Nigerian Journal of Technology (NIJOTECH)
Vol. 41, No. 4 July, 2022, pp.789 –796
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
http://dx.doi.org/10.4314/njt.v41i4.16

A Comparative Analysis of Haemoglobin Variants using Machine Learning Algorithms

A. A. Okandeji[1, *], O. F. Odeyinka[2], A. A. Sogbesan [3], N. O. Ogunye[2]

[1]*Department of Electrical and Electronics Engineering, University of Lagos, Akoka, Lagos State, NIGERIA.*
[2]*Department of Systems Engineering, University of Lagos, Akoka, Lagos State, NIGERIA.*
[3]*Department of Electrical/Electronic Engineering. DS Adegbenro ICT Polytechnic, Itori, Ogun State, NIGERIA*

Article history: Received 25 April 2022; Revised form 14 June 2022; Accepted 5 July 2022; Available online 10 September 2022

Abstract

In medical sciences, to ascertain the origin of a sickness, professionals utilize their expertise and knowledge to analyze a person's symptoms and indications. These symptoms (indicators) are threshold values that health specialists use to determine the cause of the illness by comparing a specific proportion of measurements to where a healthy population would fall. Consequently, diagnostic mistakes occur as a result of inaccuracy and imprecision. This study utilizes machine learning to categorize haemoglobin variations. Specifically, the data set used in this study includes 752 complete blood count laboratory analyses of adult patients aged eighteen and above obtained from Lagos State University Teaching Hospital (LASUTH). Multiple machine learning methods were utilized for classification from which five of the methods employed were examined and assessed. Comparative analysis was done using the five algorithms (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB)). Contrary to work done by previous researchers, it was observed that the SVM model showed the best classification accuracy of 94.7%, with an F1-score of 94.5%, precision of 94.8%, recall of 94.7%, specificity of 97.3%, and area under curve (AUC) of 99.0%. Among the other models considered, the RF model gave the least accuracy result of 87.4%. The study shows that the support vector machine algorithm outperforms the other classifiers in terms of accuracy when predicting haemoglobin variants given the haematological parameters.

Keywords: Haemoglobin variant, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Random Forest

1.0 INTRODUCTION

Until recently, machine learning was applicable in genomics and molecular biology, a terminology known as next-generation sequencing. The incorporation of clinical radiologic and genomic data based on deep learning, a subfield of machine learning that is based on neural networks comprising recognition method in the nested layer of networks, is thought to be useful in the advancement of pathology to accurately diagnose diseases and predict patient prognosis [1].

Genomics is a branch of molecular biology focused on studying all aspects of the genome on a complete set of genes within a particular organism [2]. Pathology can be divided into anatomic pathology, clinical pathology, and molecular pathology.

Clinical pathology is a medical specialty that deals with the analysis of body fluids in the laboratory such as blood, urine, spinal cerebrospinal fluid, and body tissues. Some sub-specialties of clinical pathology include chemical pathology, immunology, and haematology. Gunčar et al. [3] describe medical diagnosis as the method of deciding which disease better describes the symptoms and signs of an individual. Physicians utilize their medical knowledge, abilities, and experiences with laboratory analysis, which measures various elements in the blood, to find the diseases that best explain an individual's symptoms and signs [4]. A patient's medical history is indispensable in gathering information and data for diagnosis. Laboratory tests are also used to classify diseases and assess them to guide therapies. Nonetheless, laboratory test results are often underestimated because clinical laboratories tend to report test results as individual numerical or categorical values, and Doctors focus on those values that fall outside a given reference range [3]. Up to 70% of all medical decisions are based on laboratory tests [5]. Furthermore, diagnostic errors contribute to about 10% of patients' death and close to 17% of hospital complications [6]. To address these challenges, many researchers and organizations are working on introducing machine learning to diagnostic applications to

*Corresponding author **(Tel:** +234 (0) 9039383220**)**
Email addresses: aokandeji@unilag.edu.ng (A.A. Okandeji), *Olumide.odeyinka@unilag.edu.ng* (O. F. Odeyinka), *sogbesanadebiyi@gmail.com* (A.A. Sogbesan), and *nathaniel.ogunye@live.unilag.edu* (N. O. Ogunye)

improve the speed and accuracy of diagnosis which will produce results that are similar to previous applications of diagnoses as well as improve the reliability of the process [7].

Innovations and advancements in technology, machine vision, and other machine learning technologies have been designed to replace the efforts traditionally left only to pathologists with microscopes [8]. Since the way pathologists diagnose diseases, relying on manual observation of images under the microscope or the translation of colors into concentration, has remained unchanged for over a century. The interest of this study therefore lies within the field of haematology i.e., classification of haemoglobin variants using machine learning.

Several studies have been carried out using machine learning algorithms in medical diagnosis or for detecting blood-related disorders. Ayyıldız and Tuncer [9] proposed a way of diagnosing Iron Deficiency Anemia (IDA) and β-thalassemia using red blood cell indices and machine learning techniques such as Support Vector Machine (SVM) and k-Nearest Neighbor. They conducted a comparative evaluation of both algorithms to determine their effectiveness and concluded that both algorithms had exceptional performance. Additionally, their study revealed that complete blood count (CBC) parameters were effective in discriminating between IDA and β-thalassemia for patients. Oikonomou et al. [10] built a model that can predict the percentage fatal haemoglobin (HbF%) of patients. The authors explored and compared the accuracy of various machine learning algorithms like Decision Tree, Gradient Boosted Trees, Linear Regression, K-Nearest Neighbors, Neural Network, Random Forest, and Gaussian Process. The dataset for the study contained 465 patients of which 63.66% of the data set were used as the training set and 36.34% were the evaluation set. The results obtained showed that the K-Nearest Neighbors algorithm has the best performance with an accuracy of 87.25% and a mean error of 33.33%. El-kenawy et al. [11] conducted a study on estimating the haemoglobin level of mild and normal COVID-19 patients. In the study, the authors analyzed five different machine learning models for regression which are Artificial Neural Network, Support Vector Machine, Random Forest, Average Ensemble and K-Nearest Neighbors and recommended that machine learning should be used to estimate clinical test criteria rather than using trial and error to estimate the clinical test result. Yıldız et al. [12] constructed a model using four different machine learning algorithms which are Artificial Neural Networks, Support Vector Machine, Naive Bayes, and Decision Tree. The proposed model was evaluated with a dataset of 1663 samples and the accuracy of each model was determined. They concluded that the Decision Tree algorithm has the highest accuracy of 85.6%.

As surmised by Borah et al. [13], a large dataset will help in improving the accuracy of a model. Therefore, the use of machine learning approach in blood laboratory-based diagnosis could lead to a fundamental change in differential diagnosis and result in the modification of the currently accepted guidelines.

In contrast to existing result, this study aims to utilize machine learning to categorize haemoglobin variations. Specifically, the dataset used in this study includes 752 complete blood count laboratory analyses of adult patients aged eighteen and above. Using five methods namely (k-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB)), a comparative analysis was done to ascertain the best classification accuracy.

2.0 MATERIALS AND METHOD

2.1 Materials

The data in this study was obtained from the records of Lagos State University Teaching Hospital (LASUTH). The data was raw and cannot be used directly, but were manually gathered and organized into a spreadsheet.

The data collected contained haematological parameters that were used as relevant features to form the classifications. The features include Haemoglobin (Hb), Haematocrit (HCT), Red Blood Cells (RBC), White Blood Cells (WBC), Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), Mean Corpuscular, Haemoglobin Concentration (MCHC), Platelets Count (PLT), Packed Cell Volume (PCV) and Haemoglobin Genotypes variants as shown in Table 1. Also, from Figure 1, it is shown that the haemoglobin genotype variants used in this study are AA, AC, AS, SC, SS. The haemoglobin genotype variants were obtained from the analytical records of the Electrophoretic method using Helena Electrophoretic

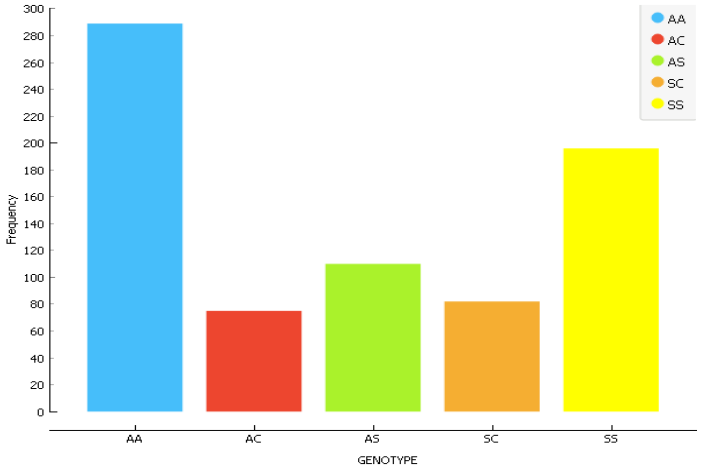


Figure 1: Distribution of target class for haemoglobin variants AA, AC, AS, SC, SS.

Machine [14], which has also been documented and cited in [15]-[20]. The recorded values of the haematological parameters were obtained from the automated analysis using Sysmex 5-part differential hematology analyzer. Figure 1 and Table 1, shows the distribution of the Hb variants.

Figure 1 shows the frequency of haemoglobin genotype classes with a total sum of 752 samples. It can be seen that the genotype AA has the highest frequency count of 289. Genotype AC has the least count with a total frequency count of 75, genotype SC slightly higher than genotype AC with a frequency count of 82. In contrast, genotype AS has a frequency count of 110, and finally, SS has the second highest with a count of 178.

Table 1: Haematological Parameters, Data Types and Genotype Class

Features	Hb	HCT	RBC	WBC	MCV	MCH	MCHC	PLT	PCV	HG
Description	g/dl	%	µ/dl	µl	fl	pg	g/dl	µl	%	Categorical code (1-AA, 2-AC, 3-AS, 4-SC, 5-SS)

2.2 Software

The software and tools used for this study include MATLAB R2019a, Go programming language (version go1.16.5), GoLearn library, Python (version 3.9.5), Scikit-learn (version 1.1.0) and Spreadsheet SaaS such as Google Sheets.

Google Sheets were used in collecting the raw data and converted it into meaningful information which was then exported as a comma-separated-values (CVS) file to be used as the dataset in the study. MATLAB was used for data analysis and data visualization. Go, with the aid of the GoLearn library, was used as the primary choice to develop the machine learning models, while Python and Scikit-learn were used to develop machine learning models that are not available in the GoLearn library.

2.3 Methods

2.3.1 Machine Learning Classification Algorithms

For effective analysis, multiple machine learning classification algorithms were used. Of all the algorithms used, only five were analyzed and evaluated. These five algorithms are as follows: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB). In general, the dataset is distributed in a multidimensional space. This space comprises of values that can be used for classification and are clustered in different regions due to their categorical difference.

Consequently, the distance between individual classes is small. KNN uses the distance function to calculate the distance between the test data and the training data, then votes for the most frequent or average category. In contrast, SVM separates distinct classes of data by creating a decision surface in a multidimensional space that includes the values of the features. DT uses a tree representation in which each leaf node corresponds to all possible solutions to a class label based on certain conditions. On the other hand, RF builds a large number of decision trees and merges them to get a more accurate and stable prediction. NB is based on the premise that a class's function is to forecast the values of features for its members.

The development, analysis, and evaluation of the classification models as illustrated in Figure 2 consist of the following stages:

(i) ***Data acquisition*****:** This includes acquiring dataset from the repository, spreadsheet, in such a way that it can be used.

(ii) ***Data preprocessing*****:** This refers to the processing involving data manipulation, sampling the data before it is used to enhance the performance. During this process, the data was split into the training data and test data.

(iii) ***Model Development:*** This process involves developing probabilistic models that best describes the relationship between features and classes. These models were already developed and provided by researchers through the open-source community.

(iv) ***Model Evaluation:*** This is an important aspect of the model creation process. It assists in determining the optimality of the model and how well that model will perform in the future.

(v) ***Hyperparameter optimization:*** This refers to the process of determining which hyperparameters for the model will produce the best results when tested on a validation set.

(vi) ***Analysis and summary*****:** This involve data analysis and summary of the model evaluation.

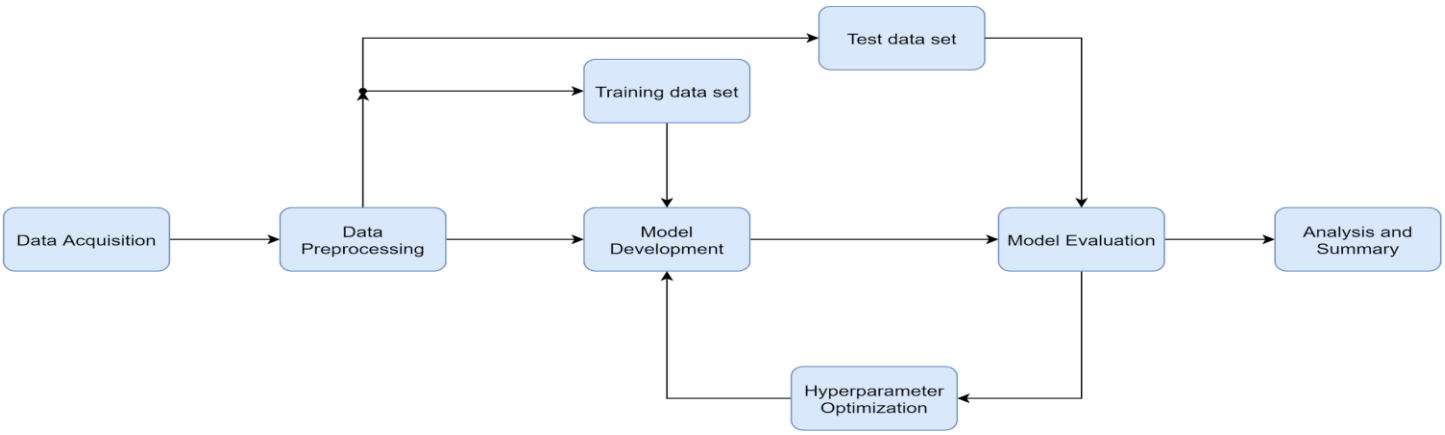


Figure 2: The methodology workflow

2.3.2 Model Performance Evaluation

Evaluation of the performance is an important part of the machine learning process. It is however a difficult task. As a result, it must be carried out with caution for the classification model to be reliable. To quantify model performance, model evaluation metrics are required. The evaluation metrics used are determined by the machine learning task at hand which in this case is classification. The performance metrics used in this study include the confusion matrix, the classification accuracy, precision, recall, specificity, F1 score and receiver operating characteristics (ROC) curve.

A confusion matrix, as shown in Figures 3-7, provides a breakdown of the correct and incorrect classification for each class. The confusion matrix shows the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) for each class.

Classification accuracy is the ratio of the total correct predictions made to the total of all predictions made.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision is the ratio of the total correctly predicted positives class to the total predicted positives.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall is the measure of our model correctly identifying true positives. It is also known as sensitivity.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Specificity is the metric that evaluates a model's ability to predict the true negatives of each available category.

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

F1 score is a measure of a model's accuracy on a data set. F1 scores measure the weighted average of precision and recall.

$$F1\ score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (5)$$

ROC is a performance metric that assesses the classifier's ability to distinguish between positive and negative classes. The ROC curve is a graph plotted against true positive rate (TPR) and false positive rate (FTR).

3.0 RESULTS AND DISCUSSION

Machine learning is a method for computing conceptual frameworks from large datasets generally in the form of an algorithm, and producing results that would be difficult for people to achieve due to the increasing data volume and complexity. The use of machine learning techniques in medicine for sorting and classifying health data is rapidly increasing.

The evaluation of the KNN model is shown in Table 2, and in Figs. 3-7. It can be observed that the KNN model had a Precision of 0.8787, Sensitivity of 0.8873, Specificity of 0.9660, Accuracy of 0.8763 and F1-Score of 0.8720. Table 2 shows the evaluation of the SVM model. It can be observed that the SVM model had a Precision of 0.8980, Sensitivity of 0.8905, Specificity of 0.9789, Accuracy of 0.9233 and F1-Score of 0.8941. Table 2 also shows the evaluation of the DT model. It can be observed that the DT model had a Precision of 0.8641, Sensitivity of 0.8496, Specificity of 0.9636, Accuracy of 0.8638 and F1-Score of 0.8508. In addition, Table 2 shows the evaluation of the RF model. It can be observed that the RF model had a Precision of 0.8286, Sensitivity of 0.8173, Specificity of 0.9582, Accuracy of 0.8472 and F1-Score of 0.8182.

Finally, Table 2 shows the evaluation of the NB model. It can be observed that the NB model had a Precision of 0.9020, Sensitivity of 0.8903, Specificity of 0.9717, Accuracy of 0.9004 and F1-Score of 0.8954.

Of the five-classification algorithm used, the support vector machine had the highest classification accuracy of 92.3%, while the random forest had the lowest classification accuracy of 84.7%. The classification accuracy of the other models, k nearest neighbors, decision tree, and naive Bayes, was 87.6%, 86.4%, and 90.0%, respectively, as shown in Table 2.

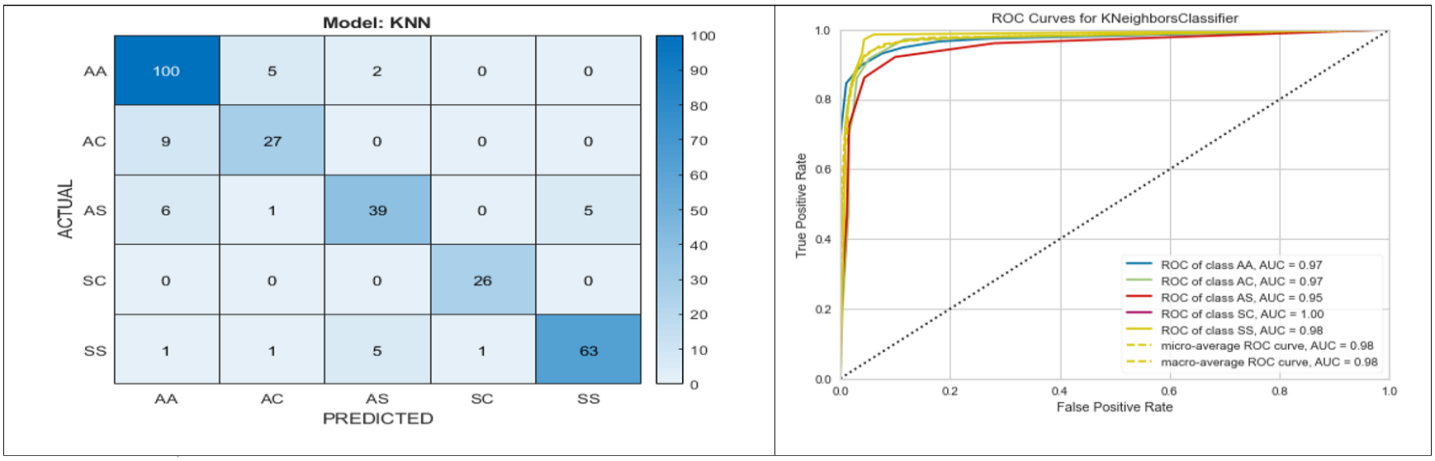
However, these results give an accurate and deep understanding of the performance of the models. It can be noted that the overall performance of the models depends on the size of the dataset and the distinction of features [13].

The study findings are promising and encourage additional research into the application of machine learning to the broader field of medicine.

Note that the area under curve (AUC) column of Table 2 is obtained from the macro average value ROC curve. This means that the probability that a randomly chosen positive instance (dataset) is ranked higher than a randomly negative instance.

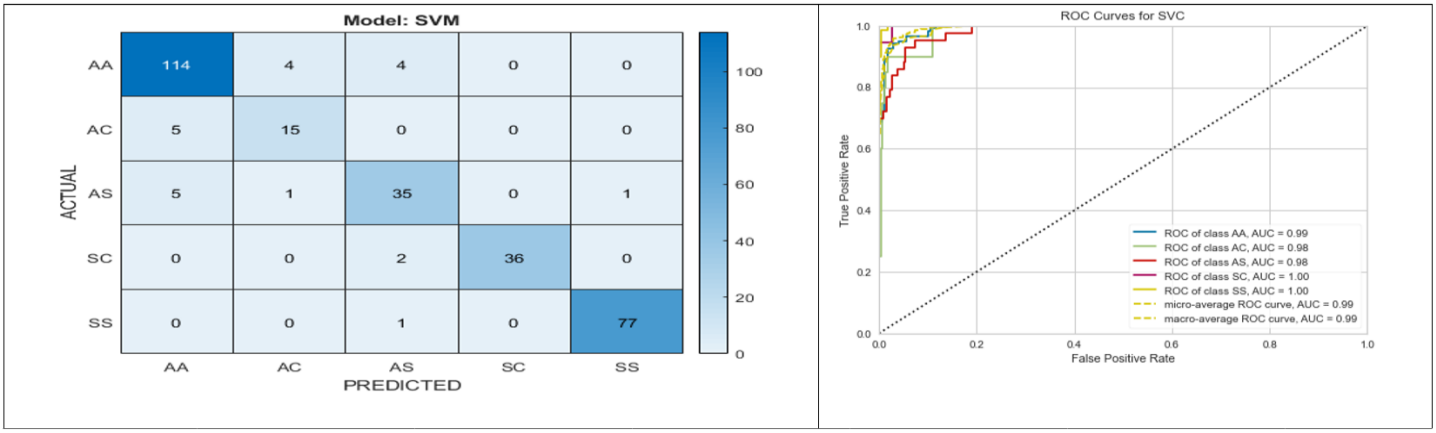
Table 2: Evaluation of the results of the models

Model	Area Under Curve	Accuracy	Precision	Recall	F1- Score	Specificity
KNN	0.980	0.876	0.879	0.867	0.867	0.966
SVM	0.992	0.923	0.898	0.891	0.894	0.979
DT	0.940	0.864	0.864	0.849	0.851	0.964
RF	0.950	0.847	0.829	0.817	0.818	0.958
NB	0.990	0.900	0.902	0.890	0.895	0.972



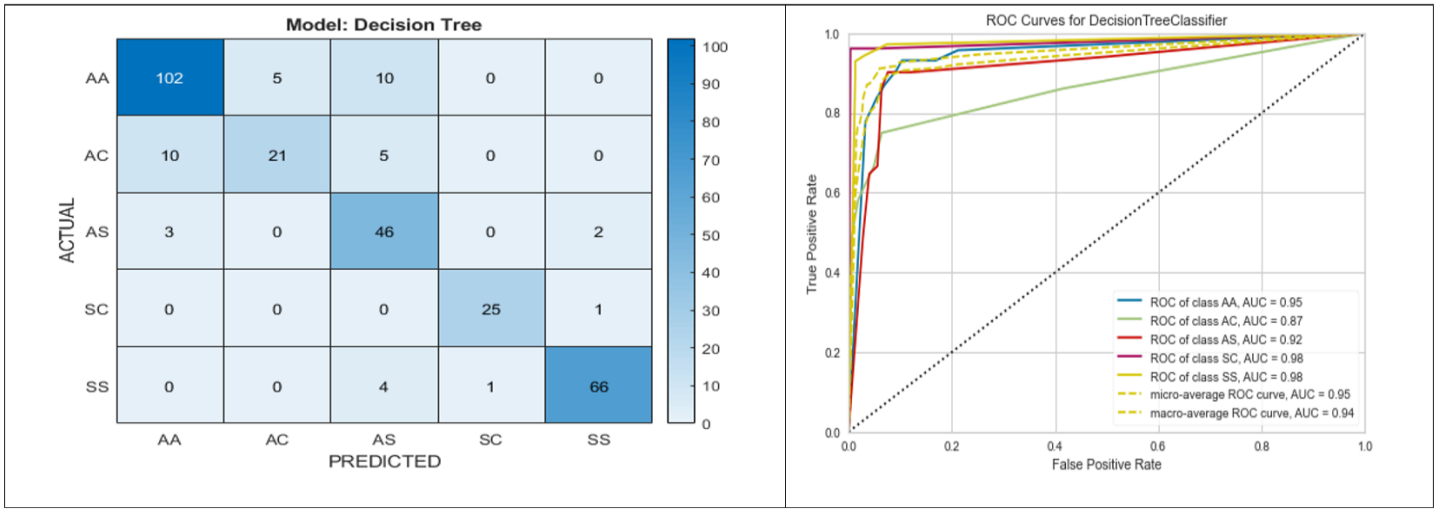
(a). Confusion matrix of the data set (b). ROC curve

Figure 3: Confusion matrix and ROC Curve of the KNN model



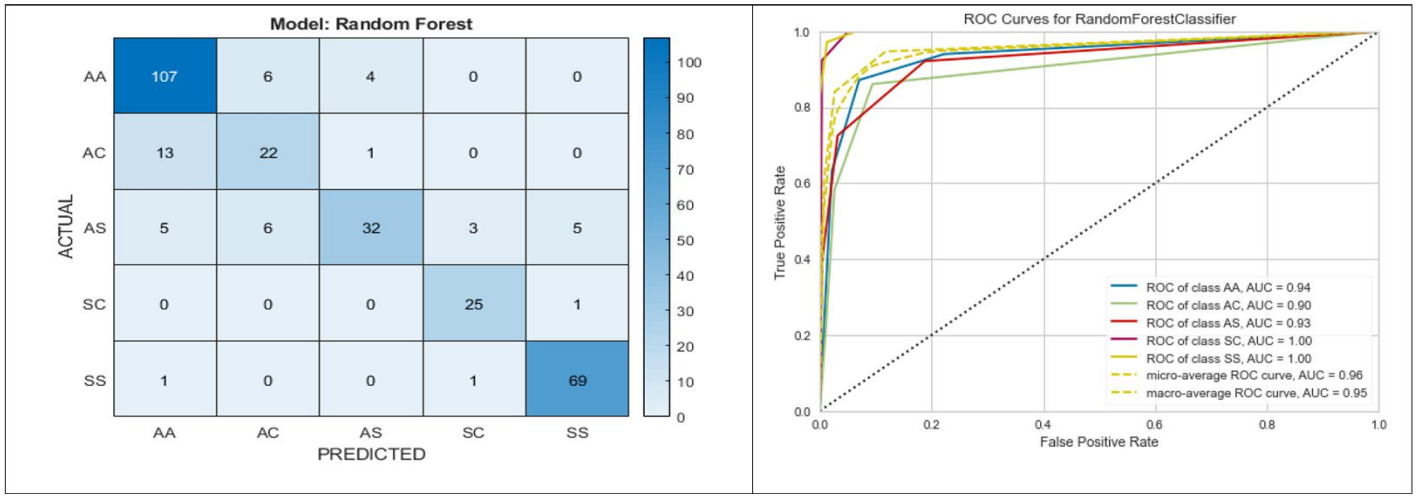
(a). Confusion matrix of the data set (b). ROC curve

Figure 4: Confusion matrix and ROC Curve of the SVM model



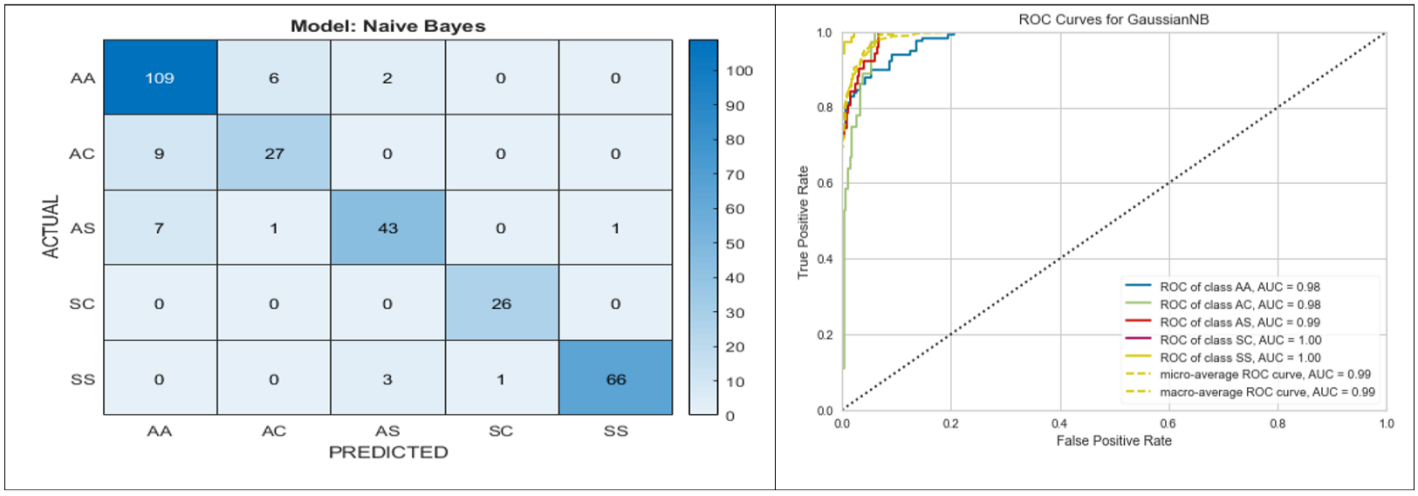
(a). Confusion matrix of the data set (b). ROC curve

Figure 5: Confusion matrix and ROC Curve of the Decision Tree model



(a). Confusion matrix of the data set (b). ROC curve

Figure 6: Confusion matrix and ROC Curve of the Random Forest model



(a). Confusion matrix of the data set (b). ROC curve

Figure 7: Confusion matrix and ROC of the Naive Bayesian model

4.0 CONCLUSION

Machine learning algorithms have a lot of potential in the classification of haemoglobin variants. Analysis of the results showed that the utilized algorithms succeeded admirably in certain circumstances while failing miserably in others. By analyzing the performance of the five different classifiers, it was concluded that the support vector machine model outperforms the other classifiers in terms of accuracy with a classification accuracy of 92.3% when classifying haemoglobin variants. The NB model had the second-highest classification accuracy of 90.0%, the KNN and DT model performed relatively close with a classification accuracy of 87.6%, 86.4%, respectively. The RF model had the least classification accuracy of 84.7%.

This study can be extended to other areas of studies such as in predicting the occurrence of blood diseases. Further improvements can also be larger data sets employing deep learning algorithms to make the predictions. The results obtained can be compared to classical machine learning methods such as linear regression, logistic regression, to ascertain better performance.

Funding opportunities
This research work has not received any funding.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Chang, H. Y., Jung, C. K., Woo, J. I., Lee, S., Cho, J., Kim, S. W., and Kwak, T. Y. "Artificial Intelligence in Pathology", *Journal of pathology and translational medicine*, 53(1), 2019, pp. 1–12.

[2] Lesk, A. M. Introduction to genomics. Oxford University Press, 2017.

[3] Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., and Notar, M, "An application of machine learning to haematological diagnosis, *Scientific reports,* 8(1), 2018, pp. 1-12.

[4] Cuomo, R., Cargiolli, M., Cassarano, S., Carabotti, M., and Annibale, B. Treatment of diverticular disease, targeting symptoms or underlying mechanisms, *Current Opinion in Pharmacology*, 43, 2018, pp. 124-131

[5] Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., and Green, R. "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods", *Academic Pathology*, 6, 2019, pp. 1-17. doi:10.1177/2374289519873088

[6] Cifra, C. L., Custer, J. W., Singh, H., and Fackler, J. C. "Diagnostic errors in pediatric critical care: a systematic review", *Pediatric Critical Care Medicine*, 22(8), 2021, pp. 701-712.

[7] Faggella, Daniel. "Machine Learning for Medical Diagnostics - 4 Current Applications." *Emerj Artificial Intelligence Research*, https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/. Accessed 29 Mar. 2022.

[8] El-kenawy, E. S. "A Machine Learning Model for Hemoglobin Estimation and Anemia Classification", *International Journal of Computer Science and Information Security (IJCSIS)*, 17(2), 2019, pp.100-108.

[9] Ayyıldız, H., and Tuncer, S. A. Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anaemia and beta-thalassemia via *Neighborhood Component Analysis Feature Selection-Based machine learning. Chemometrics and Intelligent Laboratory Systems*, 196, 2020, 103886

[10] Oikonomou, K., Steinhofel, K., and Menzel, S. "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. Springer-Verlag Berlin Heidelberg, 235, 2021, pp.79-91

[11] El-kenawy, E. S. M. T., Eid, M. M., and Ibrahim, A. "Anaemia estimation for covid-19 patients using a machine learning model", *Journal of Computer Science and Information Systems*, 2(1), 2021, pp.1-7.

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

[13] Borah, M.S., Bhuyan, B.P., Pathak, M.S., and Bhattacharya, P., "Machine Learning in Predicting Hemoglobin Variants", *International Journal of Machine Learning and Computing*, 8(2), 2018, pp. 140-143.

[14] Schneider, R.G. "Methods for Detection of Hemoglobin Variants and Hemoglobinopathies in the Routine Clinical Laboratory", *CRC Critical Reviews in Clinical Laboratory Sciences*, 9(3), 1978, pp. 243-271.

[15] Clark, B. E., Cunningham, J. C., Wild, B. J., and Thein, S. L. "A Tool for Predicting Haemoglobin Variants-an Aid to DNA Diagnostics", 104(11), 2004, pp. 3735.

[16] Rudra Kumar, M., Pathak, R., and Gunjan, V. K. "Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach", *In proc. Computational Intelligence in Machine Learning, Singapore, Springer*, 2022, pp. 123-133.

[17] Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1), 2019, pp.1-16.

[18] Joshi, K. K., Gupta, K. K., and Agrawal, J. "A Review on Application of Machine Learning in Medical Diagnosis. In proc. IEEE 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, Bhopal, India.

[19] Devanath, A., Akter, S., Karmaker, P., and Sattar, A. Thalassemia Prediction using Machine Learning Approaches. In proc. IEEE 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 1166-1174, Erode, India.

[20] Joshi, K. K., Joshi, N., and Chaudhari, R. R. "Machine Learning–Learning Techniques, CNN, Languages and APIs", *International journal of scientific research in computer science, engineering and information technology*. 6(3), 2020, pp. 23-30.