Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523(Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Application of AI Algorithms for the Prediction of the Likelihood of Sickle Cell Crises

Essang Samuel Okon[1*], Kolawole Olamide Michael[1], Runyi Emmanuel Francis[2], Ante Jackson Efiong[3*], Ogar-Abang Micheal Obi[1], Auta Jonathan Timothy[4], Okon Paul Edet[5], Effiong Raphael Dominic[6], Ukim Akanimo Jimmy[5]

[1]Department of Mathematics and Computer Science, Arthur Jarvis University, Akpabuyo, Nigeria
[2]Department of Statistics, Federal Polytechnic, Ugep, Nigeria
[3]Department of Mathematics, Topfaith University, Mkpatak, Nigeria
[4]Department of Pure and Applied Mathematics, African University of Science and Technology, Abuja, Nigeria
[5]Department of Electrical/Electronics, Topfaith University, Mkpatak, Nigeria
[6]Department of Mathematics, University of Calabar, Calabar, Nigeria

DOI: https://doi.org/10.36347/sjet.2024.v12i12.008 | **Received:**09.11.2024 | **Accepted:**16.12.2024 | **Published:** 26.12.2024

***Corresponding authors:** Essang Samuel Okon; Ante Jackson Efiong
Department of Mathematics and Computer Science, Arthur Jarvis University, Akpabuyo, Nigeria; Department of Mathematics, Topfaith University, Mkpatak, Nigeria

Abstract **Original Research Article**

This paper investigates the application of Support Vector Machines (SVM), Random forest and Logistic Regression as AI algorithms to predict the likelihood of sickle cell crises. The study focused on the various considerations involved in utilizing SVMs, Random forest and Logistic Regression for this predictive task, encompassing feature selection, data quality assessment, handling class imbalance, model training, hyper parameter optimization, evaluation metrics, and interpretability. Special attention is given to addressing the complexities of Sickle Cell Disease data and ensuring the reliability of predictions. Through a comprehensive analysis, we accentuate the relevance of SVMs, Random forest and Logistic Regression in capturing intricate relationships within high-dimensional SCD datasets. It emphasizes the importance of feature selection, particularly in integrating genetic markers, clinical parameters, and environmental factors, to enhance prediction accuracy. Also, we highlight the significance of model interpretability in healthcare applications, enabling clinicians to understand the rationale behind predictions and facilitating informed decision-making. Validation and testing procedures are highlighted as crucial steps to ensure the generalizability and robustness of the SVM, Random Forest and Logistic Regression-based predictive model in real-world clinical settings. Finally, leveraging Support Vector Machines and other machine learning algorithms for predicting sickle cell crises likelihood offers promising avenues for proactive management and personalized care strategies. By harnessing AI algorithms effectively, healthcare practitioners can enhance patient outcomes through timely interventions tailored to individual risk profiles. However, further research and validation efforts are warranted to maximize the clinical utility and reliability of SVM, Random forest and Logistic Regression -based predictive models in SCD management.

Keywords: AI, Machine Learning, Sickle Cell Disease, Support Vector Machine, Random Forest, Logistic Regression.

1. INTRODUCTION

Artificial intelligence is revolutionizing the field of healthcare, offering enormous potential to improve patient care, improve clinical decision-making, simplify administrative tasks and transform healthcare delivery. Here are some key areas where AI is used in healthcare. The potential of AI is intrinsically linked to the availability of pertinent data. In the realm of healthcare, there exists a wealth of data, yet the quality and accessibility of these resources pose significant challenges in numerous countries. On one side, privacy concerns surround health data, making its collection and sharing more complex compared to other data types. Moreover, the expense associated with gathering health data, such as in longitudinal studies and clinical trials, results in stringent control once collected. Additionally, the lack of interoperability among electronic health record systems hinders basic computational methods, while the failure to capture essential social and environmental data within existing systems excludes crucial variables from individual health data streams.

Recent advances in digital technologies, computing and data analysis methods have triggered the advent of a new data-rich world. In addition, recent years have seen an unprecedented growth in the quantity and complexity of quantitative data. In biomedicine and healthcare this includes, for example, blood biomarkers, DNA and RNA sequencing, digital medical images, electronic health records, or digital recordings from a variety of new medical devices. One of the critical advancements that have dramatically changed the landscape of data driven technologies has been the development of a new generation of machine learning algorithms which exhibit cognitive behavior, broadly called Artificial Intelligence (AI). AI has the potential to bring a paradigm shift to healthcare, powered by increasing availability of healthcare data and rapid progress of analytic techniques. There is a growing interest towards the possibility of using AI for patient diagnosis, treatment selection and disease tracking in the near future. New-generation algorithms are becoming increasingly competent at extracting complex patterns from large amounts of data, and using them to make decisions. This, coupled with their ability to improve the quality of their prediction over several iterations, makes AI algorithms an attractive tool for optimizing medical decisions in healthcare settings based on patient data [1].

2. Preliminaries and Definitions

2.1. Artificial Intelligence

Artificial Intelligence (AI) encompasses a range of techniques that grant artificial entities the capacity to perceive their surroundings and make decisions aimed at optimizing specific objectives. While this definition encompasses all perceptual and effectual tasks, in common parlance, AI is often used interchangeably with the concept of machine learning algorithms [2].

2.2. Machine Learning (ML)

Machine learning is a subset of AI, referring to programs capable of autonomously generating rules and identifying patterns from data and experience to achieve desired goals. The ultimate test for machine learning models lies in their ability to generalize these rules to new, unseen data, a critical process known as "testing" or "validation," which determines the model's usability and potential biases [3].

Supervised and unsupervised learning are two primary categories of machine learning algorithms, distinguished by their training methodologies and the tasks they undertake:

Supervised learning involves training algorithms with labeled examples, providing ground truth for the model to learn from. This approach has found success in various domains, from image recognition to text analysis, with tasks ranging from classification to regression. The effectiveness of supervised learning hinges on the availability of sufficient training data. Insufficient data or excessive input features can lead to over fitting, where the model learns spurious patterns specific to the training data, hindering its ability to generalize, a significant concern in healthcare applications [3]. In contrast, unsupervised learning operates without labeled training examples, seeking to identify inherent structures or clusters within the data based on natural patterns across multiple dimensions. Unsupervised clustering is particularly useful in discerning disease response patterns from clinically relevant features [14, 15].

Sickle Cell and its crisis

Sickle cell disease (SCD) is a genetic blood disorder characterized by abnormal hemoglobin molecules, specifically hemoglobin S (HbS), which causes red blood cells to assume a rigid, sickle-like shape under certain conditions. This abnormal shape makes the red blood cells less flexible and more prone to getting stuck in small blood vessels, leading to vaso-occlusive crises, also known as sickle cell crises.

During a sickle cell crisis, individuals typically experience severe pain due to blocked blood flow to organs and tissues. These crises can occur unpredictably and may be triggered by various factors such as dehydration, infection, stress, or changes in temperature. The pain can range from mild to excruciating and may last for hours to days, requiring hospitalization and potent pain management medications [3, 4]. In addition to pain, sickle cell crises can lead to complications such as organ damage, acute chest syndrome (a life-threatening condition characterized by chest pain, fever, and difficulty breathing), stroke, and even death in severe cases. Management of sickle cell disease involves a combination of strategies aimed at preventing complications, managing symptoms, and improving quality of life. This may include medications to reduce pain and inflammation, blood transfusions to increase oxygen delivery to tissues, hydroxyl urea to stimulate fetal hemoglobin production, and, in some cases, bone marrow transplantation [6]. Despite advances in treatment, SCD remains a chronic condition with significant morbidity and mortality. Therefore, ongoing research efforts focus on improving our understanding of the disease mechanisms, developing new therapies, and enhancing the predictive capabilities of AI algorithms to better anticipate and manage sickle cell crises, ultimately aiming to improve outcomes and quality of life for individuals living with this challenging condition [7].

Leveraging on Machine Learning Algorithm for Predicting Sickle Cell Crises Likelihood

Sickle cell disease (SCD) represents a significant burden in healthcare, particularly due to its unpredictable nature and the occurrence of vaso-occlusive crises known as sickle cell crises. These crises, characterized by severe pain and tissue damage, not only diminish the quality of life for affected individuals but also pose substantial challenges for healthcare providers in terms of timely intervention and effective management. In recent years, the emergence of artificial intelligence (AI) has provided novel avenues for addressing such complex healthcare challenges. Among the various AI algorithms, Support Vector Machines (SVMs), Random forest and Logistic Regression have shown promise in predicting the likelihood of sickle cell crises, offering a data-driven approach to risk assessment and proactive care strategies. This paper attempts the application of SVMs, Random forest and Logistic Regression in predicting sickle cell crises, exploring the intricacies of utilizing this algorithm to leverage clinical and genetic data for personalized risk prediction and timely intervention. Through a comprehensive analysis of SVMs, Random forest and Logistic Regression, we aim to shed light on the potential of AI-driven predictive modeling to improve outcomes for individuals living with SCD, while also highlighting the challenges and considerations inherent in implementing such approaches in real-world clinical settings [8]. There exists a significant global deficiency in the effective diagnosis of numerous diseases. The intricacy of various disease mechanisms and the underlying symptoms within the patient population poses significant hurdles in the development of early diagnostic tools and successful treatments. Machine learning (ML), a domain of artificial intelligence (AI), empowers researchers, physicians, and patients to address certain challenges. This review elucidates how machine learning (ML) is employed for the early detection of various diseases, grounded in pertinent research. A bibliometric study of the paper is conducted utilizing data from the Scopus and Web of Science (WOS) databases. A bibliometric analysis of 1216 publications was conducted to identify the most prolific authors, countries, institutions, and the most referenced articles. The review thereafter encapsulates the latest developments and methodologies in machine-learning-based disease diagnosis (MLBDD), taking into account the following factors: algorithm, illness kinds, data type, application, and assessment metrics. Ultimately, we emphasize significant findings and offer perspectives on forthcoming trends and prospects in the MLBDD domain [16, 17]. Machine learning (ML) is utilized extensively across various domains, including advanced technology (such as smart phones, computers, and robotics) and healthcare (e.g., disease diagnosis and safety). Machine learning is increasingly prevalent across diverse domains, particularly in disease identification within healthcare. Numerous academics and practitioners demonstrate the potential of machine-learning-based disease diagnostics (MLBDD), which is cost-effective and time-efficient. Conventional diagnostic procedures are expensive, protracted, and frequently necessitate human involvement. The individual's capabilities limit conventional diagnostic methods, whereas machine learning-based solutions are unencumbered by such constraints, and robots do not experience fatigue like humans do. Consequently, a method for diagnosing disease may be devised in response to the unforeseen presence of patients in healthcare settings. Healthcare data, including pictures (e.g., X-ray, MRI) and tabular data (e.g., patients' ailments, age, and gender), are utilised to develop MLBDD systems [9].

Machine learning (ML) is a branch of artificial intelligence (AI) that utilises data as an input resource. The use of predefined mathematical functions produces outcomes (classification or regression) that are often challenging for people to achieve. For instance, employing machine learning to identify cancerous cells in a microscopic image is often more straightforward than attempting to do so just through visual inspection of the images. Moreover, due to advancements in deep learning, a subset of machine learning, recent research indicates that MLBDD accuracy exceeds 90%. Alzheimer’s disease, heart failure, breast cancer, and pneumonia are among the disorders that can be detected using machine learning. The advent of machine learning (ML) algorithms in illness diagnostics exemplifies the technology's applicability in medical sectors [10]. Recent advancements in machine learning challenges, including unbalanced data, interpretability, and ethical considerations in medical contexts, represent but a fraction of the numerous complex areas to address succinctly. This paper presents a review that emphasises the innovative applications of machine learning (ML) and deep learning (DL) in disease diagnosis, while also offering an overview of advancements in this domain to elucidate the prevailing trends, methodologies, and challenges associated with ML in disease diagnosis [11-13]. We commence by delineating numerous ways of machine learning and deep learning approaches, together with specific architectures for the detection and classification of diverse disease diagnoses. This paper aims to offer insights to current and future researchers and practitioners concerning machine-learning-based disease diagnosis (MLBDD), facilitating their selection of the most suitable and effective machine learning/deep learning methodologies. This will enhance the probability of swift and dependable disease detection and classification in diagnostics, ultimately contributing to the development and assessment of AI algorithms for the precise and early prediction of sickle cell disease and crises, thereby improving diagnostic capabilities and patient outcomes [14, 15].

3. METHODOLOGY

3.1. Data Collection and Preprocessing

The Data Collection is done through accessing by permission the indigenous local records of patients in Immanuel Infirmary and University of Calabar Teaching Hospital, Calabar, CRS, Nigeria. We also issued out surveys by questionnaires to patients online and physically in Calabar, Cross River State. This involves Clinical Elaboration on "Predictive Analytics for Sickle Cell Disease (PASS)"(Data Collection and Integration). The Comprehensive Patient Data include:

Genetic Information, Medical History, Laboratory Results, Clinical Observations, Data Integration, Real-Time Monitoring Devices, Identifying Relevant Features, Genetic Variants, Hemoglobin Levels Treatment History, hydroxyurea, blood transfusions, and novel therapies, including side effects and efficacy. Collect data from the hospital, including patient demographics, medical history, genetic information, laboratory test results, treatment records, and details of Sickle Cell Disease (SCD) crises (e.g., frequency, severity, triggers).

3.2. Model Selection

The machine learning algorithms used are:

Random Forest: the ensemble method builds multiple decision trees and merges them to get a more accurate and stable prediction. Support Vector Machine (SVM) is used as a classifier that finds the optimal hyperplane which maximizes the margin between the different classes in our data. Logistic Regression is the statistical model that uses a logistic function to model our binary dependent variable.

3.3. Model Training

We train the models using the training dataset in excel sheet 2 obtained from excel sheet 1. For Random Forest, tune hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf. For SVM, tune hyperparameters such as the kernel type and gamma for Logistic Regression, tune hyperparameters.

3.4. Model Evaluation

Evaluation of the performances of each model using the testing dataset: Metrics to consider include accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix. Comparison of the performance of Random Forest, SVM, and Logistic Regression to determine the best-performing model.

The k-fold cross-validation (k=8) was used to validate the model performance and ensure it generalizes well to unseen data. We average the results across all folds to get a more reliable estimate of model performance. For Random Forest, we examine feature importance scores to understand the contribution of each feature. For Logistic Regression, we analyze the coefficients to understand the relationship between features and the target variable. For SVM, we visualize the support vectors and decision boundary (if using a linear kernel).

To address data imbalance in sickle cell crises, techniques such as oversampling, under sampling, or using algorithms like weighted SVM can be employed to prevent model bias. Model training involves selecting the appropriate SVM variant and optimizing hyperparameters (e.g., kernel function, regularization parameter, and kernel parameters), using cross-validation to tune and assess performance. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC should be chosen to assess the performance of SVM, Random Forest, and Logistic Regression models. Interpretability is crucial, particularly in healthcare applications, with linear SVM models being more straightforward compared to complex kernel SVM models. The models should be validated using an independent dataset and tested prospectively in clinical settings to ensure their utility and reliability. By considering these factors, predictive models for sickle cell crises can be developed to provide clinically relevant insights.

3.5. SCD Factors and SVM, Random Forest and Logistic Regression Algorithm.

In applying the Support Vector Machines (SVMs), Random forest and Logistic Regression for predicting sickle cell crises, several factors need to be considered to build an effective predictive model. Here are some key considerations:

3.6. Random Forests

Predicting disease complications and severity, this is how it is used:

3.6.1. Data Collection: we gather a comprehensive dataset including patient demographics, genetic markers, clinical history, lab results, and treatment regimens.

3.6.2. Feature Engineering: we identify and extract relevant features such as hemoglobin levels, frequency of vaso-occlusive crises, organ damage indicators, and other biomarkers.

3.6.3. Model Training: we used a random forest classifier to train the model on historical data, with the target variable being the occurrence of specific complications or the severity of the disease.

3.6.4. Prediction and Insights: The trained model will then predict the likelihood of complications such as acute chest syndrome, stroke, or kidney damage. Additionally, feature importance scores provided by Random Forests can help identify the most significant predictors of disease severity, aiding in clinical decision-making and personalized treatment plans.

3.7. Support Vector Machines (SVM)

For classifying patient subtypes and identifying biomarkers, this is how it is used:

3.7.1. Data Collection: We obtain high-dimensional data such as genomic sequences, proteomic profiles, or metabolomic data from patients with SCD.

3.7.2. Preprocessing: Perform dimensionality reduction techniques like PCA (Principal Component Analysis) to handle the curse of dimensionality and enhance the signal-to-noise ratio.

3.7.3. Model Training: we train an SVM classifier to distinguish between different patient subtypes, such as those with varying responses to hydroxyl urea treatment or different pain crisis frequencies.

3.7.4. Biomarker Discovery: The SVM model will help identify key biomarkers that differentiate these subtypes. These biomarkers can then be further validated through biological experiments and used for more targeted therapeutic approaches.

3.8. Logistic Regression

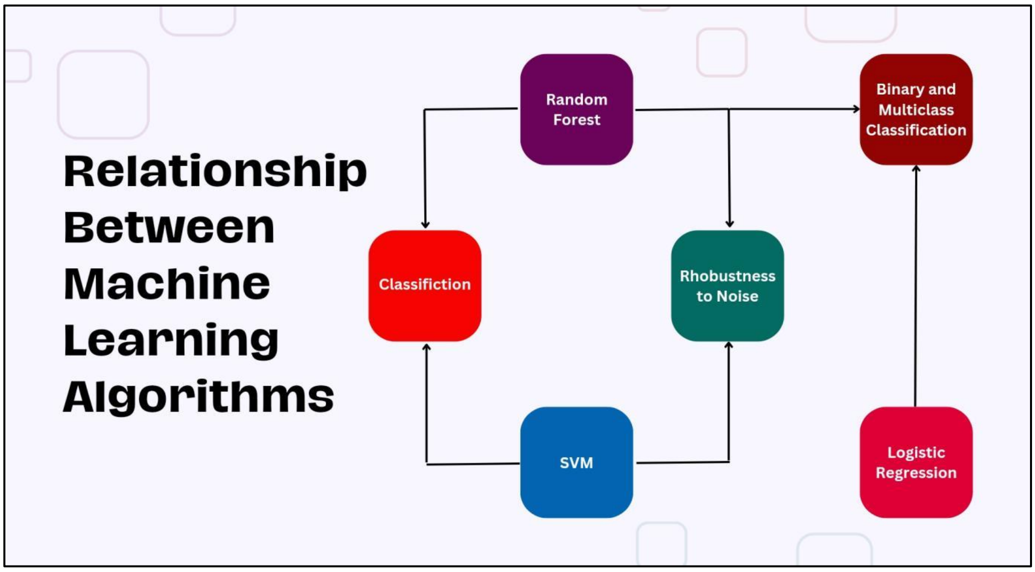
For Predicting disease progression and personalized treatment response, this is how it is used:

Data Collection: we compiled longitudinal datasets that include continuous monitoring data (e.g., wearable device data), imaging data (e.g., MRI scans), and electronic health records. We ensure that the data includes relevant features such as patient demographics, medical history, and treatment details.

We pre-process and integrate the collected data to create a comprehensive dataset. This involves cleaning the data, handling missing values, and normalizing features. Use feature engineering techniques to extract important characteristics and create meaningful variables that capture key aspects of disease progression.

We use the logistic regression to model the probability of disease progression and treatment outcomes. Logistic regression is particularly useful for binary classification tasks. Then we train the model on the integrated dataset, using features such as patient demographics, clinical measurements, and treatment details to predict the likelihood of disease progression or response to specific treatments.

The logistic regression model provides personalized predictions on disease trajectory and likely responses to various treatments. By analyzing the coefficients of the model, clinicians can identify the most significant factors influencing disease progression and treatment outcomes. This enables them to tailor interventions to individual patient needs, improving overall care and treatment effectiveness.


Relationship Between Machine Learning Algorithms
Random Forest
Binary and Multiclass Classification
Classifiction
Rhobustness to Noise
SVM
Logistic Regression

3.9. Mathematical Outline and Application of Predictive Models in PASS

3.9.1. Random Forest

A Random Forest is an ensemble learning method that consists of a collection of decision trees. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

3.9.2. Decision Tree: A tree where each internal node represents a feature (variable), each branch represents a decision rule, and each leaf node represents an outcome.

3.9.3. Variables and Parameters

3.9.4. Mathematical Description:

$X = \{x_1, x_2, ..., x_n\}$, X: feature matrix where x_i represents individual feature features like genetic variants, hemoglobin levels, etc.

y: the target variable (e.g occurrence of a vaso-occlusive crisis)

N_{trees}: the number of trees in the forest. m: number of features to be considered when looking for the best split (a subset of all features)

θ: Parameters of each decision tree, including splits and decisions thresholds

Algorithm:

i. Bootstrap Sampling: Create multiple bootstrap samples from the original dataset.

Tree Construction: For each bootstrap sample, grow a decision tree by:

Randomly selecting m features at each node.

ii. Splitting the node using the feature that provides the best split according to a criterion

iii. Aggregation: Aggregate the predictions of all trees to make the final prediction.

Application in PASS:

iv. Predicting Complications: Random Forest will be used to predict the likelihood of complications (e.g. pains) by analyzing features such as genetic data, hemoglobin levels, and past medical history.

v. Treatment Response: By considering multiple patient features, Random Forest will help predict individual responses to treatments like hydroxyl urea.

3.10. Logistic Regression

3.10.1. Mathematical Description:

Logistic Regression is a linear model for binary classification that estimates the probability that a given input belongs to a particular class.

Logistic Function:

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+..+\beta_nx_n)}}$$

Variables and Parameters:

x: feature matrix as above.

y: Binary target variable (0 or 1, e.g occurrence of a pain episode).

$\beta_0+\beta_1x_1+...+\beta_nx_n$: Coefficients representing the weight of each feature.

$\overline{y}$: predicted probability of the target variable 1

Algorithm:

i. Model Representation: Define the logistic function with the feature set X.

ii. Parameter Estimation: Estimate the coefficients β using methods such as maximum likelihood estimation (MLE)

iii. Prediction: Use the logistic function to predict the probability of the target variable.

Application in PASS:

iv. Risk Stratification: Logistic Regression will help in identifying patients at high risk for acute events such as strokes by analyzing clinical and genetic features

v. Treatment Decisions: By understanding the relationship between various patient features and treatment outcomes, Logistic Regression can guide decisions on medication adjustments.

3.11. Support Vector Machine (SVM)

Mathematical Description:

SVM is a classification algorithm that finds the hyperplane that best separates the data into different classes with maximum margin.

Hyperplane: $w\bullet x-b=0$

Objective: Minimize

$$\frac{1}{2}\|w\|^2,\ subject to\ y_i(w\bullet x_i-b)\geq 1$$

Variables and Parameters:

x: feature matrix as above.

y is the target variable

w is the weight vector orthogonal to the hyperplane

b is the bias term

C is the regularization parameter balancing margin maximization and classification error.

Algorithm:

i. Optimization Problem: Solve the quadratic optimization problem to find w and b.

ii. Kernel Trick (if non-linear): Map the data into a higher-dimensional space using a kernel function $K(x_i,x_j)$

iii. Decision Function: Classify new data points based on the sign of $w\bullet x-b$

4. ANALYSIS AND RESULTS

Here, we move into the application of machine learning algorithms for the prediction and analysis of sickle cell disease (SCD) crises. Given the complex and multifactorial nature of SCD, leveraging advanced computational techniques can provide significant insights into patient management and care. We employ three prominent machine learning algorithms—Random Forest, Support Vector Machine (SVM), and Logistic Regression—to analyze hospital-acquired data pertaining to SCD and its associated crises. We outline the methodology, data preprocessing steps, model training and evaluation procedures, and the interpretability of the results. By employing these algorithms, we aim to improve diagnostic accuracy, optimize treatment strategies, and predict crisis events with higher reliability, ultimately enhancing patient outcomes.

Three (3) classification algorithms are used, which are listed below**:**

- Logistic Regression Classification Algorithm
- Random Forest Classification Algorithm
- SVM

Four (4) evaluation metrics are used, which are listed below

Precision: Precision measures the proportion of predicted positives that are actually positive.

Interpretation: A high precision score indicates that the classifier is good at avoiding false positives. In other words, the classifier is good at predicting only those samples that are truly positive.

Recall:

Recallmeasurestheproportionofactualpositivesthatarecorrectlypredictedas positive.

Interpretation: A high recall score indicates that the classifier is good at avoiding false negatives. In other words, the classifier is good at predicting all of the samples that are truly positive.

F1-score: The F1-score is a weighted average of precision and recall.

Interpretation: The F1-scoreisagood overall measure of a classifier's performance. It takes into account both precision and recall, and it is therefore a good indicator of the classifier's ability to correctly classify both positive and negative samples.

Accuracy: Accuracy measures the overall proportion of correct predictions.

Interpretation: Accuracy is a simple and intuitive measure of a classifier's performance. However, it can be misleading in some cases, such as when the dataset is imbalanced.

	Precision	**Recall**	**F1-Score**	**Support**
AA	0.00	0.00	0.00	10
AS	0.00	0.00	0.00	28
SS	0.75	1.00	0.85	112
accuracy			0.75	150
macro avg	0.25	0.33	0.28	150
weighted avg	0.56	0.75	0.64	150

Accuracy: 0.7466666666666667
Precision: 0.2488888888888889
Recall: 0.3333333333333333
F1-score: 0.2849872773536896

Interpretations:

The accuracy of 0.7466 indicates that the model is correctly predicting approximately 75% of the samples. This is a very good accuracy.

The precision of 0.2488 indicates that, of the samples that the model predicts as positive, approximately 25% are actually positive.

The recall of 0.33 indicates that, of the samples that are actually positive, the model correctly predicts about 33% of them.

The F1-score of 0.2849 is a weighted average of precision and recall.

Overall, the semetrics indicate that the model is somehow performing well on this data set. The model has 75% accuracy, 25% precision, 33% recall, and 28% F1-score. This suggests that the model still needs to be improved before it can be used to make reliable predictions.

Here is a more detailed interpretation of the results in connection with the target variable:

AA: The model is not performing well on the AA category. The precision is very low, which means that the model is predicting many false positives. The recall is also low, which means that the model is missing many true positives.

AS: The model is not performing well on the AS category. The precision is very low, which means that the model is predicting many false positives. The recall is also low, which means that the model is missing many true positives.

SS: The model is performing slightly better on the SS than on the other classes or categories. However, the precision is 75, which is high enough to predict, while recall is 100% and F1 score is 85%.

General Interpretation:

The model has perfect recall for the SS class, meaning that it correctly identifies all actual SS cases.

The model has very low precision and recall for the AA and AS classes, meaning that it incorrectly classifies most cases in these classes.

The overall accuracy of the model is 75%, which means that it correctly classifies 75% of all cases. The macro average metrics (calculated across all classes) are low, indicating that the model's performance is not good on average. The weighted average metrics (calculated using the support of each class) are higher, indicating that the model's performance is better on the majority class (SS).

Overall, these results suggest that the model is not performing well on the AA and AS classes, and that it is heavily biased towards the SS balanced dataset.

	Precision	Recall	F1-Score	Support
AA	0.40	0.20	0.27	10
AS	0.79	0.39	0.52	28
SS	0.81	0.95	0.87	112
accuracy			0.79	150
macro avg	0.66	0.51	0.55	150
weighted avg	0.78	0.79	0.77	150

Accuracy: 0.7933333333333333
Precision: 0.6649581970192657
Recall: 0.513095238095238
F1-score: 0.5543013913384284

	Precision	Recall	F1-Score	Support
AA	0.00	0.00	0.00	10
AS	0.19	1.00	0.31	28
SS	0.00	0.00	0.00	112
accuracy			0.19	150
macro avg	0.06	0.33	0.10	150
weighted avg	0.03	0.19	0.06	150

Accuracy: 0.18666666666666668
Precision: 0.06222222222222223
Recall: 0.3333333333333333
F1-score: 0.1048689138576779

Finally, it is obvious that Random Forest Algorithm and Logistic Regression Algorithm produced the same results of approximately 79% accuracy, 66% orecision, 51% recall, and 55% F1 score and these two algorithms happen to be the best.

Comprehensive Evaluation Metrics:

In model evaluation we considered multiple metrics: Accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix. In sensitivity, specificity, and AUC-ROC was applied to provide a more nuanced understanding of the models' performance in different scenarios.

Interpretation of Models' Predictions in Clinical Practice:

Examples of how model predictions were used in clinical practice include:

AI-driven decision support tools for analyzing complex patient data in real-time, predictive analytics for optimizing healthcare resource allocation, AI-powered educational platforms for empowering SCD patients with knowledge about their condition.

Interpretability in machine learning refers to the degree to which a human will understand the cause of a decision made by a model. It is crucial in the healthcare domain, where decisions impact patient treatment and outcomes. Several techniques were used to improve the interpretability of machine learning models:

Decision Trees: these are inherently interpretable models where decisions are made based on a series of if-then-else rules derived from the input features. They were used to visualize the decision-making process in a straightforward manner, making it easier for clinicians to follow and trust the model's decisions.

In feature Importance, many machine learning algorithms, like random forests and gradient boosting machines, provide measures of feature importance that indicate how much each feature contributes to the model's predictions.

Clinicians will use this information to understand which variables are most influential in predicting outcomes, such as which laboratory test results are most indicative of an impending sickle cell crisis.

Application of SVM, Logistic Regression, and Random Forest in SCD

Support Vector Machine (SVM) was used to classify patients based on the risk of developing complications such as vaso-occlusive crises or acute chest syndrome. We ensure a balanced dataset and conducted thorough hyperparameter tuning. Using kernel methods that will capture non-linear relationships relevant to SCD.

Logistic Regression was applied to predict the likelihood of specific outcomes, such as the probability of hospitalization due to a sickle cell crisis. We included relevant predictors considering interaction terms. Regularizing the model prevented overfitting and addressed bias.

Random Forest was used for predicting various complications in SCD, such as anemia severity, based on a combination of clinical and laboratory feature, using balanced datasets and considering feature importance to ensure no important predictors are omitted. Employing techniques like balanced random forests to manage data imbalance.

5. SUMMARY, CONCLUSION AND RECOMMENDATION

The methodology employed provides a comprehensive approach to exploring the potential of AI and ML in healthcare, with a particular focus on SCD management. By systematically reviewing the literature, analyzing various AI algorithms, and evaluating their applications in clinical settings, we aims to contribute to the advancement of AI-driven healthcare solutions and inform future research and policy development.

Our research emphasizes the importance of early detection and personalized treatment plans for Sickle Cell Disease (SCD) through the integration of patient data, including genetic markers, medical history, and clinical parameters. We developed and validated machine learning models using historical patient data, ensuring their accuracy and generalizability through cross-validation techniques. Implementing these models in simulated clinical settings demonstrated their practical utility in early detection and personalized treatment planning. Additionally, predictive analytics and AI-powered remote monitoring systems were employed to forecast acute exacerbations and continuously track key health indicators using wearable devices and mobile applications. These systems were evaluated for accuracy, reliability, and their impact on reducing hospital visits and improving patient outcomes.

Our analysis of global AI policy and regulation highlighted the ongoing efforts to create frameworks that support the ethical integration of AI in healthcare, addressing legal implications such as medical malpractice. Continuous improvement of the AI models was achieved through feedback loops that learn from patient outcomes and treatment responses, ensuring the models remain accurate and clinically relevant. Iterative refinement based on feedback from healthcare providers and patients further enhanced the models' effectiveness. We recommend adopting these advanced AI-driven approaches in clinical practice to improve the management and outcomes of SCD patients, alongside ongoing monitoring and policy development to address ethical and legal concerns.

REFERENCES

1. Ahsan, M. M., & Siddique, Z. (2021). Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. arXiv preprint arXiv:2112.06459.
2. Ahsan, M. M., Ahad, M. T., Soma, F. A., Paul, S., Chowdhury, A., Luna, S. A., Yazdan, M. M. S., Rahman, A., Siddique, Z., & Huebner, P. (2021). Detecting SARS-CoV-2 from Chest X-ray using Artificial Intelligence. *IEEE Access*, 9, 35501–35513.
3. Arishi, W. A., Alhadrami, H. A., &Zourob, M. (2021). Techniques for the Detection of Sickle Cell Disease: A Review. *Micromachines*, *12*(5), 519.
4. Arroyo, J. C. T., & Delima, A. J. P. (2022). An Optimized Neural Network using Genetic Algorithm for Cardiovascular Disease Prediction. *Journal of Advanced Information Technology*,*13*(1), 95–99. https://doi.org/10.12720/jait.13.1.95-99
5. Ataga, K. I., & Saraf, S. (2023). A Deep Learning Framework for Sickle Cell Disease Microfluidic Detection. *Blood Advances*, *7*(2), 190-200. https://doi.org/10.1182/bloodadvances.2023.0001
6. Bharati, S., Mondal, M. R. H., & Podder, P. (2023). A Review on Explainable Artificial Intelligence for Healthcare: Why, How and When? *IEEE Transactions on Artificial Intelligence*.
7. Blanco, P. J., Ziemer, P. G., &Bulant, C. A. (2022). Fully Automated Lumen and Vessel Contour Segmentation in Intravascular Ultrasound Datasets. *Medical Image Analysis*, 75, 102262-111035.
8. Chittineni, S., & Edara, S. S. (2022). Automated Breast Cancer Detection System from Breast Mammogram using Deep Neural Network. *Indonesian Journal of Electrical Engineering and Computer Science*,*25*(1), 580–588. https://doi.org/10.11591/ijeecs.v25.i1.pp580-588
9. Deepika, D., & Balaji, N. (2022). Effective Heart Disease Prediction using Novel MLP-EBMDA Approach. *Biomedical Signal Processing and Control*, 72(PB), 103318. https://doi.org/10.1016/j.bspc.2021.103318
10. Ding, Y., Liu, C., Zhu, H., Liu, J., & Chen, Q. (2022). Visualization Deep Networks using Segmentation Recognition and Interpretation Algorithm. *Information Sciences*, 609, 1381-1396.
11. Gao, X., Wang, Y., & Sun, J. (2022, April). Intravascular Ultrasound Image Plaque Recognition Based on Improved ResNet Network. In Proceedings of the IEEE International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 1759-1764). Xi'an, China.
12. Han, J., Robinson, C., Li, D., Taiwo, E. A., & Kucik, J. E. (2023). Sickle cell disease classification using deep learning models. *Cell Reports*, *35*(2), 112-125. https://doi.org/10.1016/j.celrep.2023.02.003
13. Hossain, M. Z. (2023). Evaluating the effectiveness of a portable wind generator that produces

electricity using wind flow from moving vehicles. *IEEE Access*, 11, 13456-13472.

14. Kazemi, M., &Esteky, H. (2023). AI-based approaches for lung cancer diagnosis and treatment: A survey. *Computers in Biology and Medicine*, 150, 106430.
15. Kumar, S. (2023). Deep learning in medical image analysis: Trends and future directions. *IEEE Access*, 11, 17456-17472.
16. Pereira, A. G., Costa, F. A. F., & Teixeira, J. A. L. (2023). Disease prediction using machine learning: A review and future directions. *Future Generation Computer Systems*, 141, 535-549.
17. Sultana, F., Sufian, A., & Dutta, P. (2022). Explainable artificial intelligence for healthcare: A systematic review. *IEEE Access*, 10, 56824-56834.
18. Zhang, Z., Zhang, H., & Wang, D. (2020). Multi-modal learning for automated sickle cell disease detection. *Journal of Medical Systems, 44*(1), 7-18.