University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

The Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 15 No. 1, September 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

A Machine Learning Framework for Classifying Haemoglobin Levels in Sickle Cell Anaemia Patients

[1]✉Olajide O. B., [1]Sakpere A. B., [1]Adeyemo A. B., [2]Ogbole G. I., [3]Arekete S. A. and [4]Aribisala S. B.

[1]Department of Computer Science, University of Ibadan, Nigeria
[2]Department of Radiology, University of Ibadan College of Medicine, Ibadan, Nigeria
[3]Department of Computer Science, Redeemer's University, Ede, Nigeria
[4]Department of Computer Science, Lagos State University, Nigeria

[1]oolajide4174@stu.ui.edu.ng, *[1]ab.sakpere@ui.edu.ng*, *[1]sesanadeyemo2014@gmail.com*, *[2]gogbole@gmail.com*, [3]areketes@run.edu.ng, *[4]aribisala@uchicago.edu*

Abstract

Sickle Cell Anaemia (SCA) significantly impacts haemoglobin (HGB) levels, leading to severe health complications with high mortality rates. In Nigeria, about 2% of newborns, approximately 150,000 annually, are diagnosed with SCA. Accurate HGB monitoring is essential for effective disease management, yet traditional methods are labour-intensive and prone to errors. This necessitates automated and reliable diagnostic techniques like machine learning (ML) for improved SCA management. This study classifies HGB levels in SCA patients using clinical records and ML techniques. A dataset of 364 records (203 female population) was obtained from Kaggle; a public data repository containing eleven (11) features namely: age, sex, red blood cell (RBC) count, packed cell volume (PCV), mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC), red cell distribution width (RDW), total leukocyte count (TLC), platelets per cubic millimeter (PLT/mm³), and haemoglobin (HGB). Two ML models, Logistic Regression (LR) and Support Vector Machine (SVM), were used with two feature selection methods: all features and selected features. The latter identified age, RBC, PCV, MCV, and HGB as key predictors. Continuous HGB values were categorized into (1) low, (2) normal, and (3) high using standard medical metrics. SMOTE analysis was also carried out to mitigate class imbalance. SVM with a Radial Basis Function (RBF) kernel achieved 84.90% accuracy and AUC-ROC of 93.40%, while LR underperformed with 79.50% accuracy and AUC-ROC of 90.90%. Using all feature selection, SVM improved to 91.80% accuracy and AUC-ROC of 98.20%, with LR achieving accuracy of 93.20% and AUC-ROC of 98.90%. Both models demonstrated high accuracy, with LR excelling using all features, while SVM performed better with selected features. Future work will involve the use of primary datasets, additional feature selection techniques and ML algorithms, and incorporate the use of Haemoglobin variants to provide further insight into SCA progression and in turn offer personalized treatment.

Keywords: *Haemoglobin level classification, Logistic Regression, Machine learning models, Sickle cell anaemia, Support Vector Machine*

1. Introduction

Sickle Cell Anaemia (SCA) is a hereditary blood disease that is characterized by sickled red blood cell (RBC) usual caused by an abnormal gene mutation [1]. This abnormality is caused by a genetic substitution of glutamic acid with valine which alters the nature and function of RBCs. Unlike normal RBCs which are biconcave in nature, a SCA's patient RBCs are sickle-shaped and deformed. This morphological abnormality hinders the seamless flow of blood, thereby leading to acute complications [2]. This structural abnormality disrupts oxygen transport, precipitating vaso-occlusion, chronic haemolysis, and ischemic tissue damage, which cumulatively contribute to severe complications such as stroke, organ dysfunction, and acute chest syndrome [3, 4]. The disease predominantly affects individuals of African, Mediterranean, Middle Eastern, and Indian ancestry, with Nigeria bearing the highest

Olajide O. B., Sakpere A. B., Adeyemo A. B., Ogbole G. I. Arekete S. A. and Aribisala S. B. (2025). A Machine Learning Framework for Classifying Haemoglobin Levels in Sickle Cell Anaemia Patients. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 15 No. 1, pp. 1 - 12

global burden, where approximately 150,000 infants are born annually with the condition [5,6]. Despite advancements in treatment, SCA continues to pose a significant public health challenge due to its high morbidity and mortality, particularly in low-resource settings with limited access to specialized healthcare.

The severity of sickle cell disease varies greatly among varied populations, necessitating proper classification. This severity is driven by a variety of radiological variables, including RBC shortage and the frequency of clinical signs such as acute chest syndrome [7,8]. Haemoglobin (HGB) levels are a good predictor of SCA severity and ongoing clinical symptoms. Patients who suffer these ongoing manifestations always require appropriate and attentive care to improve their life expectancy [9,10]. Traditional techniques of SCA severity categorization rely exclusively on blood smear investigation of the patient. However, the process can be tasking and susceptible human error especially in low resourced area where access to top notch diagnostic tools is unavailable [11, 12].

The increasing availability of large-scale clinical datasets and advancements in computational methodologies have facilitated the application of artificial intelligence (AI) and machine learning (ML) techniques in medical diagnostics, offering a promising paradigm shift in disease classification and predictive analytics [11, 3].

ML models have demonstrated remarkable efficacy in various healthcare applications, including disease detection, prognostic modelling, and personalized medicine, by leveraging high-dimensional data to uncover intricate patterns that may not be immediately apparent through conventional statistical methods [11, 12, 14]. In the context of SCA, ML-based classification systems hold significant potential for improving the accuracy and efficiency of HGB level estimation, thereby enabling more timely and targeted therapeutic interventions [14]. Among the array of ML algorithms explored for classification tasks, Support Vector Machines (SVM) and Logistic Regression (LR) have emerged as viable candidates due to their robust decision-making capabilities and interpretability in clinical settings [14]. While SVM excels in constructing optimal hyperplanes to delineate complex classification boundaries, LR provides a probabilistic framework for modelling the likelihood of different disease states based on predictive variables, both of which are critical for reliable HGB classification [14,15].

Despite the transformative potential of ML models in haematological diagnostics, research in this domain remains relatively nascent, necessitating further empirical investigations to establish their clinical validity and real-world applicability. Existing studies have primarily focused on model optimization through hyperparameter tuning and feature engineering to enhance classification performance [14, 15, 16]. However, a persistent challenge in ML-based classification lies in feature selection, where the inclusion of extraneous variables may introduce noise and compromise model interpretability [16]. To address this limitation, this study employed a dual-feature selection strategy comprising all feature selection and selected features. The former leverages all available features to maximize predictive power, while the latter prioritizes clinically relevant variables to improve interpretability and facilitate integration into existing diagnostic workflows. A comparative evaluation of these models was conducted based on key performance metrics, including accuracy, precision, recall, and computational efficiency, to ascertain the optimal balance between model complexity and diagnostic utility.

By integrating ML techniques into SCA management, this study aims to classify HGB levels in SCA patients using SVM and LR models. By leveraging machine learning techniques, we aim to improve the accuracy and reliability of HGB level classification in SCA cases, provide improved patient outcomes and reduced healthcare burdens associated with chronic haemoglobinopathies classification.

2. Related Works

An increasingly large number of studies have been devoted to the deployment of machine learning (ML) methods in healthcare diagnosis, especially in the case of hematologic diagnosis categories. These ML approaches have demonstrated great potential in the process of

correctly predicting and diagnosing blood disorders like leukaemia, anaemia and SCA through the examination of complete blood count (CBC) parameters and other pertinent biomarkers [7]. Though these breakthroughs have largely influenced the diagnosis of medical conditions, specifically haematological diseases, there exists a lapse in the actual area of haemoglobin (HGB) classification amongst the carriers of SCA. Precise classification of HGB level is essential, as it is one of the central indicators of SCA severity.

HGB levels classification are of clinical benefits in SCA management [6, 7, 18]. One of the most significant factors of the disease progression and life-threatening risks assessment is the level of haemoglobin concentration that is used to individualize transfusion guidelines [8]. Patients who have low levels of HGB portray high danger of critical anaemia, multi-organ injury and high morbidity [7, 8, 9]. Furthermore, molecular analysis can give rise to a more accurate classification that allows clinicians to develop a more patient-specific treatment plan, maximize transfusion plan and accurately monitor the effectiveness of treatment. Timely and correct classification also has the benefits of lowering hospitalization rates and long-term outcomes of patients.

The ability to handle a high dimension dataset and influence robustness makes SVM and LR some of the most used ML algorithms. SVM is highly efficient in its performance with binary classification task and is used in areas such as oncology and cardiovascular diseases diagnostics [8,16]. Likewise, LR has been used clinically to draw and understand patient outcome based on input variable and it had been seen to be especially instrumental in monitoring disease progression [9,10,17]. Empirical research has confirmed the capability of these models to appropriately categorize medical conditions with various arrays of biomarkers. Indeed, SVM has been demonstrated to be useful in classification of various anaemia types based on blood analysis [10]. Similarly, in the COVID-19 patients, it has been found that LR was helpful in estimating the level of haemoglobin and other important haematological features that assisted in categorizing the health profile of patients [11, 18].

In contrast to other ML-based works that have studied either the diagnosis of SCA or the detection of sickle cell traits, little attention has been given to the task of classifying the level of HGB to assist with clinical decision-making [6,7,16]. Bhatia *et al.* [17] used deep learning techniques to classify red blood cell morphologies in SCA patients with 81% accuracy in reference to the cell types including sickle cells and ovalocytes. Nonetheless, they focused on the morphology of the cells instead of an actual estimation of the HGB levels. A similar work was done by Srivastava *et al.* [18] in which ML models based on spectroscopy data were used to diagnose SCA with high sensitivity and specificity, although this work did not solve the problem of quantifying HGB.

Additionally, Ekong *et al.* [19] designed a classification system that diagnoses SCA in adolescents using a Bayesian network with 99 percent accuracy. Still, they emphasized only disease identification, but nothing was done on HGB classification. Concurrent to that, Alzubaidi *et al.* [20] presented deep learning lightweight models capable of classifying erythrocytes into normal, sickle, and miscellaneous categories. Nonetheless, in such method quantitative assessments of haemoglobin parameters were excluded, which rendered the model inapplicable in treatment monitoring.

Though such studies reinforce the idea of the viability of ML in hematological diagnostics, the direct classification of HGB levels has not been studied well. Another work is that of Dada *et al.* [21] who applied convolutional neural networks (CNNs) to study peripheral blood smears in a children population and obtained an anaemia detection of 92% precise rate although they did not measure the level of haemoglobin in a blood sample. Likewise, Zemariam *et al.* [22] used a variety of ML classifiers to determine the anaemia prevalence rate among Ethiopian adolescent girls, and Random Forest scored an area under the curve (AUC) of 82%. However, they were concerned with general anaemia prediction, but not HGB level. Hybrid ML models with the attention mechanism were also suggested by Ramzan *et al.* [23] in respect

of the anaemia detection, which obtained a promising model high accuracy but without focusing on HGB at SCA patients.

With the view of mitigating this gap, the current study focused on determining the performance of SVM and LR models in classifying the HGB level using a well-known CBC data primarily gathered by Mendeley team which was obtained from Kaggle repository. Focusing only on the issue of HGB classification in SCA patients, the study aims develop a machine learning framework for classifying haemoglobin levels in sickle cell anaemia patients. The knowledge obtained can be used as a useful input to the developing area of ML application in medical diagnosis and further encourage extensive research in the field of SCA management.

3. Methodology

The approach used to develop the SCA HGB level classification model are discussed here. The process involved five main stages as shown in Figure 1; each stage is structured to ensure a systematic workflow for model development, optimization, and thorough assessment.

3.1 Data Collection and Description

The dataset used in this study is the Mendeley Complete Blood Count dataset source from Kaggle, an open-source standard dataset repository. The dataset contained 364 patient records and 11 features seen in Table 1.

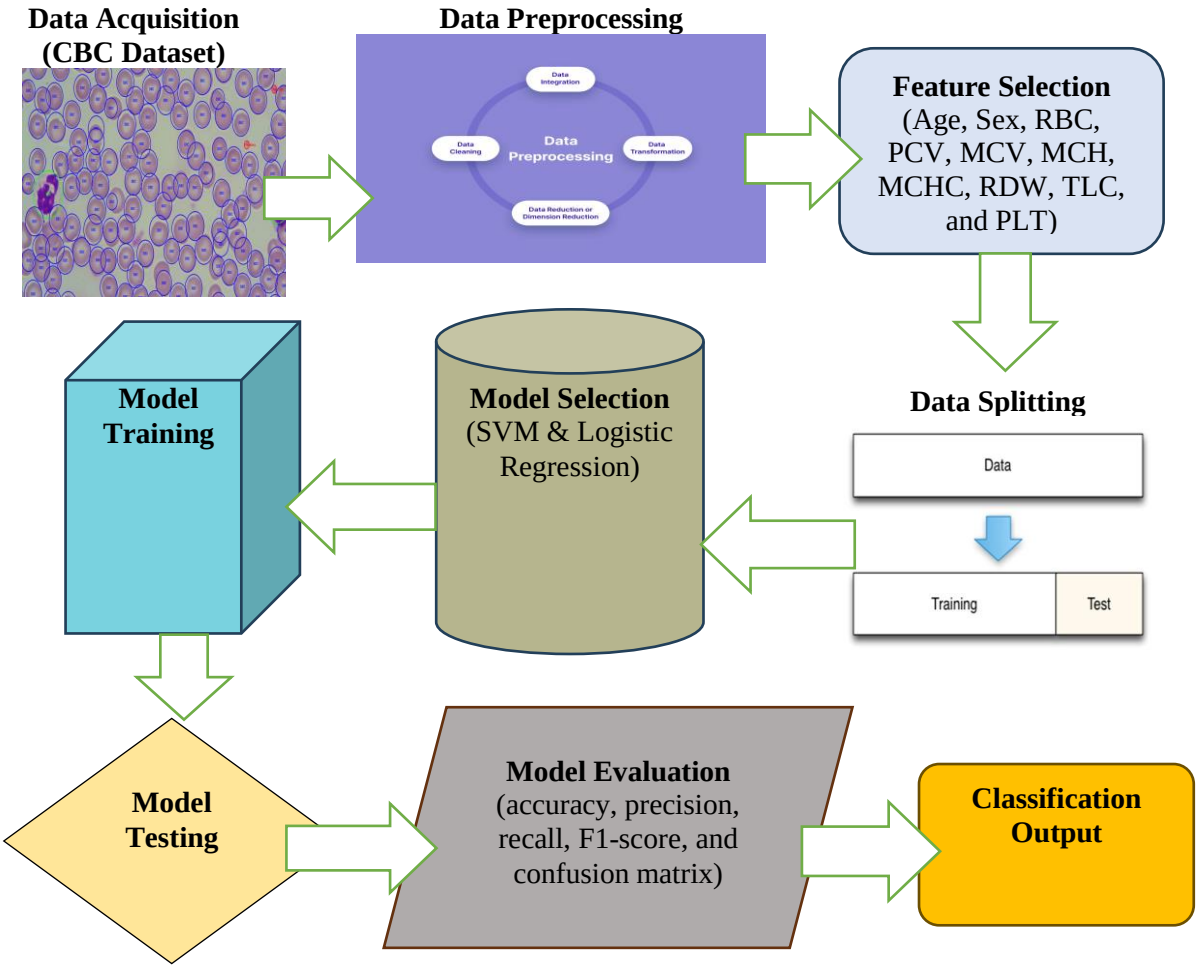

Figure 1: Classification Architecture

Table 1: CBC Dataset Feature Description

S/N	Feature	Description	Data Completeness (%)
1	Age (years)	The chronological age of the individual in years.	100
2	Sex	The biological classification of the individual as either male or female.	100
3	Red Blood Count (RBC)	The number of red blood cells present per unit volume of blood, typically measured in millions per microliter.	100
4	Packed Cell Volume (PCV)	Also known as hematocrit, it represents the percentage of blood volume occupied by red blood cells.	100
5	Mean Corpuscular Volume (MCV)	The average volume of a single red blood cell, measured in femtoliters (fL).	100
6	Mean Corpuscular Haemoglobin (MCH)	The average amount of Haemoglobin in a single red blood cell, measured in picograms (pg).	100
7	Mean Corpuscular Haemoglobin Concentration (MCHC)	The average concentration of Haemoglobin in each volume of packed red blood cells, measured in g/dL.	100
8	Red Cell Distribution Width (RDW)	A measure of the variation in red blood cell size, expressed as a percentage.	100
9	Total Leukocyte Count (TLC)	The total number of white blood cells in each volume of blood, measured in thousands per microliter ($\times 10^3/\mu L$).	100
10	Per Cubic Millimetre (PLT /mm3)	The platelet count, indicating the number of platelets per cubic millimeter of blood.	100
11	Haemoglobin (HGB)	The concentration of Haemoglobin in the blood, measured in grams per deciliter (g/dL), which is crucial for oxygen transport.	100

3.2 Data Preprocessing and Feature Selection

In preparing the dataset for machine learning functionalities, the sex feature, with categorical variables were transformed into numerical values using One-Hot Encoding (OHE), ensuring that qualitative features were represented in a binary format without imposing an artificial ordinal relationship [11,12,24]. SMOTE analysis was also employed to handle class imbalance to improve model generalizability.

Also, HGB levels, originally recorded as a continuous numerical variable within the range of 4.2 g/dL to 19.6 g/dL, were discretized to enhance clinical interpretability and model performance. The discretization process was conducted using a binning technique based on established medical thresholds for haemoglobin classification [12,24]. The HGB level was mapped using clinically relevant metrics such that Low is (0–12 g/dL), Normal is (12–16 g/dL), and High is (>16 g/dL). This was done using the binning which partitions the range into predefined bins and assigns discrete labels accordingly as shown below:

Mathematically, the binning function can be expressed as:

$$f(HGB) = \begin{cases} 0, if\ HGB < 12\ g/dL \\ 1, if\ 12 \leq HGB \leq 16\ \frac{g}{dL} \\ 2, if\ HGB > 16\ g/dL \end{cases} \quad (1)$$

Where *f(HGB)* represents the discretized haemoglobin category. This technique was tailored to mirror real-world medical diagnostic ranges in alignment with the classification models to ensure realistic results [24]. This will

improve clinical relevance and model's predictive capabilities.

Feature Selection Technique

All feature (AF), and literature selected features (SF) sets were used to ascertain the significance of features on model's performance and for comparison purposes. Clinical and hematological features that are important for predicting HGB levels were tagged as SF based on literature [12]. This informed the study to select features like age, RBC, PCV, MCV, and HGB as key predictors that are critical in the diagnosis of conditions like anaemia, including sickle cell anaemia.

Data Splitting

To avoid model train-test bias and maintain a balanced distribution of HGB level dataset between the train and test sets, this study employed a stratified sampling techniques and the dataset was split into 70:30 percentile [8, 12].

$$n_h = \frac{N_h}{N} \times n. \quad (2)$$

Where n_h is the sample size for stratum h, N_h is the population size for stratum, N is the population size and n are the desired sample size. This ensures that each subgroup (stratum) is proportionally represented in the sample, maintaining the distribution of the population.

3.3 Model Development and Implementation

Four models were built, namely SVM using the entire features, SVM using selected features, LR using the entire features and LR using selected features. The SVM and LR models contributes uniquely to the overall performance of HGB level classification, ensuring a robust and well-balanced predictive approach as shown in Figure 2. In the context of this study, SVM RBF kernel was employed to handle non-linear dataset attributes. While the LR model, L1 regularization was employed. RF n-estimators was set, and maximum tree was used. LR and SVM were precisely selected based on their efficacy in handling medical data classification tasks [19, 24]. SVM adapts well to datasets with high dimensionality and good in capturing nonlinear relationships within datasets especially when using kernel function [19]. However, LR is a simple but powerful model that uses statistical probabilistic interpretation ability to predict results. The feature selection techniques were also added to understand how different features contribute to HGB levels prediction. Furthermore, the models were trained independently to allow their individual strengths contribute to a reliable and well-balanced classification system.

3.4 Evaluation Metrics

The following performance metrics were used to access the performance of the two ML models:

1. **Accuracy** was used to determine the model's total correct prediction in relation to the actual values. This is paramount to ensure efficient classification task.

$$Accuracy = \frac{TP+FP}{TP+TN+FP+FN} \quad (3)$$

2. **Precision** was used to determine the proportion of true positive predicted correctly and evaluate the model's ability to minimize false positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

3. **Recall also known as sensitivity** was used was employed to measure the proportion of actual positive cases that were correctly identified by the mode.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

4. **F1-Score** was employed to measure the balance between precision and recall especially when data imbalance is present.

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

5. **AUC (Area Under the Curve)** was employed to measure the developed model's ability to differentiate between positive and negative predictions.

$$FPR = \frac{FP}{FP+TN} \quad (7)$$

Note: *TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative and FPR is False Positive Rate for the equations above.*

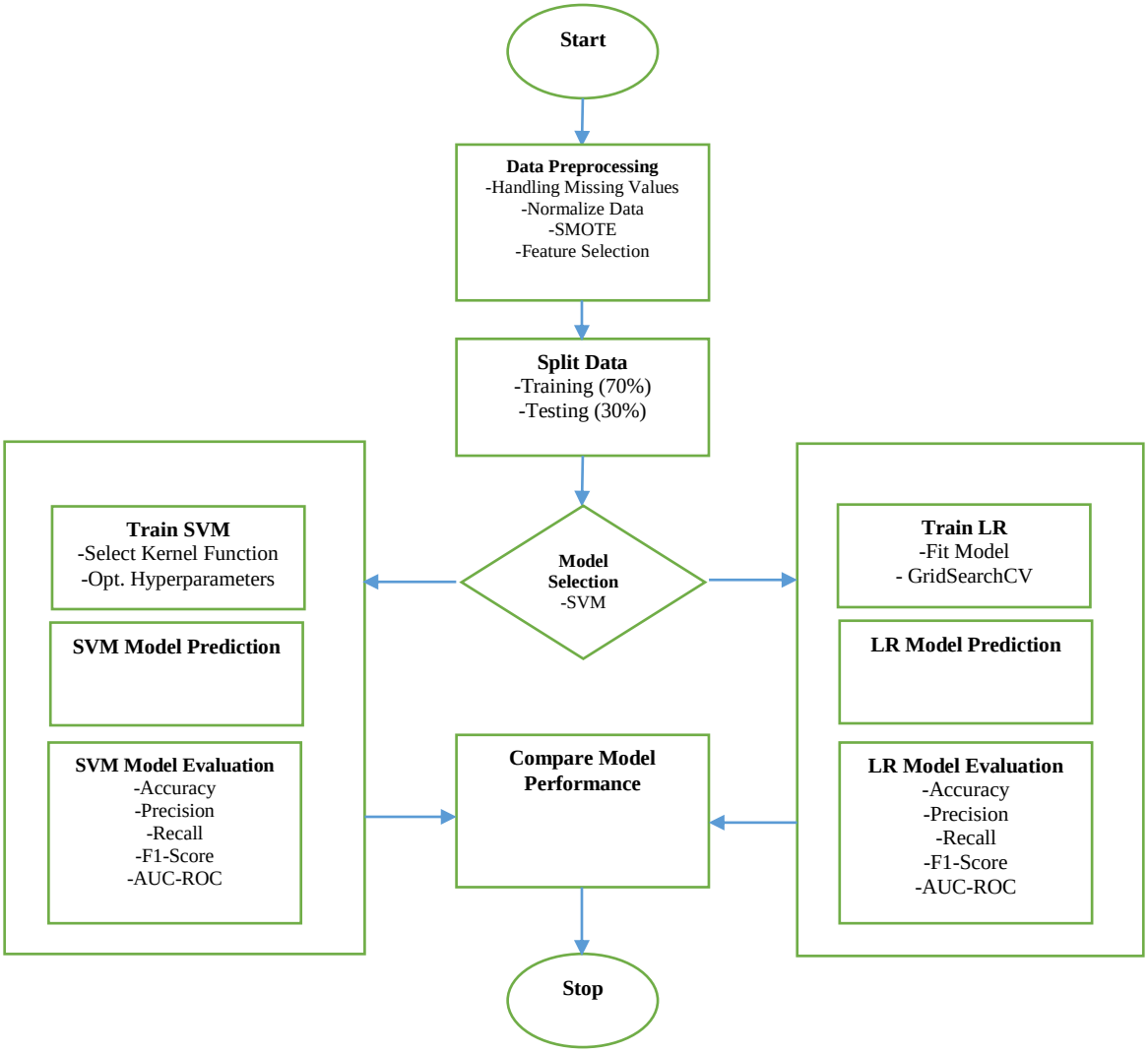

Figure 2: SVM-LR Sickle Cell Anaemia HGB Level Classification Flowchart

4. Results and Discussion

4.1 Results

Two ML models namely, Logistic Regression (LR) and Support Vector Machine (SVM) were used for the HGB level classification. Also, both models were trained using All Feature (AF) and Selected Feature (SF) sets. Evaluation metrics such as accuracy, precision, recall (sensitivity), f1-score and AUC-ROC were used to access the ML models performance.

The results in *Table 2* shows that when both models were evaluated using SF, SVM demonstrated superior performance an accuracy of 84.90%, precision of 73.90%, recall of 77.30%, F1-score of 75.60% and AUC-ROC of 93.40%. However, LR performed poorly under the same experimental condition with an accuracy of 79.50%, precision of 65.20%, recall of 68.20%, F1-score of 66.70%, and AUC-ROC of 90.90%.

Furthermore, when both models were evaluated using AF, they had an improved classification performance with LR coming out top as opposed to the results when using SF. LR had an accuracy of 93.20%, precision of 90.50%, recall of 86.40%, F1-score of 84.40% and AUC-ROC of 98.90%. Also, SVM experienced a notable increase in performance with an accuracy of 91.80%, precision of 90.00%, recall of 81.00%, F1-score of 85.70% and AUC-ROC of 98.20%

These findings showed that SVM had a strong classification performance on both SF and AF sets. However, LR showed better performance all when AF were used. This posits that LR

model has pronounced dependency on availability of inexhaustive feature sets.

The ability of SVM to maintain a good classification performance under reduced number of features connotes its robust and effective characteristics to offer desired results when faced with computational constraints and limited data scenarios. However, the rise in performance of LR when all feature sets were introduced showed its optimal ability to explore high dimensional spaces where it can leverage on richer patterns and interactions across wide range of variables. This observation suggested that LR is highly effective and reliable for clinical settings, where a wide range of haematological and demographic features contribute significantly to disease classification.

The results also gave a clear and precise description of the influence of feature selection techniques on the model's predictive performance in classifying SCA haemoglobin level. It was further observed that HGB distribution in Figure 4 showed that Low HGB cases were predominant as compared to High HGB cases. This informed the study, to handle class imbalance using SMOTE analysis so that both models can effectively detect accurate percentages of both cases.

This context made the analysis in Table 2 particularly important, as it evaluates how each model performs across a range of metrics (accuracy, precision, recall, F1-score, and AUC-ROC) under different feature configurations. High precision and recall for minority classes, for instance, suggested a model's robustness in identifying patients at risk. Therefore, the results in Table 2 did more than rank models but highlighted how well each model addressed real-world challenges of imbalanced data in a clinically meaningful classification task. Also, these findings emphasized the necessity of rigorous feature selection and extraction techniques when developing machine learning models, positing that well-engineered features can substantially improve classification accuracy and overall model robustness.

4.2 Discussion

The sole aim of this study was to develop a state-of-the-art ML model for classifying HGB level in SCA patients because accurate diagnosis and severity level classification of SCA ensures that patients are offered timely interventions, optimized treatment plan and personalized treatment strategies. To achieve these, two ML models namely LR and SVM were implemented using all feature and selected feature selection techniques to evaluate their efficacy in classifying various levels of HGB in SCA patients. Also, a comparison of both models was carried out to ascertain the best model and feature selection approach suitable for SCA timely diagnosis and monitoring.

The results obtained from our study showed that LR outperformed SVM in its accuracy and AUC-ROC metrics when all feature sets were used. Also, it was noticed that both models performed exceptionally well when trained using all feature sets when compared with using the literature selected feature sets. LR achieved the highest classification accuracy of 93.20% and AUC-ROC of 98.90% when all features were used, surpassing the 91.80% accuracy and AUC-ROC of 98.20% attained by SVM under similar conditions. However, with selected feature sets, both LR and SVM had an accuracy of 84.90%, and 79.50% respectively. The disparities in results while using all feature sets and selected feature sets suggested that the models perform better when been trained with a broader features spectrum.

The optimal performance displayed by LR model in our study can be attributed to the ability of the model to adapt to both non-linear and linear relationships especially where there is high dimensionality in feature spaces. Although SVM has always shown strong performance in diverse ML classification tasks but its reliance on the radial basial functions might have hindered its efficiency when dealing with datasets that possess intricate relationships. It could also be deduced that there was significance improvement in both models when all feature sets were introduced. This, however, suggests that for the dataset used feature selection techniques could have hampered the performance of both models because there may have been exclusion of variables useful for a model superior performance.

Several studies have used ML models in predicting the level of haemoglobin in patients with SCA which is in line with this study. The first work by Oikonomou *et al.* [25] experimented on using genetic biomarkers lime BCL11A, Xmm1-HGB2 and HBS1L-MYB to predict the percentage of Haemoglobin patients with SCA. However, the model was only able to predict a small size of clinical trials, but our model aims to classify SCA Haemoglobin level across broader spectrum.

Table 2: Classification Report Breakdown

Support Vector Machine		
	SF	**AF**
Accuracy (%)	84.90	91.80
Precision (%)	73.90	90.00
Recall (%)	77.30	81.80
F1-Score (%)	75.60	85.70
AUC-ROC (%)	93.40	98.20
Logistic Regression		
	SF	**AF**
Accuracy (%)	79.50	93.20
Precision (%)	65.20	90.50
Recall (%)	68.20	86.40
F1-Score (%)	66.70	84.40
AUC-ROC (%)	90.90	98.90

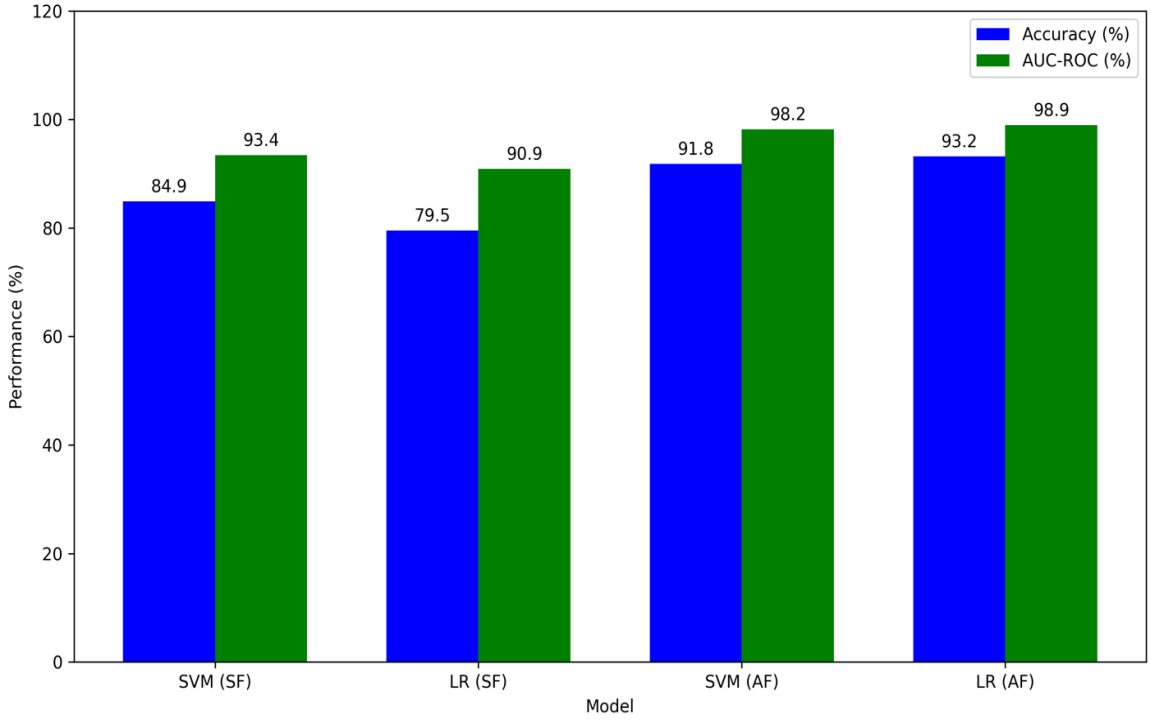

Figure 3: SVM-LR Performance Comparison

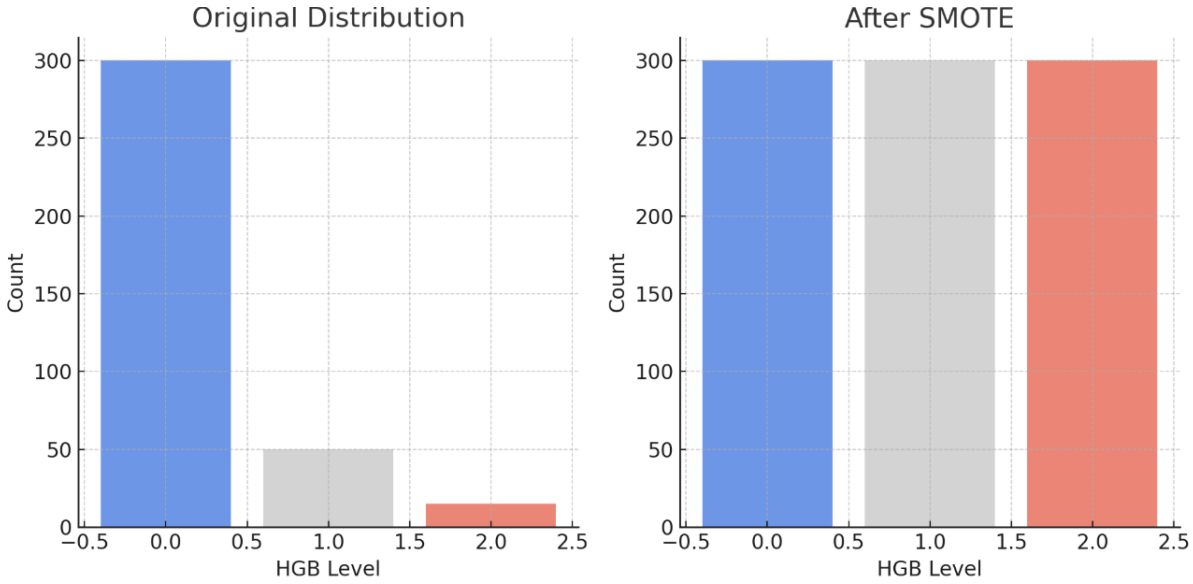

Figure 4: HGB Level Distribution Before and After SMOTE Analysis

Another study by Odigwe *et al.* [26] used artificial neural networks to predict a patient's response to hydroxyurea therapy, a treatment that elevates haemoglobin levels. The model achieved a high accuracy of 92.6% in predicting HbF levels post-treatment. Unlike this targeted therapy response prediction, our model focuses on general severity classification using a broader feature set and dataset.

The third study was HgbNet developed by Zhi *et al.* [27], a model leveraging electronic health records (EHRs) to predict haemoglobin levels and anaemia severity was developed. By handling missing values and using attention mechanisms, HgbNet provided a robust method for anaemia diagnosis. While this study emphasized feature engineering for EHR-based predictions, our work explores the impact of feature selection on model performance, demonstrating that using all available features significantly improves classification accuracy.

This study contributes to the present body of knowledge by effectively evaluating the effect of using feature selection techniques on model's predictive performance when compared to previous research studies. The results indicate that models using all features (AF) significantly outperform those with selected features (SF), with LR (AF) achieving an accuracy of 93.20% and AUC-ROC of 98.90% and SVM (AF) reaching 91.80% accuracy and AUC-ROC of 98.20%. This reinforces the importance of comprehensive feature selection in improving classification accuracy for HGB level classification.

The significant benefit of this study was that it explored the use of all feature sets and literature selected feature which served as a benchmark for comprehensive comparison between ML models. Also, the study took into cognizance the advantages of HGB level distribution for cross evaluation of true positive and false negative rates, which were important for carrying efficient diagnosis of SCA. Utilizing the above techniques, ensured that our findings were reliable and clinically applicable for SCA real-world problem management.

Despite the above-mentioned strengths of this study, it still possesses some limitations. Though the dataset used was adequate, there is the possibility that it might not fully represent the variability found amongst diverse SCA patient population. Future work should focus on incorporating independent datasets to the framework as an external validation for model generalizability. Although, LR showed strong classification prowess, exploring other ML algorithms or ensemble models could further broaden the scope of the study, improve prediction capability and accuracy by capturing complex relationships within diverse dataset [28].

5. Conclusion

This study highlights ML potential in SCA-HGB level classification. LR performed best with the full feature set, highlighting comprehensive data's role in its accuracy. Also, the findings of this study, will contribute to the body of knowledge in the field of haematology as regards improving the diagnosis and management of SCA. Future research should explore deep learning techniques, alternative feature selection such as ANOVA, Chi-Square and the likes, exploring diverse datasets and extending the dataset by collecting more samples alongside augmentation techniques to enhance generalizability of the model and its applicability in clinical settings.

Acknowledgement

It is with deepest gratitude that I thank the staff of the Faculty of Computing at the University of Ibadan for their unwavering support and guidance throughout this project. I would like to also express my deep appreciation for the insightful discussions and collaborative efforts of my mentors, supervisors and colleagues. Furthermore, I am indebted to the medical professionals, including haematologists, laboratory technicians, and clinicians, whose expertise supported the study's medical context. I also want to appreciate DATICAN (www.datican.org) for their financial assistance, that made this study was possible. Finally, I would like to thank the Kaggle repository for providing access to the dataset used in this study.

References

[1] Johnston, J. D., Reinman, L. C., Bills, S. E., & Schatz, J. C. (2022). Sleep and fatigue among youth with sickle cell disease: A daily diary study. Journal of Behavioural Medicine, 1-11.

[2] Luo, L., King, A. A., Carroll, Y., Baumann, A. A., Brambilla, D., Carpenter, C. R., Colla, J., Gibson, R. W., Gollan, S., Hall, G., Klesges, L., Kutlar, A., Lyon, M., Melvin, C. L., Norell, S., Mueller, M., Potter, M. B., Richesson, R., Richardson, L. D., Ryan, G., Siewny, L., Treadwell, M., Zun, L., Armstrong-Brown, J., Cox, L. & Tanabe P. (2021). Electronic Health Record-Embedded Individualized Pain Plans for Emergency Department Treatment of Vaso-occlusive Episodes in Adults with Sickle Cell Disease: Protocol for a Pre-implementation and Postimplementation Study. JMIR Res Protoc. 2021 Apr 16;10(4).

[3] Bou-Fakhredin, R., De Franceschi, L., Motta, I., Cappellini, M. D., & Taher, A. T. (2022). Pharmacological Induction of Fetal Haemoglobin in β-Thalassemia and Sickle Cell Disease: An Updated Perspective. Pharmaceuticals, 15(6), 753.

[4] Acharya, B., Mishra, D. P., Barik, B., Mohapatra, R. K., & Sarangi, A. K. (2023). Recent progress in the treatment of sickle cell disease: an up-to-date review. Beni-Suef University Journal of Basic and Applied Sciences, 12(1), 38

[5] Tebbi, C. K. (2022). Sickle cell disease, a review. *Hemato*, *3*(2), 341-366.

[6] Ramachandran, P., Perisetti, A., Kathirvelu, B., Gajendran, M., Ghanta, S., Onukogu, I., Lao, T. & Anwer, F. (2020). Low Morbidity and Mortality with COVID-19 in Sickle Cell Anaemia: A Single Centre Experience. eJHaem, 1, 608-614.

[7] Obeagu, E. I., Adias, T. C., & Obeagu, G. U. (2024). Advancing life: innovative approaches to enhance survival in sickle cell anaemia patients. *Annals of Medicine and Surgery*, *86*(10), 6021-6036.

[8] Olatunji, S. O., Khan, M. A. A., Alanazi, F., Yaanallah, R., Alghamdi, S., Alshammari, R., ... & Ahmed, M. I. B. (2024). Machine Learning-Based Models for the Preemptive Diagnosis of Sickle Cell Anaemia Using Clinical Data. In *Finance and Law in the Metaverse World* (pp. 101-112). Springer, Cham.

[9] Muhsen, I. N., Shyr, D., Sung, A. D., & Hashmi, S. K. (2021). Machine learning applications in the diagnosis of benign and malignant haematological diseases. *Clinical Haematology International*, *3*(1), 13-20.

[10] Santos-Silva, M. A., Sousa, N., & Sousa, J. C. (2024). Artificial intelligence in routine blood tests. *Frontiers in Medical Engineering*, *2*, 1369265.

[11] Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, *15*(4), 235.

[12] Nenova, Z., & Shang, J. (2022). Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics. *Production and Operations Management*, *31*(1), 259-280.

[13] Yıldız, T. K., Yurtay, N., & Öneç, B. (2021). Classifying anaemia types using artificial learning methods. *Engineering Science and Technology, an International Journal*, *24*(1), 50-70.

[14] Zhu, J., Sun, R., Liu, H., Wang, T., Cai, L., Chen, Z., & Heng, B. (2023). A Non-Invasive Haemoglobin Detection Device Based on Multispectral Photoplethysmography. *Biosensors*, *14*(1), 22.

[15] Raza, A., Eid, F., Montero, E. C., Noya, I. D., & Ashraf, I. (2024). Enhanced interpretable thyroid disease diagnosis by leveraging

synthetic oversampling and machine learning models. *BMC Medical Informatics and Decision Making*, *24*(1), 364

[16] Bakır, H., & Ceviz, Ö. (2024). Empirical enhancement of intrusion detection systems: a comprehensive approach with genetic algorithm-based hyperparameter tuning and hybrid feature selection. *Arabian Journal for Science and Engineering*, *49*(9), 13025-13043.

[17] Bhatia, M., Meena, B., Rathi, V. K., Tiwari, P., Jaiswal, A. K., Ansari, S. M., ... & Marttinen, P. (2023). A novel deep learning-based model for erythrocytes classification and quantification in sickle cell disease.

[18] Srivastava, S., Srinivasan, R., Nambison, N. K., & Gorthi, S. S. (2021). Diagnosis of sickle cell anaemia using AutoML on UV-Vis absorbance spectroscopy data.

[19] Ekong, B., Ekong, O., Silas, A., Edet, A. E., & William, B. (2023). Machine Learning Approach for Classification of Sickle Cell Anaemia in Teenagers Based on Bayesian Network. *Journal of Information Systems and Informatics*, *5*(4), 1793-1808.

[20] Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., & Duan, Y. (2020). Deep Learning Models for Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anaemia Diagnosis. *Electronics*, *9*(3), 427.

[21] Dada, E. G., Oyewola, D. O., & Joseph, S. B. (2022). Deep convolutional neural network model for detection of sickle cell anaemia in peripheral blood images. *Communication in Physical Sciences*, *8*(1).

[22] Zemariam, A. B., Yimer, A., Abebe, G. K., Wondie, W. T., Abate, B. B., Alamaw, A. W., Yilak, G., Melaku, T. M., & Ngusie, H. S. (2024). Employing supervised machine learning algorithms for classification and prediction of anaemia among youth girls in Ethiopia. *Scientific reports*, *14*(1), 9080.

[23] Ramzan, M., Sheng, J., Saeed, M.U. *et al.* Revolutionizing anaemia detection: integrative machine learning models and advanced attention mechanisms. *Vis. Comput. Ind. Biomed. Art* **7**, 18 (2024).

[24] Faye, L. M., Magwaza, C., Dlatu, N., & Apalata, T. (2025). Exploring Determinants and Predictive Models of Latent Tuberculosis Infection Outcomes in Rural Areas of the Eastern Cape: A Pilot Comparative Analysis of Logistic Regression and Machine Learning Approaches. *Information*, *16*(3), 239.

[25] Oikonomou, K., Steinhöfel, K., & Menzel, S. (2021, September). A Machine Learning Model for Predicting Fetal Haemoglobin Levels in Sickle Cell Disease Patients. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1* (pp. 79-91). Singapore: Springer Singapore.

[26] Odigwe, B. E., Eyitayo, J. S., Odigwe, C. I., & Valafar, H. (2019). Modelling of sickle cell anaemia patients' response to hydroxyurea using artificial neural networks. *arXiv preprint arXiv:1911.10978*.

[27] Zhi, Z., Elbadawi, M., Daneshmend, A., Orlu, M., Basit, A., Demosthenous, A., & Rodrigues, M. (2024). HgbNet: predicting Haemoglobin level/anaemia degree from EHR data. *arXiv preprint arXiv:2401.12002*.

[28] Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, *36*(21), 12655-12699.