An Ensemble Model of Machine Learning Algorithms for the Severity of Sickle Cell Disease (Scd) Among Paediatrics

[1]Balogun Jeremiah Ademola, [2]Aderounmu Temilade, Egejuru Ngozi Chidozie and [1]Idowu, Peter Adebayo

[1]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

[2]Department of Paediatrics and Child Health, Obafemi Awolowo University Teaching Hospital Complex (OAUTHC), Ile-Ife, Nigeria.

Abstract

This study was motivated at developing an ensemble of 3 supervised machine learning algorithms for the assessment of the severity of sickle cell disease (SCD) among paediatric patients. The study collected data from a tertiary hospital in south-western Nigeria following the identification of variables required for assessing the severity of SCD. The study also adopted the use of 3 supervised machine learning algorithms namely: naïve Bayes (NB), C4.5 decision trees (DT) and support vector machines (SVM) for creating the ensemble model using a 10-fold cross validation technique. The models were created by adopting the algorithms in isolation and in combination of 2 and 3 which were compared. The developed models were evaluated in order to present the model with the best performance. The results of the study showed that using an ensemble of DT and NB alone provided the best performance. The study has implications in presenting a model for improving the assessment of the severity of SCD among paediatric patients in Nigeria.

Keywords: *Sickle Cell Disease (SCD), Disease severity, Stack-Ensemble Model, Naïve Bayes, Decision Trees, Multi-Layer Perceptron.*

1. Introduction

Sickle cell disease (SCD) is a genetic blood disorder – a structural variant of normal haemoglobin which affects the red blood cells of humans and has led to high morbidity and mortality rates thereby becoming a global public health concern (Chakravorty & Williams, 2015). According to the World Health Organization (WHO, 2006), it was recommended that 50% of member states establish SCD control programs by the year 2020. Aliyu *et al.* (2008), reported that there are between 20 and 25 million people worldwide living with SCD among which 12 to 15 million live in Africa.

According to Agasa *et al*. (2010), the highest prevalence of sickle-cell trait (SCT) in Africa usually occur around tropical areas which lie between latitudes of 15° North and 20° south with SCD prevalence which range between 10 and 40 percent of the population. It was also estimated that 240,000 children were born with SCD annually in sub-Saharan Africa (Makani *et al*., 2011). It was also estimated that 75 to 85 percent of children born with SCD were born in Africa, where mortality rates for those under the age of 5 years range from 50 to 80 percent. In 2012, it was reported that SCD had affected 20 to 25 million people globally among which 50 to 80 percent of infants born with SCD in Africa die before the age of 5 years (Aygun & Odame, 2012).

Life expectancy in SCD was substantially reduced especially in those with severe disease as reported in a 10-year retrospective study which revealed that the mean age of SCD patients was found to reduce, suggesting reduced life expectancy. Anemia is a major cause of morbidity and mortality in SCD, and many patients die in hospital emergency rooms and wards before blood transfusions can be initiated (Ikefuna & Emodi, 2007). It has been suggested that one factor associated with the high incidence of SCD in tropical Africa is the protection against Plasmodium malaria associated with having the SCD (Aygun & Odame, 2012).

Machine learning (ML) is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization tools to learn from past examples and afterwards use the prior training to classify new data, identify new patterns or predict novel trends (Mitchell 1997). Machine Learning has also been extensively adopted in

medical research to generate knowledge from complex clinical data required for improving clinical decision making process (Jaree *et al.,* 2013). Observational studies show that data mining and machine learning prediction techniques have been widely used to determine patterns and how these patterns can be used by physicians to determine diagnoses, prognosis and apply treatment for patients (Boris and Milan, 2012).

The adoption of machine learning into healthcare research has also shown success in the prediction and diagnosis of various diseases thus increasing the accuracy of diagnosis and provide answers to physicians about affected patients (Jiang *et al.*, 2017). Ensemble learning refers to the procedures employed to train multiple learning machines and combine their outputs, treating them as a committee of decision makers (Joshi & Srivastava, 2014). The success of the ensemble approach depends on the diversity in the individual classifiers with respect to misclassified instances (Simidjievski *et al.*, 2016). There exist numerous methods for model combination which includes: linear combiner, the product combiner, and the voting combiner are by far the most commonly used in practice.

In Nigeria today, the number of children with sickle cell disorder (SCD) is increasing with every recorded birth thus leading to an increase in the number of deaths associated with SCD. The number of deaths associated with SCD has been in part as a results of poor management of SCD patients by uneducated and ignorant parents leading to the number of emergency visits to the hospitals for treatment for anaemia and crisis episodes. Related studies in the area of SCD have been targeted at either monitoring the risk of SCD or at the survival of SCD with but not to the management of SCD.

A number of related works which have adopted the use of machine learning to the healthcare data management have shown that a single classifier may have varying performance over a variety of data which can be removed by combining more than one classifier. This challenge has paved way for the development of ensemble methods which combine one or more machine learning algorithms for the development of predictive models. There is a need for the development of an ensemble of machine learning algorithms aimed at improving the assessment of the severity of SCD among paediatric patients in Nigeria, hence this study.

2. Related Works

A number of study have been reviewed in this study, among which include the application of machine learning to healthcare research and the adoption of the ensemble of various machine learning algorithms for predictive modeling.

Xiao *et al.* (2018), worked on the development of a deep learning-based multi-model ensemble method for cancer prediction. The study applied deep learning to an ensemble approach that incorporated 5 different machine learning models by supplying informative gene data selected by differential gene expression analysis to five different classification models. The results revealed that the proposed method showed an average accuracy of 98% however was limited to the demonstration of the advantage of voting ensemble learning over traditional machine learning techniques.

Xu *et al.* (2017), applied neural networks to the classification of red blood cells among SCD patients. The study proposed a method for automating the high-throughput Red Blood Cell (RBC) shape classification using a neural network framework. The presented a feature extraction of the region of interest (ROI) from RBC images following which the images captured were normalized. A convolutional neural network based classification system was formulated using 7000 single RBC images via 5 fold cross validation collected from 8 SCD patients. The results showed that the model was able to classify RBC with an accuracy of 67.5%. The study was limited to the classification of red blood cells.

Goyal and Kaur (2016), worked on a survey of application of ensemble modeling for loan prediction. The study identified that there were various techniques of ensemble which include: bagging, stacking, boosting, voting and using bucket of models to mention a few. The results of the survey showed that the development of

ensemble models guarantee better forecasting, a more constant model, better results and error reduction. The study was limited to a survey of the application of ensemble modeling for improving model performance.

King (2015) applied ensemble learning methods using various machine learning algorithms to structured and unstructured data which were collected, pre-processed, analyzed and followed by model evaluation. The study developed an ensemble model for classifying profitable campaigns thereby maximizing overall campaign portfolio profits. The study adopted the use of 4 traditional classifiers and 4 ensemble learning techniques to build models for identifying pay-per-click campaigns. The results if the study showed that using an ensemble configuration produced the highest campaign portfolio profit. The study was limited to the application of ensemble modeling to marketing data.

Milton *et al.* (2014) performed the prediction of fetal hemoglobin in sickle cell aneamia using an ensemble of genetic risk prediction models. The study developed a collection of 14 models with genetic risk score (GRS) composed of different numbers of single nucleotide polymorphisms (SNPs), and use the ensemble of these models to predict HbF in sickle cell anemia patients. The models were trained in 841 sickle cell anemia patients and were tested in three independent cohorts. The ensemble of 14 models explained 23.4% of the variability in HbF in the discovery cohort, while the correlation between predicted and observed HbF in the 3 independent cohorts ranged between 0.28 and 0.44.

Otaigbe (2013) performed a study on the prevalence of blood transfusion in sickle cell anemia patients in south-south Nigeria over a two year period. The study involved the collection of data from the files of patients seen in clinic or admitted in the Pediatrics Department of the University of Port Harcourt Teaching Hospital within 2 years. Of the 131 cases observed, 130 had genotype Hb SS and 1 had genotype Hb SC. The results of the study showed that 57% had received at least one blood transfusion with the commonest indication been severe anaemia. The study concluded that efforts must be made to reduce the frequency of blood transfusion by monitoring the level of hematocrit in SCD patients.

3. Materials and Methods

This section identified the material and methods that were adopted for the development of the ensemble model required for assessing the severity of SCD among paediatric patients receiving treatments. It consists of a sequence of methods which started with the identification and the collection of data containing the features alongside the target classes of SCD severity. The ensemble model was formulated for the severity of aneamia based on the data collected using a combination of the C4.5 decision trees (DT), Naïve Bayes (NB) and Support Vector Machines (SVM) classifiers. The ensemble model of classifiers was simulated using the Waikato Environment for Knowledge Analysis (WEKA) followed by a performance evaluation of model required for validating the ensemble model required for assessing the severity of SCD patients.

3.1 Method of data identification and collection

Following the review of related works of literature in the body of knowledge of SCD and its severity, a number of variables required for determining the severity of SCD were also determined. The identified variables for assessing the severity of SCD among paediatrics were validated by the medical experts with more than 10 years' experience was interviewed before the data was collected from the medical records office of the Wesley Guilds, Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) in Ilesha, Osun State.

The data was collected from paediatric SCD patients aged below 15 years. Information about the aforementioned variables was collected and stored into electronic format from the information stored in the files located at the medical records department of Wesley Guilds OAUTHC, Ilesha, Nigeria. The data collected was used for the formulation of the predictive model for determining the severity of SCD among paediatrics. A description of the variables are presented as follows.

a. **Gender (sex) of the patient:** was used to identify the gender of the SCD patients which was recorded as a nominal value Male (M) or female (F);
b. **Age of patients**: was used to identify the present age of the which was recorded as a numeric value (measured in years);
c. **Age at diagnosis:** was used to identify the age at which the SCD patients was screened for the presence and identification of SCD which was recorded as a numeric value (measured in years);
d. **Ethnicity:** was used to identify the ethnic tribe to which an SCD patient belonged to. It was measured as a nominal variable with values: Yoruba, Hausa, Ibo and others;
e. **Religion of SCD patient:** was used to identify the religion of the SCD patient that is receiving treatment and was measured using a nominal variable with values: Yoruba, Hausa, Ibo and others;
f. **Body Mass Index (BMI):** was used to identify the nutritional status of an SCD patients based on the values of the weight (measured in Kg) and the height (measured in meters). The BMI is a numeric values measured in Kg/m^2 and can also be classified into nominal values such as: underweight, normal, obese and overweight;
g. **Clinical variables for assessing anemia risk:** These are a class of other variables which will be identified by the physician interviewed and will be used to estimate the risk of anemia among SCD patients receiving treatment. Such includes: the packet cell volume (PCV), frequency of aneamia crisis, frequency of blood transfusions and so on.

3.2 Formulation of ensemble model of machine learning algorithms

This study adopted the development of an ensemble of three (3) machine learning algorithms which was formulated based on historical data collected from SCD patients. Figure 1 shows a diagram of how the ensemble model combined the 3 classifiers selected for this study. The ensemble model was formulated following the standard process of dividing the dataset into two (2) parts namely training and testing which adopted the 10-fold cross validation training technique. Therefore, the training dataset was used to formulate the models following which the testing dataset was used to validate the model.

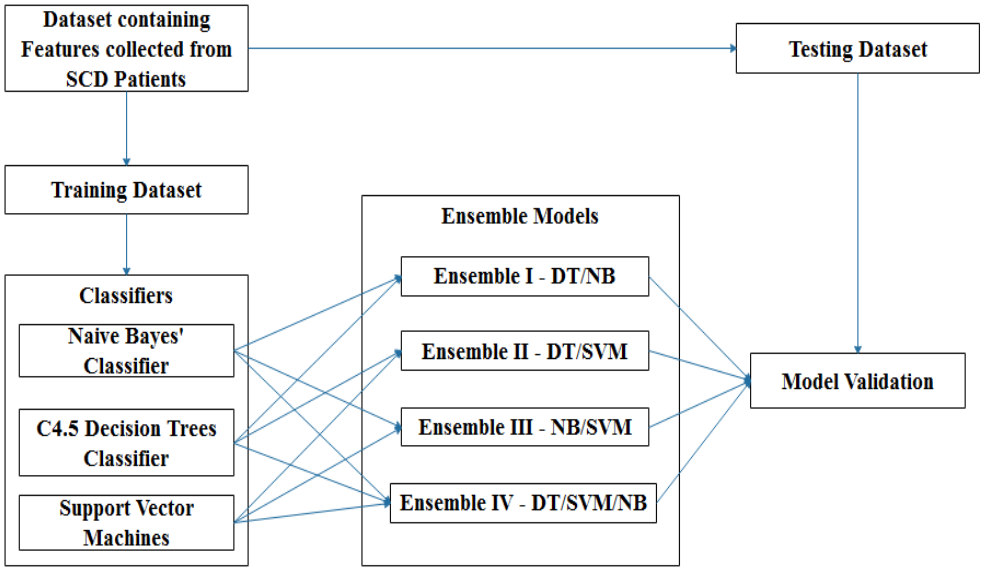

Figure 1: Ensemble Model for Anaemia Risk

Equation (1) shows the mapping function that describes the relationship between the causative features and the target class used to assess the severity of SCD using the ensemble model φ. The equation shows the relationship between the set of causative features represented by a vector, X consisting of the values of i variables and the label Y which defines the severity of SCD identified as Low, Moderate and High. Assuming the values of the set of variable for a SCD patient is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each variable, i

= 1 to i; then the mapping φ used to represent the predictive model for the risk of anemia maps the variables of each SCD patient to their respective risk of aneamia according to equation (2).

$$\varphi: X \rightarrow Y \quad (1)$$

$$defined\ as: \varphi(X) = Y$$

$$\varphi(X) = \begin{cases} Low \\ Moderate \\ High \end{cases} \quad (2)$$

This study adopted a process of developing the predictive model for the severity of SCD using naïve Bayes' (NB), support vector machines (SVM) and the C4.5 decision trees (DT) algorithms in isolation following which the ensemble model which combined the classifiers using a voting technique was formulated. Therefore, the first ensemble named Ensemble I combined DT and NB followed by Ensemble II which combined DT and SVM and Ensemble II which combined NB and SVM. Also, an ensemble model which combined all 3 classifiers, namely: DT, NB and SVM was also formulated called Ensemble IV. Following the formulation of the ensemble model by combining the respective classifiers, the testing dataset was used to validate the performance of the predictive model for the risk of anemia using a number of performance evaluation metrics. The algorithms adopted are presented in the following paragraphs.

- **C4.5 Decision Trees (DT) classifier**

The C4.5 decision trees classifier represents model evaluated from dataset as a hierarchical tree structure using a splitting criteria called the gain ratio. During the training process of model development using the historical dataset collected, the pattern was learned by the tree by splitting the training dataset into subsets based on an attribute value test for each input variables; the process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion was completed when the subset at a node had all the same value of the target class, or when splitting no longer adds value to the predictions. The criteria used by the C4.5 decision trees for tree split is the gain ratio which required the use of the information gain in equation (3) and split criteria in equation (4) to determine the gain ratio by dividing equation (3) by equation (4).

$$IG(X_i) = H(X_i) - \sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \quad (3)$$

Where:

$$H(X_i) = -\sum_{t\epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \quad (4)$$

- **Naïve Bayes' (NB) classifier**

Naive Bayes' Classifier is a probabilistic model based on Bayes' theorem. It is defined as a statistical classifier. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X_{ij} be a dataset sample containing records (or instances) of i number of risks factors (attributes/features) alongside their respective severity of SCD, C (target class) collected for j number of records/patients and $H_k = \{H_1 = Low\ Risk, H_2 = Moderate\ Risk, H_3 = High\ Risk\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the risk of anaemia given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following:

- $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 3$.
- $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 3$;
- $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), i; and
- $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Therefore, the posteriori probability of a hypothesis H_k is defined according to Bayes' theorem as follows:

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2,3 \qquad (5)$$

Hence, the severity of SCD for a record is thus:

$$max.\left[P(H_1|X_{ij}), P(H_2|X_{ij}), P(H_3|X_{ij})\right] \qquad (6)$$

- **Support vector Machines (SVM)**

An SVM model is a representation of the examples (data records) as points in space which were mapped so that the examples of the separate categories: Low, Moderate and High Risk were divided by a clear gap that is as wide as possible. In formal terms, the SVM was used to construct a hyper-plane in a high-dimensional space, and adopted for classification using the sequential minimum optimization (SMO) algorithms. A good separation was achieved by the hyperplane $< \boldsymbol{w}, \boldsymbol{x} > +b = 0$ that has the largest distance $\frac{2}{||w||}$ to the neighbouring data points of either classes at opposite ends, since in general the larger the margin the lower the generalization error of the SVM classifier. A hyperplane created is defined as $< \boldsymbol{w}, \boldsymbol{x} > +b = 0$ where $\boldsymbol{w} \epsilon \mathbb{R}^p$ and $b \epsilon \mathbb{R}$ while $< \boldsymbol{w}, \boldsymbol{x} > +b = -1$ and $< \boldsymbol{w}, \boldsymbol{x} > +b = 1$ are the margins required for the separation *w* of support vectors *x* within the *n* variables.

3.3 Simulation of ensemble model of machine learning algorithms

The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model following the collection of data about paediatric SCD patients. The dataset collected was divided into two parts: training and testing data – the training data was used to formulate the model while the test data was used to validate the model. The process of training and testing predictive model according to literature is a very difficult experience especially with the various available validation procedures.

For this classification problem, it was required to measure a classifier's performance in terms of the error rate. In order to predict the performance of a classifier on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the classifier. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data using the 10-fold cross validation technique.

The 10-fold cross validation technique involved the process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. It involved dividing the whole datasets into 10 folds (or partitions) such that each partition was selected for testing with the remaining 9 partitions used for training. Each new partition was used for testing with the remaining successive 9 partitions (including the first partition used or testing) used for training until all 10 partitions had been selected for testing.

3.4 Validation of ensemble of machine learning algorithms

During the course of evaluating the predictive model, a number of metrics were used to quantify the model's performance for model validation following model simulation using WEKA. These results of the correct and

incorrect classifications made by the ensemble model on the testing dataset was presented on a confusion matrix. For this study, the confusion matrix is a 3 x 3 confusion matrix table owing to the three (3) labels of the output class as shown in Figure 2. Using the confusion matrix, correct classifications were plotted along the diagonal from the north-west position for Low risk predicted as Low risk (A), followed by Moderate risk predicted as Moderate risk (E) and High risk predicted as High risk (I) on the south-east corner.

The incorrect classifications were plotted in the remaining cells of the confusion matrix. Also, the actual Low risk cases are A+B+C, actual Moderate risk cases are D+E+F, while actual High risk cases are G+H+I and the predicted Low risk cases are A+D+G, predicted Moderate risk cases are D+E+F and predicted High risk cases are G+H+I. The developed model was validated a number of performance metrics based on the values of *A – I* in the confusion matrix for each predictive model.

Low	Moderate	High	<- Predicted as
A	B	C	Low
D	E	F	Moderate
G	H	I	High

Figure 3: Confusion Matrix for Model Performance Evaluation

a. **Accuracy**: the total number of correct classification

$$Accuracy = \frac{A + E + I}{A + B + C + D + E + F + G + H + I} \quad (7)$$

b. **True positive rate (recall/sensitivity)**: the proportion of actual cases correctly classified

$$TP_{Low} = \frac{A}{A + B + C} \quad (8a)$$

$$TP_{Moderate} = \frac{E}{D + E + F} \quad (8b)$$

$$TP_{High} = \frac{I}{G + H + I} \quad (8c)$$

c. **False positive (false alarm/1-specificity):** the proportion of negative cases incorrectly classified as positive

$$FP_{Low} = \frac{B + C}{D + E + F + G + H + I} \quad (9a)$$

$$FP_{Moderate} = \frac{D + F}{A + B + C + G + H + I} \quad (9b)$$

$$FP_{High} = \frac{G + H}{A + B + C + D + E + F} \quad (9c)$$

d. **Precision**: the proportion of predictions that are correct

$$Precision_{Low} = \frac{A}{A + D + G} \quad (10a)$$

$$Precision_{Moderate} = \frac{E}{D + E + F} \quad (10b)$$

$$Precision_{High} = \frac{I}{C + F + I} \quad (10c)$$

Using the aforementioned performance metrics, the performance of the predictive model for the classification of risk of aneamia can be evaluated by validation using a historical dataset collected based on the information provided in the questionnaire. The TP rate and precision lie within the interval [0, 1], accuracy within the interval of [0, 100] % while the FP rate lies within an interval of [0, 1]. The closer the accuracy is to 100% the better the model, the closer the value of the TP rate and precision is to 1 the better while the closer the value of FP rate is to 0 the better. Therefore, the evaluation of an effective model has a high TP/Precision rates and a low FP rates.

4. Results

This section presents the results of the methods adopted for the development of the ensemble model for the severity of SCD using 3 machine learning algorithms. The results of the description of the nominal and numeric attributes within the dataset were also identified. Following the presentation of the results of the description of attributes identified in this study for the development of the classification model for the severity of anemia among pediatric SCD patients the presentation of the results of the different models formulated based on the individual and ensemble of selected classifiers. The results of the evaluation of the performance of the predictive model was done based on the outcome of the testing phase using accuracy, true positive (TP) rate, false positive (FP) rate and precision with the model with the best performance presented.

4.1 Results of data identification and collection

The data considered in this study which was collected from South-west Nigeria contained demographic and clinical information about pediatric SCD patients. Based on the data collected from the patients, the severity of anemia among SCD patients was measured and recorded by the experts. The results of the data collection process showed that, majority of the SCD patients had Moderate severity owing for a proportion of 55.7% followed by SCD patients that had Low severity with a proportion of 33.9%.

Table 1 gives a description of the results of the data collected about SCD patients in terms of the severity of anemia among pediatric SCD patients. Table 2 shows a summary of the description of the nominal attributes among the identified attributes using a frequency distribution table in terms of the frequency of distribution and percentage of total. Based on the data presented in Table 2, it was observed that majority of the data collected consisted of male patients which was about 65% owing for a ratio of about 2 to 1 for male to female SCD patients.

Table 1: Results of the Distribution of the Severity of Anemia

Variable	Frequency	Percentage (%)
Low Risk	39	33.9
Moderate Risk	64	55.7
High Risk	12	10.4
Total	**115**	**100.0**

The results of the data collected about the occupation of the parents of SCD patients showed that majority of the mothers were traders owing for a proportion of 46% followed by either artisans of teachers/civil servants with a proportion of 23.55 each. Regarding the occupation of the fathers, it the results showed that majority of

the fathers were either artisans or teacher/civil servants with a proportion of 32% each. The results regarding the parent's education based on the results showed that at least 30% of the parents had university education. The results showed that majority of the parents were of upper class owing for a proportion of 40% followed by middle class parents with proportion of 34%.

Table 2: Results of the Description of Nominal Attributes

Variable Name	Values	Score	Frequency	Percentage (%)
Sex	**Male**	1	75	**65.2**
	Female	2	40	**34.8**
Mother's Education	**No formal education**	1	6	**5.2**
	Primary	2	16	**13.9**
	Secondary	3	54	**47.0**
	University	4	39	**33.9**
Mother's Occupation	**Full housewife**	1	4	**3.5**
	Artisan	2	27	**23.5**
	Teacher/civil servant	3	27	**23.5**
	Large scale business	4	0	**0.0**
	Professional	5	2	**1.7**
	Student	6	2	**1.7**
	Trader	7	53	**46.1**
Father's Education	**No formal education**	1	5	**4.3**
	Primary	2	6	**5.2**
	Secondary	3	52	**45.2**
	University	4	52	**45.2**
Father's & Occupation	**Student**	1	0	**0.0**
	Artisan	2	37	**32.2**
	Teacher/civil servant	3	36	**31.3**
	Large scale business	4	1	**0.9**
	Professionals	5	13	**11.3**
	Trader	6	17	**14.8**
	Driver	7	11	**9.6**

Social Class	**Upper Class**	1	20	**17.4**
		2	26	**22.6**
	Middle Class	3	39	**33.9**
	Lower Class	4	23	**20.0**
		5	7	**6.1**
CVD Lifetime Incidence	**Yes**	1	5	**4.3**
	No	2	110	**95.7**
Acute chest syndrome Lifetime Incidence	**Yes**	1	23	**20.0**
	No	2	92	**80.0**
AVN Lifetime Incidence	**Yes**	1	5	**4.3**
	No	2	110	**95.7**
Pneumococcal Meningitis Lifetime Incidence	**Yes**	1	2	**1.7**
	No	2	113	**98.3**
Gall Stone Lifetime Incidence	**Yes**	1	3	**2.6**
	No	2	112	**97.4**
Osteomyelitis Lifetime Incidence	**Yes**	1	29	**25.2**
	No	2	86	**74.8**
Chronic Leg Ulcer Lifetime Incidence	**Yes**	1	2	**1.7**
	No	2	113	**98.3**
Priapism Lifetime Incidence	**Yes**	1	2	**1.7**
	No	2	113	**98.3**

As shown in Table 4.2, it was observed that majority of the SCD patients assessed did not have any of the associated lifetime incidence related to CVD with a proportion of at least 95%, acute chest syndrome with a proportion of at least 80%, pneumococcal meningitis with a proportion of at least 98%, gall stone with a proportion of at least 97%, osteomyelitis with a proportion of at least 74%, chronic leg ulcer with a proportion of at least 98% and priapism with a proportion of at least 98%. The results of the distribution of the numeric clinical data is also presented in Table 3 in terms of the minimum, maximum, mean and standard deviation of the features assessed.

Table 3: Description of the Numeric Attributes collected

Variable Name	Minimum	Maximum	Mean	Standard Deviation
Present Age (in years)	1	15	6.60	3.795
Age at first diagnosis (months)	4	156	29.62	28.113
Frequency of Painful Crisis	0	20	3.87	3.671
Frequency of Blood Transfusions	0	8	1.24	1.490
Frequency of Hospitalization	0	12	2.23	2.271
Spleen size (in cm)	0	18	4.29	4.432
Liver size (in cm)	0	11	3.96	2.921
Hematocrit Level (PCV) (%)	6.0	37.0	23.08	4.620
White Blood Cell (WBC) Count (/mm^3)	55	87000	15111.87	12528.441
HbF Level	1.1	10.3	5.52	2.461

The results of the data collection showed that the minimum age of SCD patients assessed is 1 year with a maximum of 15 years which yielded an average age of 6 years for SCD patients assessed. It was also observed in the results that the minimum age of diagnosis of SCD was 4 months with a maximum of 156 months which yielded an average age of 29 months. The frequency of blood transfusions, hospitalization painful crisis was evaluated based on the number of episodes in the last year. The results showed that regarding the number of painful crisis in the last year, the maximum recorded was 20 painful episodes within a distribution with a mean of 3 episodes and standard deviation of 3 episodes. The results showed that regarding the number of blood transfusions in the last year, the maximum recorded was 8 transfusion episodes within a distribution with a mean of 1 episode and standard deviation of 1 episode. The results showed that regarding the number of hospitalizations episodes in the last year, the maximum recorded was 12 within a distribution with a mean of 2 episodes and standard deviation of 2 episodes.

4.2 Results of ensemble model formulation and simulation

This section presents the results of the process of formulating and simulating the classification models required for assessing the severity of anemia among SCD patients using 3 supervised machine learning algorithms. In one part, the classification model was formulated using the supervised machine learning algorithms in isolation. In the other part, the classification model was formulated using an ensemble of the supervised machine learning algorithms. The simulation of the classification models was simulated using a 10-fold cross validation technique via the Waikato Environment for Knowledge Analysis (WEKA).

- **Results of the formulation and simulation of isolated models**

Based on the results of the application of the C4.5 Decision Trees algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 3 and 4 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The

presentation of the number of correct and incorrect classification of each target class for the severity of anemia is presented in Figure 4 (left). The results of the performance of the application of the C4.5 decision trees algorithm showed an accuracy of 84.3%.

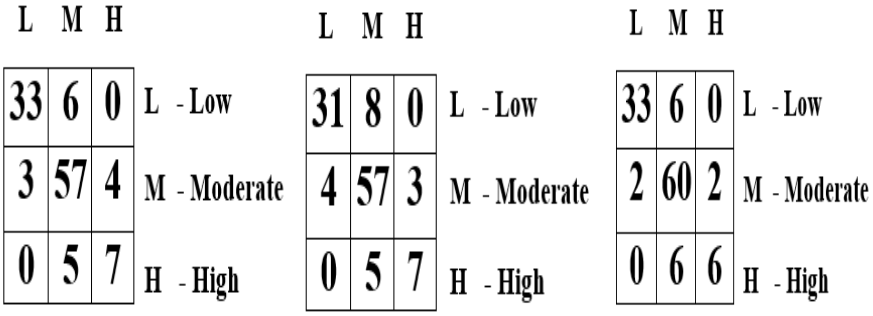

Figure 4: Results of the Isolated Classifiers

Regarding the results of the application of the naïve Bayes' algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 31 were correctly classified while 8 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 4 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia is presented in Figure 4 (center). The results of the performance of the application of the naïve Bayes' algorithm showed an accuracy of 82.6%.

Regarding the results of the application of the support vector machine (SVM) using the Sequential Minimal Optimization (SMO) algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 60 were correctly classified while 2 were misclassified as each of low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 6 were correctly classified while 6 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia is presented in Figure 4 (right). The results of the performance of the application of the SVM algorithm showed an accuracy of 86.1%.

- **Results of the formulation and simulation of ensemble of 2 models**

Based on the results of the application of ensemble of algorithms, using the ensemble of C4.5 Decision Trees (DT) and naïve Bayes' (NB) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 59 were correctly classified while 2 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 8 were correctly classified while 4 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT and NB is presented in Figure 5 (left). The results of the performance of the ensemble of DT and NM algorithms showed an accuracy of 87%.

Regarding the results of the application of ensemble of algorithms, using the ensemble of C4.5 Decision Trees (DT) and support vector machines (SVM) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 3 and 4 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number

of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT and SVM is presented in Figure 5 (center). The results of the performance of the ensemble of DT and SVM algorithm showed an accuracy of 84%.

Regarding the results of the application of ensemble of algorithms, using the ensemble of naïve Bayes' (NB) and support vector machines (SVM) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 56 were correctly classified while 5 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 5 were correctly classified while 7 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of NB and SVM is presented in Figure 5 (right). The results of the performance of the ensemble of DT and SVM algorithm showed an accuracy of 83%.

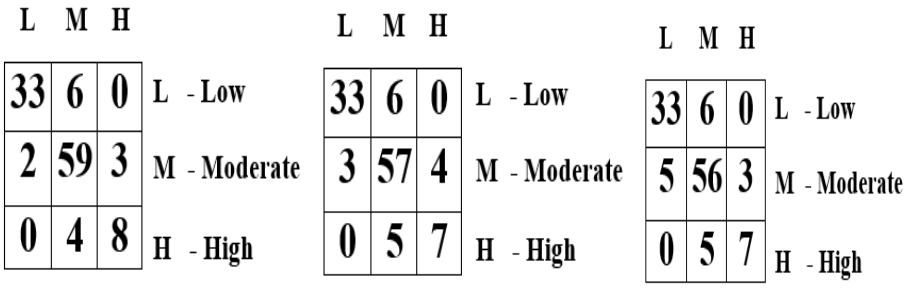

Figure 5: Results of the Ensemble of Two (2) Classifiers

- **Results of the formulation and simulation of ensemble of all models**

The results of the performance of the ensemble of all the machine learning algorithms showed that out of the 39 actual low severe cases, 32 were correctly classified while 7 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 59 were correctly classified while 3 and 2 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT, SVM and NB is presented in Figure 6. The results of the performance of the ensemble of DT, SVM and NB algorithms showed an accuracy of 85%.

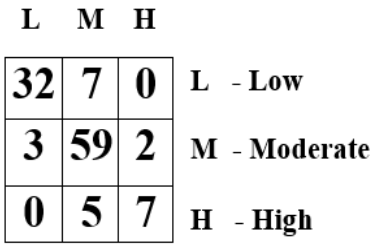

Figure 6: Results of the Ensemble of the Three (3) Classifiers

4.3 Results of ensemble model validation

Based on the results of the performance of the DT algorithm, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the model developed using the DT also showed that an average of 84% of the

predictions made by the algorithm was correct. Based on the results of the performance of the NB algorithm, it was observed that the model developed classified correctly on average about 83% of the actual cases alongside misclassification rate of an average of 16% of actual cases. The results of the model developed using the NB also showed that an average of 82% of the predictions made by the algorithm was correct.

Based on the results of the performance of the SVM algorithm, it was observed that the model developed classified correctly on average about 86% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the model developed using the SVM also showed that an average of 86% of the predictions made by the algorithm was correct. It was also observed from the results of the isolated algorithms that among the adopted algorithms for the classification of the severity of SCD patients, SVM had the highest capability of predicting correctly the severity of anemia and the lowest ability of misclassifying the severity of SCD patients.

Based on the results of the performance of the ensemble of DT and NB algorithms, it was observed that the model developed classified correctly on average about 87% of the actual cases alongside misclassification rate of an average of 12% of actual cases. The results of the model developed using the ensemble of DT and NB also showed that an average of 87% of the predictions made by the algorithm was correct.

Based on the results of the performance of the ensemble of DT and SVM algorithms, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the ensemble model developed using the DT and SVM also showed that an average of 84% of the predictions made by the algorithm was correct.

Based on the results of the performance of the ensemble of NB and SVM algorithms, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 15% of actual cases. The results of the model developed using the ensemble of NB and SVM also showed that an average of 83% of the predictions made by the algorithm was correct.

It was also observed from the results of the ensemble of 2 algorithms that among the adopted algorithms for the classification of the severity of anemia among SCD patients, using and ensemble of DT and SVM had the highest capability of predicting correctly the severity of anemia and the lowest ability of misclassifying the severity of SCD patients.

Also, based on the ensemble of 3 algorithms using DT, SVM and NB, it was observed that the model developed classified correctly on average about 85% of the actual cases alongside misclassification rate of an average of 15% of actual cases. The results of the model developed using the ensemble of DT, SVM and NB also showed that an average of 85% of the predictions made by the algorithm was correct. The results of the study showed that out of all the models developed for the classification of the severity of anemia among SCD patients that the adoption of the ensemble of DT and NB showed the best performance out of all the proposed combinations. The results showed that although in isolation, NB and DT did not do as well as did SVM but what could not be achieved in isolation was compensated for in combination.

On the other hand, the ensembles that were developed using the combination of SVM with either NB or DT did not produce results better than that produced by the SVM in isolation. It was also observed from the results that ensemble of DT with SVM did not shows any improvement in performance over the use of DT in isolation while the ensemble of the 3 algorithms was observed to be better than the use of NB and DT in isolation. It was also observed for the results that using the ensemble of the 3 algorithms had better performance compared to the use of the ensemble of DT and SVM and that of NB and SVM. In all, the ensemble of DT and NB was observed to produce the best results among the various combination of ensemble models for the 3 supervised machine learning algorithms adopted in this study due to its high TP rate and Precision and low FP rates. The summary of the model validation results is presented in Table 4.

Table 4.4: Results of the Evaluation of the Performance of Classification Models

Classifier	Accuracy	Correct	TP rate	FP rate	Precision
C4.5 Decision Trees (DT)	84.348	97	0.843	0.137	0.844
Naïve Bayes (NB)	82.609	95	0.826	0.163	0.827
Support Vector Machines (SVM)	86.087	99	0.861	0.142	0.862
Ensemble I (DT + NB)	86.957	100	0.870	0.121	0.872
Ensemble II (DT + SVM)	84.348	97	0.843	0.137	0.844
Ensemble III (NB + SVM)	83.478	96	0.835	0.145	0.833
Ensemble IV (DT + NB + SVM)	85.217	98	0.852	0.146	0.854

5. Conclusion

The study concluded that a number of demographic and clinical variables were associated with the risk of anaemia based on the severity of SCD among patients. The study concluded that the data collected contained a majority of Moderate risk cases followed by Low risk and High risk cases. The study concluded that following the process of model formulation and simulation using the WEKA simulation environment that by using a combination of classifiers, it was observed that a better performance was detected compared to using the classifiers in isolation for model development.

The study concluded that the process of model classification using the isolated classification algorithms showed that among the isolated model, the application of SVM yielded the best results among the 3 selected algorithms selected for this study. On the other hand, following the performance of the SVM among the isolated algorithms is the application of the C4.5 decision trees algorithm followed by naïve Bayes' algorithm.

The study concluded that the model development using an ensemble of two (2) algorithms revealed that the performance of the ensemble which applied decision trees outperformed the performance of the model formulated using the isolated decision trees algorithms. The results however showed that both ensemble models created using SVM with NB and DT did not perform well as did the performance of the model formulated using the isolated SVM algorithm. Also, the results showed that all the ensemble of 2 algorithms outperformed the use of naïve Bayes' classifier in isolation.

The study concluded that using an ensemble of 3 classifiers, a performance better than that using either DT or NB in isolation and that of using and ensemble of either DT+SVM or NB+SVN was determined. On a general note, the results showed that the best classification model for determining the severity of anemia among SCD patients was developed using an ensemble of DT and NB algorithms. The study concluded that the predictive model for the risk of anaemia with the best performance was the ensemble model which combined the C4.5 decision trees (DT) and the naïve Bayes' (NB) classifiers.

References

1. Agasa, B., Bosunga, K., Opara, A., Tshilumba, K., Dupont, E., & Vertongen, F. (2010). Prevalence of sickle cell disease in a northeastern region of the Democratic Republic of Congo: What impact on transfusion policy? *Transfusion Medicine 20*(1): 62 – 65.

2. Aliyu, Z. Y., Kato, G. J., Taylor, Jt., Babadoko, A., Mamman, A. I. & Gordeuk, V. R. (2008). Sickle cell disease and pulmonary hypertension in Africa: A global perspective and review of epidemiology, pathophysiology, and management. *American Journal of Hematology 83*(1): 63–70.

3. Aygun, B. & Odame, I. (2012). A global perspective on sickle cell disease. *Pediatric Blood & Cancer 59*(2): 386 – 390.

4. Chakravorty, S. & Williams, T. N. (2015). Sickle cell disease: A neglected chronic disease of increasing global health importance. *Archives of Disease in Childhood 100*(1): 48 – 53.

5. Goyal, A. & Kaur, R. (2016). A Survey on Ensemble Model for Loan Prediction. *International Journal of Advanced research and Innovative Ideas in Education (IJARIIE), 2*(1): 623 – 628.

6. Ikefuna, A. N. & Emodi, I. J. (2007). Hospital admission of patients with sickle cell anaemia pattern and outcome in Enugu area of Nigeria. *Nigerian Journal of Clinical Practice 10*(1): 24–29.

7. Jaing, Y., Qiu, B., Xu, C. & Li, C. (2017). The Research of Clinical Decision Support System Based on Three-Layer Knowledge base Model. *Journal of Healthcare Engineering, 7*: 12 – 32.

8. Joshi, N. & Srivastava, S. (2014). Improving Classification Accuracy using Ensemble Learning Technique (using different Decision Trees). *International Journal of Computer Science and Mobile Computing, 3*(5): 727 – 732.

9. King, M.A. (2015). Ensemble learning techniques for Structured and Unstructured Data. Unpublished PhD Thesis of the Department of Business Information Technology.

10. Makani, J., Cox, S. E., Soka, D., Komba, A. N., Oruo, J. & Mwamtemi, H. (2011). Mortality in sickle cell anemia in Africa: A prospective cohort study in Tanzania. *PLoS ONE 6*(2): 1 – 12.

11. Milton, J.N., Gordeuk, V.R., Taylor, J.G., Gladwin, M.T., Steinberg, M.H. & Sebastiani, P. (2014). Prediction of Fetal Hemoglobin in Sickle Cell Anemia using an Ensemble of Genetic Risk Prediction Models. *Circulatory Cardiovascular Genetics, 7*92): 110 – 115.

12. Mitchell T. (1997). *Machine Learning*. New York: McGraw Hill

13. Simidjievski, N., Todorovski, L. & Dzeroski, S. (2016). Modeling Dynamic Systems with efficient Ensembles of Process-Based Models. *PLoS ONE Computational Biology, 11*(4): 1 – 27.

14. Xiao, Y., Wu, J., Lin, Z & Zhao, X. (2018). A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction. *Journal of computational Methods Programs and Biomedicine, 153*: 1 - 9.

15. Xu, M., papageorgiou, D.P., Abidi, S.Z., Dao, M., Zhao, H. & Karniadakis, G. (2017). A Deep Convolutional Neural Network for the Classification of Red Blood Cells in Sickle Cell Anemia. *PLoS ONE Computational Biology, 13*(10): 1 – 12.