) = 0, elongated (sickle) = 1, and others (mutants) = 2 as illustrated in the micro format in Table 2.

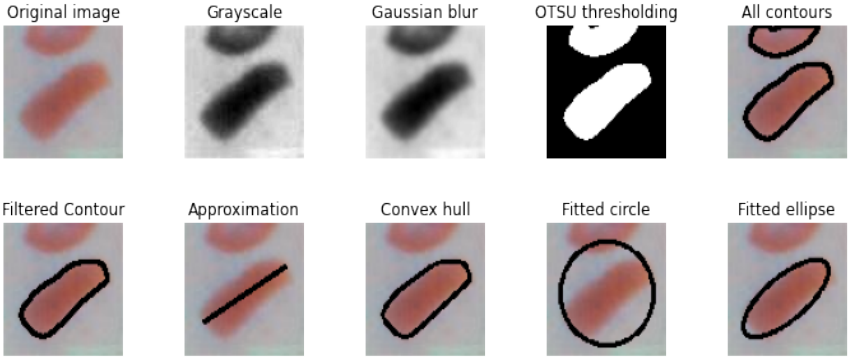

Fig. 4. Operation Process on Erythrocytes Images

4.1 MLR Algorithm Performance

When fitting an MLR Model without intercept on a dataset with constant nonzero columns, the Python programming language outputs zero coefficients for continuous nonzero columns. The area under the ROC curve is a common metric in MLR. The Binary Classification Evaluator may be used to calculate the AUC.

A Precision-Recall curve shows (precision, recall) points for various threshold values, whereas a Receiver Operating Characteristic (ROC) curve shows (precision, recall) points for various threshold values (recall, false positive rate) as depicted in Fig. 6 and the corresponding heatmaps. The model predicts better when the area under the ROC curve is near to one.

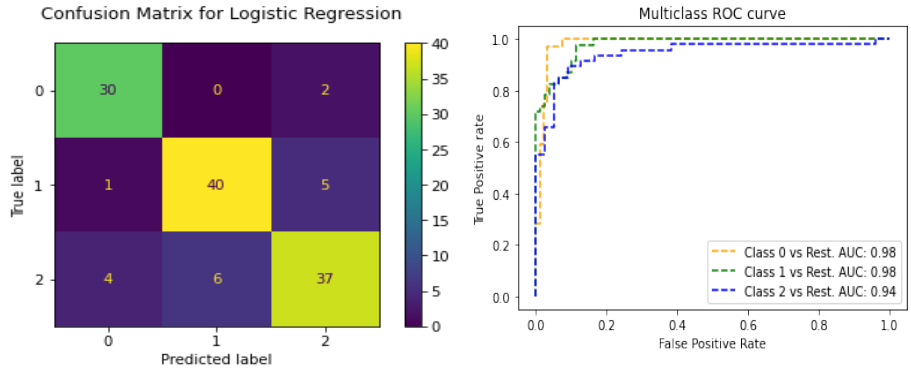

Fig. 5. Confusion Matrix and ROC-AUC curve for MLR classifier

4.2 XGBoost Algorithm Performance

XGBoost is well-known for assisting slow learners in becoming powerful learners. The XGBoost technique for hybridizing trees has the advantage of progressively increasing the weight of each branch, which strengthens the trees. It applies a machine learning technique with great performance, adaptability, and portability using a gradient boosting architecture. Due to the process of progressively analyzing datasets, the Gradient Descent Boosting technique tends to compute the model's output at a slower rate, XGBoost is supplied to boost or greatly boost the model's speed and performance.

Hence, Fig. 6 displays XGBoost's performance on the SCD dataset in terms of several performance measures. It is found that XGBoost has a higher FI score, indicating that the dataset's accuracy and recall provide a higher and extremely positive value. XGBoost also has an AUC of 0.98, indicating that it was able to enhance the performance of weak learners to classify the SCD dataset in 98 percent of cases. Similarly, the yellow distribution curve represents the positive class (circular or normal) RBCs, the green distribution curve represents the negative class (elongated or sickle) RBCs, and the blue distribution curve represents other unknown RBCs. The output of the analysis are demonstrated in Fig. 6 demonstrates ROC-AUC graph and the corresponding heatmaps respectively. As a result, the XGBoost model passes the erythrocytesIDB datasets prediction test. The ROC-AUC curve for the XGBoost model, which shows an AUC value of 99 percent for the elongated (sickle) class.

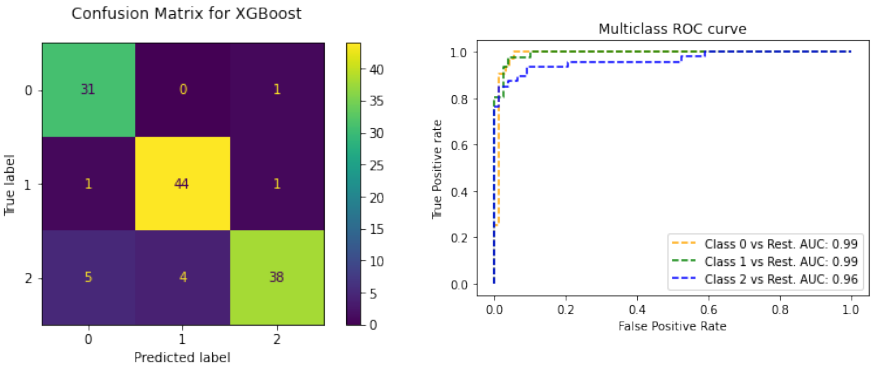

Fig. 6. Confusion Matrix and ROC-AUC curve for XGBoost classifier

4.3 Random Forest Algorithm

The RF model, which is used in conjunction with other machine learning approaches to boost performance, is another decision tree methodology used in this study. As previously explained, the "randomForest" package was used to implement this strategy in the Python programming language using the sklearn package. Fig. 7 presents the performance measures of RF on erythrocytesIDB dataset for SCD dataset with respect to the performance metrics.

As a consequence, the RF model passes the test of predicting medical SCD datasets from erythrocyteIDB datasets by offering the best and highest individual ML performance results.

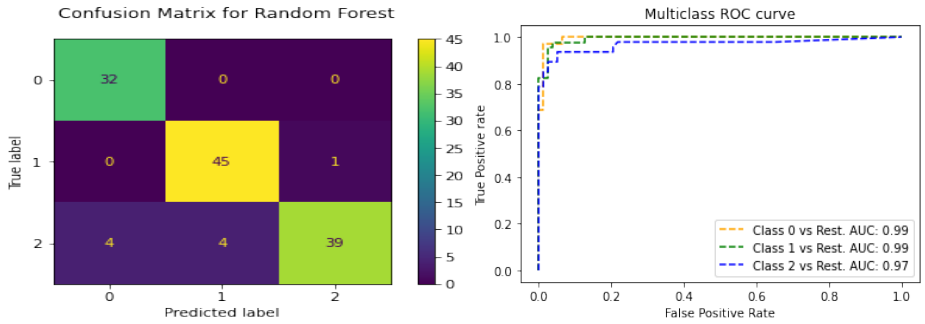

Fig. 7. Confusion Matrix and ROC-AUC curve for RF Model

Fig. 7 displays the ROC-AUC curve for the RF model, which demonstrates that the elongated (sickle) class has an AUC of 99 percent and the Confusion Matrix respectively. The RF model also has a 99 percent ROC-AUC value for predicting SCD dataset, which passes the AUC 0.9 point on a 0.5 threshold level of significance. As a result, the RF model present a better individual ML prediction test for SCD using the erythrocyteIDB datasets.Similarly, from each of the ML algorithms under study, the yellow distribution curve represents the positive class (circular or normal) RBCs, the green distribution curve represents the negative class (elongated or sickle) RBCs, and the blue distribution curve represents other unknown RBCs.

4.4 The Performance of the proposed ensemble classifiers

These functions (as shown in section 3.4) were chosen because they make it possible to keep track of how many times a tree is built, which leads to a more precise tree-based model. When employing bagging procedures in RF, these XGBoost routines are meant to aid with the development of individual trees and raise the weight of the majority vote. Therefore, by lowering the Out-of-bag (OOB) error value and increasing model accuracy, RF-XGBoost hybridization improves the RF model. This is because RF overfits the model because it is unable to calculate the ideal number of trees to utilize.

0.These XGBoost functions boost the majority vote value in RF while also contributing in the creation of individual trees in the RF bagging techniques procedure. The RF-XGBoost hybrid models presents an accuracy of 99% and an improvement over individual ML algorithms (XGBoost = 90% and RF = 93%). As a consequence, RF-XGBoost hybridization helps to increase the RF model's accuracy by decreasing the Out-of-bag (OOB) error number. This is due to RF's failure to calculate the ideal amount of trees, causing the model to overfit. Table 3 illustrates the results of the prediction using individual algorithms and their hybrid models. The performance output of the hybridization of the RF and MLR Hybrid models is shown in Table 3. In comparison to the individual MLR model, the output of the hybridization of RF-MLR provides a modest decline in accuracy in terms of performance metrics. The accuracy of the RF-MLR model is 92 percent, compared to 93 percent for individual RF. When compared to its combination with the RF model, MLR has a decline performance in terms of the performance metrics. Hence, MLR model (which has an accuracy of 87%) improves even when extra trees are added to mitigate Out-of-bag (OOB) errors that may arise from categorizing the erythrocyteIDB dataset.

Table 3. Performance Measures for Individual ML Algorithm and their Hybrid Models

	Performance Metrics (%)				
ML algorithms & Hybrids Models	Accuracy %	Sensitivity %	Specificity %	F1 Score %	ROC-AUC %
LR(MLR)	87	87	86	86	97
XGBoost	90	97	89	90	98
RF	93	97	\91	93	99

RF-MLR	92	97	91	92	99
RF-XGBoost	**99**	**99**	**96**	**97**	**99**

As discussed in earlier sections, the performance of the RF-XGBoost Hybrid model is compared to that of individual and other ML method classifiers. Table 3 illustrates the results of the RF-XGBoost hybrid model's as against other individual and hybrid models prediction based on their accuracy, sensitivity, specificity, and AUC. Fig. 8 shows a bar graph of all the ML algorithms, hybrid models and their performances.

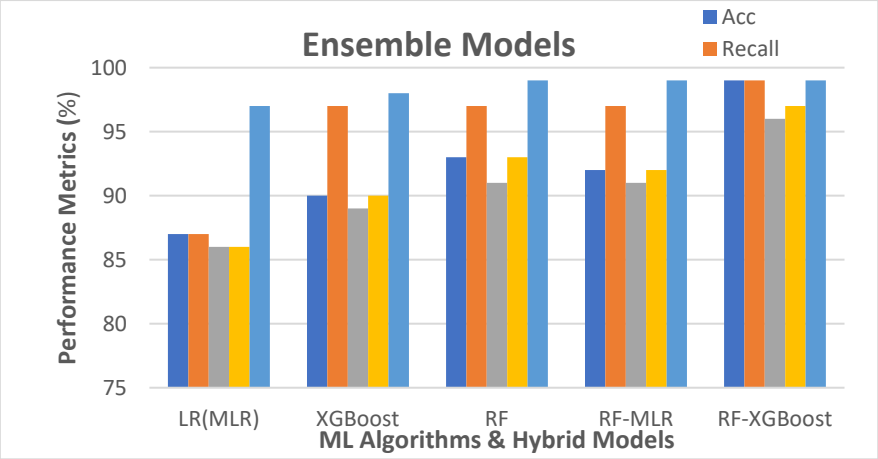

Fig. 8. Graph showing the ML Algorithms and the performance metrics

Based on the performance measurements, it can be seen that the RF-XGBoost Hybrid classifier outperformed other ML method classifiers in terms of predicting SCD from erythrocytesIDB datasets with an accuracy of 99%.

5 Discussion

Future researchers working on SCD prediction can use the findings of this study to develop more effective ways for analyzing aberrant RBCs. With exceptional forecasting skills, the proposed hybridized ML prediction approaches were able to correctly detect distinct blood problems samples contained inside the erythrocytesIDB dataset. More hybrid techniques to sickle detection might be studied in the future. The findings of this study, on the other hand, benefit healthcare providers, hematologists, data scientists, bioinformaticists, geneticists, and other medical data researchers. As a result, future researchers in this field may need to depend on Big Data and AI to deal with the computational challenges that come from sickle cell case pre-processing and prediction.

Due to the complexity of the data, selecting significant genes and accurately classifying alleles for improved performance has proven challenging in the investigation of SCD. As a result, several statistical and machine learning techniques are employed. A condensed version of the dataset was utilized for the testing due to the size of the dataset and the intensive processing needed to manage the creation of the hybrid ML models. The accessibility of regional material is another issue. With involvement of local content, the analyses might produce a better result.

The hybrid RF-XGBoost algorithm model in this work has room for improvement. Both binary and multiclass classification issues are addressed by the approach. Although the binary classification challenge in medical data research is the most frequent, this approach may also be enhanced for multiclass classification issues. Interested in is also survival analysis. A meaningful biological link between gene expression levels and the survival window or condition of certain SCD patients may be demonstrated by using the relevant parameters selected by the model as predictors in survival models. Improved machine learning methods that can help with the prediction and categorization of red blood cell abnormalities will be the focus of future research. Future work to this research include incorporating more sophisticated hybrid ML models that make use of Big Data and AI technology with emphasis on local content. Future research may concentrate on producing more output from a greater variety of datasets as the built models in this study only use two datasets.

6 Limitation and Implication

Due to the complexity of the data, selecting significant genes and accurately classifying alleles for improved performance has proven challenging in the investigation of SCD. As a result, several statistical and machine learning techniques are employed. A condensed version of the dataset was utilized for the testing due to the size of the dataset and the intensive processing needed to manage the creation of the hybrid ML models. The accessibility of regional material is another issue. With involvement of local content, the analyses might produce a better result.

7 Conclusion

The strategies used in this study improved the status of disorder prediction in a variety of ways. This study presents the various evolutions and dimensions of blood abnormalities. The study also looked at literature that offered different methods for predicting sickle cell disease. Following that, better prediction models for detecting these abnormalities were created using a methodological framework based on ensemble ML approaches. These models were trained, tested, and evaluated using erythrocytesIDB datasets. After understanding the complicated pattern in the dataset through exploratory analysis, the study emphasized the importance of using ensemble methodologies to forecast sickle cell disease. As such, RF-XGBoost hybrid model demonstrated to outperform both individual ML algorithms and RF-MLR hybrid model with a performance efficiency of 99%. Therefore, hybridizing ML algorithm has shown to offer better performance and also increase the throughput time of analysis.

Acknowledgements

Dr. Antoni Jaume-i-Capo of the Computer Graphics, Vision and Artificial Intelligence Group (UGIVIA), Department of Mathematics and Computer Science, University of the Balearic Islands, 07122 Palma de Mallorca, Spain – UIB, and supported by Universidad-de-Oriente, Cuba – UOC, were instrumental in making the RBCs erythrocytesIDB datasets available for this work and assisting with other research analyses. Dr. Bamgbose, a hematologist and ED specialist at Ekiti State University Teaching Hospital (EKSUTH), Ado Ekiti, Ekiti State, Nigeria, provided medical support under the authors' instruction and supervision. The authors received no support from EKSUTH.
Similarly, the Association of Commonwealth Universities (ACU) is thanked for organizing and sponsoring the principal researcher's participation in several Commonwealth Future Series workshops at the University of Mauritius, Reduit Campus, in 2019; and O. P. Jindal Global University, Sonipat, India, in 2020. The Carnegie Corporation of New York, in collaboration with the University of Ghana, Legon on the Pan-African Doctoral Academy, organized and funded the principal authors' candidacy at the University of Ghana, Legon six times between 2016 and 2019.
.

References

[1] L. Alzubaidi, M. A. Fadhel, O. Al-shamma, and J. Zhang, "Deep Learning Models for Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anemia Diagnosis," *Electron. MDPI*, vol. 9, no. 427, pp. 1–18, 2020.

[2] P. K. Das, S. Meher, R. Panda, and A. Abraham, "A Review of Automated Methods for the Detection of Sickle Cell Disease," *IEEE Rev. Biomed. Eng.*, vol. 13, pp. 309–324, 2020, doi: 10.1109/RBME.2019.2917780.

[3] P. L. Stephenson, M. V. Taylor, and C. Anglin, "Sickle Cell Disease," *J. Consum. Health Internet*, vol. 19, no. 2, pp. 122–131, 2015, doi: 10.1080/15398285.2015.1026706.

[4] M. W. Darlison and B. Modell, "Sickle-cell disorders: limits of descriptive epidemiology.," *Lancet (London, England)*, vol. 381, no. 9861, pp. 98–9, Jan. 2013, doi: 10.1016/S0140-6736(12)61817-0.

[5] J. Kanter and R. Kruse-Jarres, "Management of sickle cell disease from childhood through adulthood.," *Blood Rev.*, vol. 27, no. 6, pp. 279–87, Nov. 2013, doi: 10.1016/j.blre.2013.09.001.

[6] L. Alzubaidi, O. Al-Shamma, M. A. Fadhel, L. Farhan, and J. Zhang, "Classification of red blood cells in sickle cell anemia using deep convolutional neural network," *Adv. Intell. Syst. Comput. - Springer*, vol. 940, pp. 550–559, 2020, doi: 10.1007/978-3-030-16657-1_51.

[7] C. Grosan and A. Abraham, *Machine Learning*, vol. 17. 2011. doi: 10.1007/978-3-642-21004-4_10.

[8] S. W. Knox, "Machine Learning - A Coincise Introduction," *Wiley Ser. Probab. Stat.*, pp. 1–320, 2018.

[9] G. Roth, "Machine learning with Python: An introduction," *JavaWorld*, pp. 1–5, 2019, [Online]. Available: https://www.javaworld.com/article/3322898/application-development/machine-learning-with-python-an-introduction.html

[10] O. B. Ayoade, "Comparative Analysis of Selected Machine Learning Algorithms for predicting Sickle Cell Disease," *Depatertment Comput. Sci. Fac. Commun. Inf. Sci. Univ. Ilorin, Kwara State, Niger.*, vol. December, pp. 1–270, 2021.

[11] N. I. of H. NIH, "Health Information for the Public - Sickle Cell Disease (SCD)," *National Heart Lung and Blood Institute*, 2016.

[12] N. I. of H. NIH, "The Management of Sickle Cell Disease," *Natl. Hear. Lung Blood Inst.*, no. 02–2117, pp. 1–206, 2015, [Online]. Available: http://www.nhlbi.nih.gov

[13] S. E. Roger and H. R. Rodney, "Some Medical and Social Aspects of the Treatment for Genetic-Metabolic Diseases," *Ann. Am. Acad. Polit. Soc. Sci.*, vol. 399, pp. 30–37, 2017.

[14] M. Saad and Z. Salem, "Basic concepts of medical genetics , formal genetics," *Egypt. J. Med. Hum. Genet.*, vol. 15, no. 1, pp. 99–101, 2014, doi: 10.1016/j.ejmhg.2013.10.001.

[15] L. M. Gunder and S. A. Martin, *Essentials of Medical Genetics for Health Professionals*. USA: Jones & Bartlett Learning, LLC, 2011.

[16] World-Health-Organization, "Sickle-Cell Anaemia," *World*

Heal. Organ., vol. 11, no. April, pp. 1–5, 2020.
[17] X. Jiang, T. Wang, and Z. Xing, "Simulation Study of Hemodynamics of Red Blood Cells in Stenotic Microvessels," *Adv. Mater. Res. - Biomater. Bioeng.*, vol. 647, pp. 321–324, 2013, doi: 10.4028/www.scientific.net/AMR.647.321.
[18] J. R. Frost *et al.*, "Improving Sickle Cell Transitions of Care Through Health Information Technology.," *Am. J. Prev. Med.*, vol. 51, no. 1 Suppl 1, pp. S17-23, Jul. 2016, doi: 10.1016/j.amepre.2016.02.004.
[19] C. P. Rivera, A. Veneziani, R. E. Ware, and M. O. Platt, "Sickle cell anemia and pediatric strokes : Computational fluid dynamics analysis in the middle cerebral artery," *Exp. Biol. Med.*, vol. 241, pp. 755–765, 2016, doi: 10.1177/1535370216636722.
[20] S. D. Grosse, I. Odame, H. K. Atrash, D. D. Amendah, F. B. Piel, and T. N. Williams, "Sickle cell disease in Africa: A neglected cause of early childhood mortality," *Am. J. Prev. Med.*, vol. 41, no. 6 SUPPL.4, pp. S398–S405, 2011, doi: 10.1016/j.amepre.2011.09.013.
[21] B. Nisha, B. Madasamy, and J. J. Tamilselvi, "Enhanced Backpropagation Approach for Identifying Genetic Disease," *Appl. Mech. Mater.*, vol. 622, pp. 75–80, 2014, doi: 10.4028/www.scientific.net/AMM.622.75.
[22] O. S. Platt *et al.*, "Mortality in Sickle Cell Disease-Life Expentancy & Risk Factors," *N. Engl. J. Med.*, vol. 330, no. 23, pp. 1639–1644, 2012.
[23] D. Divya, K. N. Rao, Si. G. Ratnam, and D. Sowjanya, "Supervised Machine Learning Algorithms for Analysis on Sickle Cell Anemia," *High Technol. Lett.*, vol. 26, no. 11, pp. 994–1004, 2020.
[24] T. M. Sabu, "Bioinformatics," *Fundam. Concepts Bioinforma.*, pp. 1–155, 2003.
[25] A. D. Hardie, L. Ramos-Duran, and J. U. Schoepf, "Cardiac MR assessment of myocardial iron deposition in sickle cell disease : risk factors and association with cardiac function," *J. Cardiovasc. Magn. Reson.*, vol. 1, pp. 48–48, 2010, doi: 10.1186/1532-429X-12-S1-P274.
[26] G. D. Magoulas and A. Prentza, "Machine Learning in Medical Applications," *Springer*, vol. 204, no. 9, pp. 300–307, 2015, doi: 10.1007/3-540-44673-7.
[27] G.-H. Manuel, F. A. Guerrero-Peña, S. Herold-García, A. Jaume-I-Capó, and P. D. Marrero-Fernández, "Red Blood Cell Cluster Separation From Digital Images for Use in Sickle Cell Disease," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1514–1525, 2015, doi: 10.1109/JBHI.2014.2356402.
[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[29] Y. Zhang, S. Wang, and G. Ji, "A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications," vol. 2015, 2015.
[30] R. Rajbharath and L. Sankari, "Predicting Breast Cancer using Random Forest and Logistic Regression," *Int'l J. Eng. Sci. Comput.*, vol. 7, no. 4, pp. 10708–10713, 2017.
[31] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, and S. Zhang, "A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data," *Diagnosticsc - MDPI*, vol. 9, no. 178, pp. 1–21, 2019.
[32] N. Safca, D. Popescu, and L. Ichim, "Image Processing Techniques to Identify Red Blood Cells," in *International Conference on System Theory, Control and Computing*, 2018, pp. 93–98.
[33] F. Akter, A. Hossin, G. M. Daiyan, and M. Hossain, "Classification of Hematological Data Using Data Mining Technique to Predict Diseases," *J. Comput. Commun.*, vol. 6, pp. 76–83, 2018, doi: 10.4236/jcc.2018.64007.
[34] J. Carson, K. Ollingsworth, R. Datta, G. Clark, and A. Segev, "A Hybrid Decision Tree-Neural Network (DT-NN) Model for Large-Scale Classification Problems," *Univ. South Alabama*, vol. 2, no. 11, pp. 1–9, 2018.
[35] A. M. Sagir and S. Sathasivam, "Design of a modified adaptive neuro fuzzy inference system classifier for medical diagnosis of Pima Indians Diabetes," in *AIP Conf Proc.*, 2017, vol. 1, pp. 1–7. doi: 10.1063/1.4995880.
[36] T. D. Pham, N. Yokoya, J. Xia, N. T. Ha, and N. N. Le, "Comparison of Machine Learning Methods for Estimating Mangrove Above-Ground Biomass Using Multiple Source Remote Sensing Data in the Red River Delta Biosphere Reserve, Vietnam," *Remore Sens. - MDPI*, vol. 12, no. 1334, pp. 1–24, 2020.
[37] D. Uike and S. Thorat, "Computerization Method to classifying of Red Blood Cells using Boosting Technique," *Int'l J. Eng. Researcg Technol.*, vol. 9, no. 06, pp. 1572–1577, 2020.
[38] P. E. M. D. Ouglass, T. I. O. C. Onnor, and B. A. J. Avidi, "Automated sickle cell disease identification in human red blood cells using a lensless single random phase encoding biosensor and convolutional neural networks," *Opt. Express*, vol. 30, no. 20, pp. 35965–35977, 2022.
[39] M. Darrin *et al.*, "Classification of red cell dynamics with convolutional and recurrent neural networks : a sickle cell disease case study," *Sci. Rep.*, vol. 13, no. 745, pp. 1–12, 2023, doi: 10.1038/s41598-023-27718-w.
[40] A. Sada, M. Bordukova, A. Makhro, N. Navab, A. Bogdanova, and C. Marr, "RedTell : an AI tool for interpretable analysis of red blood cell morphology," *Front. Physiol.*, vol. 14:1058720, pp. 1–16, 2023, doi: 10.3389/fphys.2023.1058720.
[41] H. B. R. Alabed *et al.*, "Comparison between Sickle Cell Disease Patients and Healthy Donors : Untargeted Lipidomic Study of Erythrocytes," *Int. J. Mol. Sci.*, vol. 24, no. 2529, pp. 1–15, 2023.
[42] Y. Qiang, A. Sissoko, Z. L. Liu, T. Dong, and F. Zheng, "Microfluidic study of retention and elimination of abnormal red blood cells by human spleen with implications for sickle cell disease," *PNAS - Eng. Cell Biol.*, vol. 120, no. 6, pp. 1–12, 2023, doi: 10.1073/pnas.
[43] D. J. Weatherall *et al.*, "Global epidemiology of sickle haemoglobin in neonates : a contemporary geostatistical model-based map and population estimates," *The Lancet (London, England)*, vol. 381, no. 9861, pp. 142–151, 2013, doi: 10.1016/S0140-6736(12)61229-X.
[44] M. Zhang, X. Li, M. Xu, and Q. Li, "Image Segmentation and Classification for Sickle Cell Disease using Deformable U-Net," *Springer*, vol. 10, pp. 1–10, 2017.
[45] A. Navlani, "Understanding Logistic Regression in Python," *Mach. Learn.*, vol. 3, pp. 1–11, 2019.
[46] M. Stojiljkovic, "Logistic Regression in Python," *J. Data Sci.*, vol. 2507, no. 1, pp. 1–9, 2020.
[47] Jason Brownlee, "Logistic Regression for Machine Learning," *Machinelearningmastery.Com*, 2019.
[48] Z. Zixuan, "Boosting Algorithm Explained," *Theory, Implement. Vis.*, vol. 7, pp. 1–12, 2019.
[49] L. Zulalkha, "A Comprehensive Guide To Boosting Machine Learning Algorithms," *Edureka Res. Anal. J.*, vol. 3, no. 12, pp. 1–7, 2020.
[50] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[51] P. R. Patil and S. A. Kinariwala, "Automated Diagnosis of

Heart Disease using Random Forest Algorithm," *Int. J. Adv. Res. Ideas Innov. Technol.*, vol. 3, no. 2, pp. 579–589, 2017.
[52] F. Alam and S. Pachauri, "Usage of Data Mining Techniques for combating cyber security," *Int'l J. Eng. Comput. Sci.*, vol. 6, no. 1, pp. 20011–20016, 2017, doi: 10.18535/ijecs/v6i1.31.
[53] J. De Boer, "Applying machine learning methods for predicting 120-day hospital readmission by utilizing medical administrative patient data," *Tilbg. Unversity Res.*, vol. 6, pp. 1–35, 2019.
[54] B. Bradley and G. Brandon, "Classification Algorithms - Decision Tree," *Sch. Informatics*, vol. 1, pp. 1–6, 2020.
[55] T. Yiu, "Understanding Random Forest How the Algorithm Works and Why it Is So Effective," *Mach. Learn. Appl. An Int. J.*, vol. 6, pp. 1–9, 2019.
[56] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 6, pp. 551–560, 2013.
[57] B. Mustain and I. Nazrul, "An Early Diagnosis System for predicting Lung Cancer Risk Using _adaptive Neuro Fuzzy Inference System and Linear Discriminant Analysis," *J. MPE Mol. Pathol. Epidemiol.*, vol. 1, no. 1, pp. 1–4, 2016, [Online]. Available: http://molecular-pathological-epidemiology.imedpub.com/an-early-diagnosis-system-for-predicting-lung-cancer-risk-using-adaptive-neuro-fuzzy-inference-system-and-linear-discriminant-anal.php?aid=11320
[58] B. Bryan, "Bioinformatics Computing," *Prentice Hall - Pearson Educ. Inc.*, vol. 1st Editio, pp. 1–395, 2002.