VOLUME 02 ISSUE 02 (2023)

AMERICAN JOURNAL OF MEDICAL SCIENCE AND INNOVATION (AJMSI)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Medical Science and Innovation (AJMSI)

Volume 2 Issue 2, Year 2023
ISSN: 2836-8509 (Online)
DOI: https://doi.org/10.54536/ajmsi.v2i2.1979
https://journals.e-palli.com/home/index.php/ajmsi

Risk Prediction of Thalassemia Using Data Mining Classifiers

Khizra Ali[1*], Muhammad Saqib[1]

Article Information

Received: August 17, 2023
Accepted: September 11, 2023
Published: September 22, 2023

Keywords

Medical Data Mining, Thalassemia, J48 Decision Tree, Naïve Bayesian Network, Multilayer Perceptron Neural Network

ABSTRACT

Medical data mining is concerned with prediction knowledge, which is a useful method for extracting hidden patterns from given data for specific purposes. Thalassemia is one of the most common inherited blood hematological disorders, and this paper adopted data mining classification techniques to generate results with high performance and accuracy for risk prediction of thalassemia. The dataset for this purpose was collected from NIBD (National Institute of Blood Diseases), a well-known institute and hospital for blood diseases in Karachi, Pakistan. They provided 301 records of CBC test reports containing positive and negative statuses of diagnosis of thalassemia traits. There were many instances in the report, of which 6 were used for our research purpose, i.e. Gender, MCV, HGB, HCT, MCHC, and RDW. The dataset was divided into training and test data using the WEKA tool. Four algorithms of data mining classification, namely J48 Decision Tree, Naïve Bayesian Network, SMO algorithm, and Multilayer Perceptron Neural Network were adopted to train the model and classify the patient having traits of thalassemia from normal persons with the use of the WEKA tool. Results revealed that out of all four algorithms, Naïve Bayes provided results with the highest accuracy of 99%.

INTRODUCTION

One of the most common genetic blood disorders, thalassemia, has received excessive attention in the research field of medicine worldwide. Due to the absence or decreased production of normal globin chains, a heterogeneous group of hemoglobin disorders is called thalassemia syndromes. The disease is known to be the most common recessive worldwide as the population carriers of genetic thalassemia mutation are 1-5% globally (Jameel *et al.*, 2017; Origa, 2017). It is commonly found in Southeast Asia, the Middle East, North and Central Africa, and the Mediterranean (Herbert *et al.*, 2009). In Pakistan, thalassemia carriers are more than 10 million, with a 5-7% prevalence rate, as about 5000 children yearly are diagnosed as thalassemia major carriers (Kamil *et al.*, 2021; Khaliq, 2022). Therefore, under this context, the greatest challenge for professionals in healthcare is to differentiate between normal individuals and asymptomatic thalassemia carriers (Alaa & Shurrab, 2017; Jatoi *et al.*, 2018). Thalassemia diagnosis depends on certain characteristics derived after performing a complete blood count (CBC) test. However, the reliability of the test can lead to the misdiagnosis of thalassemia as similar characteristics can also be observed in different blood disorders (Abdullah & Al-Asmari, 2016; Jatoi *et al.*, 2018; Meena *et al.*, 2019).

Blood diseases can be of various types, such as anaemia, which is a common nutritional deficiency and blood disorder in childhood and infancy, and iron deficiency anemia (IRD) is mostly found in women and children, especially in developing countries (AlAgha *et al.*, 2018; Jatoi *et al.*, 2018). However, the most crucial type of anaemia is thalassemia, an inherited disorder whose identification or differentiation from normal patients is challenging from the CBC test (Abdullah & Al-Asmari, 2016). Therefore, the problem identified in the healthcare sector is to design a model that can predict the risk of thalassemia in patients before their CBC test. This research explored data mining techniques using the CBC dataset, examining the similarities of the different attributes categorizing normal traits or thalassemia traits for accurate detection and recognition of optimal disease patterns.

Multiple studies have been done worldwide regarding detecting different kinds of blood disorders using data mining. However, researchers have used a different set of instances of the CBC test, and therefore, the data mining classifiers' accuracy has varied (AlAgha *et al.*, 2018; Elshami & Alhalees, 2012). This also depends upon the choice of data mining classifiers; Amin and Habib (2015) conducted a research to diagnose blood diseases using age and gender as dataset labels with MLP, J48, and Naïve Bayes algorithms (Amin & Habib, 2015). Similarly, Saichanma *et al.* (2014) used the J48 decision tree algorithm to predict the abnormality of peripheral blood smear, focusing mainly on the attribute of RBC of the CBC test (Saichanma *et al.*, 2014).

The previous studies (Abdullah & Al-Asmari, 2016; Alaa & Shurrab, 2017; AlAgha *et al.*, 2018; Jatoi *et al.*, 2018; Meena *et al.*, 2019) have classified the types of anaemia or thalassemia utilizing the techniques of data mining. However, the present study focused on determining thalassemia traits' existence based on the CBC test attributes (MCV, HGB, RDW, MCHC, and HCT) for predicting the risk of thalassemia. The attributes used in this study were gender, MCV, HGB, HCT, MCHC, and RDW. The current research was aimed at risk prediction of thalassemia using data mining classification algorithms. The selected classification algorithms for this study were

[1] Department of Computer Science and Information Technology, NED University of Engineering & Technology, University Road, Karachi 75270, Pakistan
[*] Corresponding author's e-mail: khizra.ali12@gmail.com

Naïve Bayes, Multilayer Perceptron, J48, and SMO using the WEKA data-mining tool. The research also focused on identifying the most potential classification algorithm for thalassemia by determining accuracy, precision, Recall, True Positive Rate, False Positive Rate, and F-measure using WEKA Experimenter.

LITERATURE REVIEW

Data Mining in Healthcare

Three basic data mining mechanisms include Decision Rules, Analysis, and Clustering (Classification). There are various applications of data mining in healthcare that are effective for physicians to detect diseases by the extraction of knowledge from data of patients. It is a prominent field of research and has extensive potential to revolutionise the medical environment and improve the quality of clinical decisions (Asere & Botson, 2021; Saichanma *et al.*, 2014). Tomar and Agarwal (2013) presented the importance of data mining as the most interesting and motivating area of research popular in health organisations. The role played by data mining is essential in uncovering the new trends in the healthcare domain and helpful for the associated field members. Results revealed that the best accuracy and performance of the classification model are presented by feature selection (Tomar & Agarwal, 2013).

Jothi and Husain (2015) illustrated that the most important step of data mining is KDD for pattern discovery and extraction with the involvement of a huge amount of data. The paper reviewed various emerging models and techniques in this regard, and results from the paper review revealed that tasks of classification, clustering, association rules, and anomaly detection are the most used methods in the medical sector. The widely used algorithms are decision tree, discriminant analysis, swarm intelligence, artificial neural network, k-nearest neighbour, association rule, and support vector rule (Jothi & Husain, 2015). Mdaghri *et al.* (2016) elaborated that data mining has been widespread in the many applications of Clinical Decision Support Systems (CDDS) such as alerting systems, information retrieval, suggestion systems, reminders, and prediction models (Mdaghri *et al.*, 2016).

Risk Prediction of Thalassemia Using Data Mining

Jatoi *et al.* (2018) examined the core relationship between Thalassemia and Anemia from the CBC test. The relationship between the two was exploited to predict and identify the possibility of getting thalassemia for patients already suffering from anaemia. The research was conducted in Pakistan, and the dataset of CBC reports was collected from the Diagnostic and Research Laboratory of LUMHS. Researchers with the application of the Naïve Bayesian Network algorithm analysed and evaluated the dataset, and it was revealed in the final results that the Naïve Bayes algorithm has the best capability to detect or predict the core relationship between anaemia and thalassemia with an accuracy of 98%. It was evaluated that out of 400 patients, 290 were infected with the disease, where 110 were normal. Results also reflected that if patients have high or low values of MCV and MCH, then there are increased chances of thalassemia, and the conditions of Iron deficiency and Vitamin B12 are slightly different in terms of MCV and MCH (Jatoi *et al.*, 2018).

Alaa and Shurrab (2017) applied data mining techniques to identify the relations between blood tumours and characteristics of blood tests to predict disease at an early stage, as it can be used to enhance the ability to cure disease. Three data mining techniques were used, i.e. rule induction, association rules, and deep learning, to predict or classify normal patients from the ones suffering from blood tumours. Real-time data was collected from Gaza European Hospital in Palestine, and results showed that association rules provide the best relationship between blood tumour and blood test characteristics. However, deep learning classifiers are best for predicting tumour types, providing an accuracy of 79.45%. Association rules best describe both normal haematology and tumours in blood and are identified as follows in the current research using the inductive rules method (Alaa & Shurrab, 2017).

Abdullah and Al-Asmari (2016) specified the types of anaemia for anaemic patients by using classification algorithms of data mining and constructing a predictive model. The dataset was made by filtering and eliminating the variables that were not desired, and then some classification algorithms such as SMO, Multilayer Perceptron, Naïve Bayes, and J48 Decision Tree were applied to the WEKA tool of data mining to present results. After performing several experiments, the WEKA experimenter has proven that the best accuracy is provided by the J48 Decision Tree algorithm, giving the best potential classification of anaemia types. Evaluation of data from 41 patients was done with several percentage splits of the dataset, i.e. 20%, 40%, and 60%. Furthermore, comparing the results for four algorithms, it was observed that the best accuracy of 93.75% was obtained from SMO and J48 Decision Tree (Abdullah & Al-Asmari, 2016).

AlAgha *et al.* (2018) collected data on patients suffering from thalassemia from the Palestine Avenir Foundation to work on presenting a model for overcoming the challenges of thalassemia. The study aimed to predict a special type of thalassemia (β-thalassemia) based on the hybrid data mining model. Additionally, to overcome the problem of imbalanced class distribution in the dataset, a SMOTE technique was first used and applied to deal with this problem. The second step utilised four classification models to differentiate between patients carrying carriers of β-thalassemia and normal persons with the help of different evaluation metrics. The classification models used were Decision Tree, Naïve Bayes, k-nearest Neighbour (k-NN) and Multilayer Perceptron neural network. Naïve Bayes provided the results with the highest accuracy differentiating between a normal person and thalassemia patients with the sampling SMOTE ratio of more than 400%, revealing a sensitivity of 98.81% and

specificity of 99.47% (AlAgha *et al.*, 2018).

Meena *et al.* (2019) developed a decision support system using data mining techniques as data was collected from a survey conducted by the Indian Government in 2015-16 called NFHS-4. The collected data was utilised to predict anaemia among children and create a relationship between a mother's diet during pregnancy and health affecting the child's anaemic status. In this way, clinicians and parents can understand the influence of the practices of feeding by the mother are fine, and they are following guidelines to prevent anaemia. The two techniques of association rule and decision tree were applied and compared for the mentioned aim, and a model of knowledge discovery was proposed using artificial intelligence to propose systems based on medical experts' advice. The architecture of the proposed system depicted the stages of the data pre-processing tier, processing tier using decision tree and association rules, and user tier (Meena *et al.*, 2019).

Elshami and Alhalees (2012) performed different experiments to identify the most significant classifier useful for differentiating between the multiple types of thalassemia (Normal, Thal-I, Thal-M, Iron Def/Thal-T, Other). Results revealed that neural networks provided the best classification outputs for detecting or diagnosing different types of thalassemia compared to Naïve Bayes and Decision Tree. Almost all the experiments supported the main feature of MCV to indicate thalassemia existence, where the indicator value was identified to be less than 77.65. Furthermore, if the age is greater than 12 and the value of MCV is greater than 77.65, thalassemia is not possible in an individual. The results were bright, with an accuracy greater than 90% (Elshami & Alhalees, 2012).

Hasani and Hanani (2017) investigated the three types of anaemia, including α-thalassemia trait, β-thalassemia trait and iron deficiency anaemia, as detecting them is crucial due to their similar blood characteristics. Also, the tests for their diagnosis are costly and time-consuming; therefore, an accurate model for classifying and diagnosing kinds of anaemia is important. The study used five classification algorithms on WEKA software to suggest an algorithm that provides the lowest mean absolute error and highest accuracy. Results of the research indicated that combining Naïve Bayes, IBK and J48 algorithms using a voting algorithm with all the features provided the highest accuracy of 96.343% and 96.2169% (Hasani & Hanani, 2017).

Egejuru *et al.* (2017) predicted the risk of thalassemia in different age groups, and knowledge regarding the risk factors was collected by carrying out a structural interview with experienced medical personnel, and a questionnaire was used to collect data regarding the related parameters. The environment used for applying supervised machine learning algorithms was WEKA, and the predictive model was validated for diagnosing thalassemia. Data was collected from 51 patients, and dataset parameters included demographics and clinical variables. Age, Gender, Marital Status, Social Class, and ethnicity were demographic variables, whereas spleen enlargement, urine colour changes, family history, diabetes and parent carriers were clinical variables. Results showed that the distribution of risk of thalassemia was 31% high cases, 16% moderate cases, 10% low cases and 43% no cases. It was concluded in this study that for the prediction of thalassemia, multilayer perceptron will improve the process of decision-making within the healthcare sector regarding thalassemia diagnosis (Egejuru *et al.*, 2019).

MATERIALS & METHODS

Dataset and Pre-processing

Data mining also termed knowledge discovery in databases, is a useful tool for extracting knowledge containing the stages of data pre-processing, pattern recognition and classification (Singhal & Jena, 2013). Data pre-processing is about filtering the data by removing the no-interest attributes and reducing noise or inconsistencies such as missing values or outliers (Sultana *et al.*, 2016). The stages of pre-processing of data utilised in this research are shown in Figure 1 below.

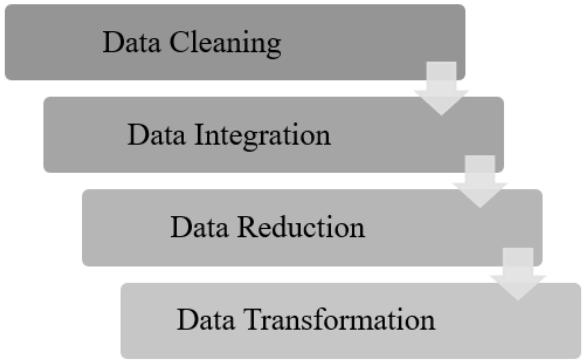

Figure 1: Flowchart of Data Pre-processing Stages

The pre-processing of data provides data quality with high accuracy, completeness, consistency, interpretability and timeliness (Singhal & Jena, 2013). As shown in Figure 1, data cleaning is the stage to check inconsistency and eliminate incorrect values, and data integration includes combining data from all databases, data reduction decreases the volume of unnecessary data and data transformation is converting into the format supported by the software used (Sharma *et al.*, 2012; Singhal & Jena, 2013).

In the context of this research, diagnosis of thalassemia was conducted by collecting data from CBC test reports from the National Institute of Blood Diseases (NIBD) Karachi. CBC (Complete Blood Count) test is commonly ordered as a simple blood test for a routine medical assessment (Ogasawara *et al.*, 2019). The attributes shown in Table 1 contain all the essential parameters of CBC tests, Blood components and their reference range. The number of RBCs, the total amount of haemoglobin in the blood, the average size of an RBC, i.e. MCV, and the amount of space taken by an RBC in the blood, i.e. Haematocrit (Mekić *et al.*, 2018; Yin *et al.*, 2020). The other measurements, i.e. the concentration (MCHC) and the amount (MCH) of haemoglobin, also contain useful

Table 1: CBC Test Parameters (Source: Dataset)

Blood component	Abbreviation	Reference range
Haemoglobin*	HGB	Male: 13.5-17.5 g/dL
		Female: 12.0-16.0 g/dL
Hematocrit*	HT	Male: 41%-53%
		Female: 36%-46%
Mean corpuscular volume	MCV	80-100 µm3
White blood cells	WBC	4500-11,000/mm3
Red blood cells*	RBC	Male: 4.3-5.9 million/mm3
		Female: 3.5-5.5 million/mm3
Mean corpuscular haemoglobin	MCH	25.4-34.6 pg/cell
Mean corpuscular haemoglobin concentration	MCHC	31%-36% Hb/cell
Red Cell Distribution Width	RDW	Male: 11.6-14.6%
		Female: 12.2 to 16.1%
Platelets	Platelets	150,000-400,000/mm3

information about RBCs. These features of the CBC test can also be explained as Red Blood Cells (RBCs) carry oxygen, Hemoglobin (HB) is in RBC as an oxygen-carrier protein, White Blood Cells (WBC) are required to fight infections in the human body, and Hematocrit (HCT) represents the red blood cells percentage into the plasma or fluid component in Platelets of blood that causes blood clotting (Ogasawara *et al.*, 2019; Yin *et al.*, 2020). The test diagnoses and monitors the different types of blood diseases (e.g. anemia, thalassemia, iron deficiency), inflammatory diseases, infection, and malignancy (Mekić *et al.*, 2018; Ogasawara *et al.*, 2019).

There are many attributes in the CBC test reports, but only 6 attributes were selected for the risk prediction of thalassemia via WEKA software shown in Table 2. These six attributes include Gender, MCV, HCT, HGB, MCHC, and RDW and the ranges considered to differentiate between normal persons and thalassemia patients are shown in Table 1. Furthermore, the dataset was transformed into CSV and ARFF file format as supported by WEKA software. The data of 301 patients was provided by NIBD, containing the CBC test results of normal persons and patients suffering from thalassemia. The dataset included 1 nominal, whereas 5 numerical values and one column presented patients' status, i.e. POSITIVE or NEGATIVE. Later, the dataset was divided into training and test data by applying class "Resample, " a supervised instance filter on WEKA software used to produce a random subsample of a dataset with or without replacement.

Table 2: Dataset Attributes Used for Classification

Attribute	Attribute Value	Attribute Category
Gender	Male	1
	Female	2
MCV	<80	Microcytic
	80-100	Normal
HGB	<10	Severe
	10-12	Normal
HCT	<37	Low
	37-50	Normal
MCHC	<32	Hypochromic
	32-36	Normochromic
RDW	>14.6	High
	11.6-14.6	Normal

Data Analysis

WEKA Software

The algorithms used to predict and diagnose thalassemia were applied to the pre-processed dataset on WEKA software. WEKA is a popular machine learning software developed at the University of Waikato, New Zealand, containing a collection of algorithms and visualization

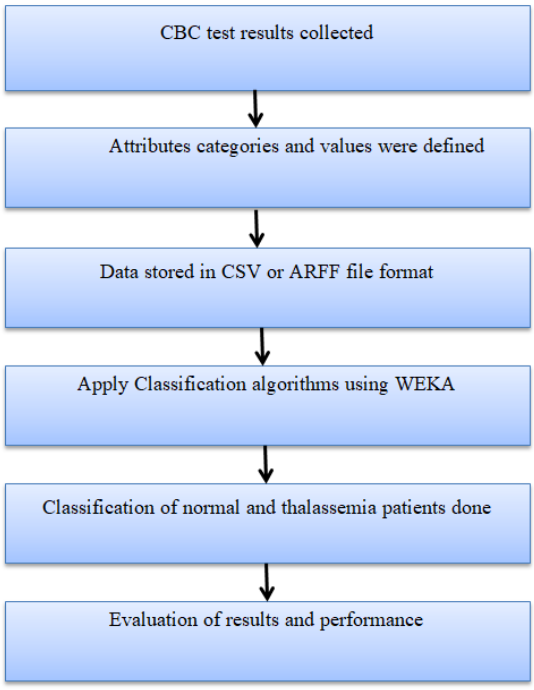

Figure 2: The Proposed Method

tools (Jovic *et al.*, 2014; Sultana *et al.*, 2016). These are used for predictive modelling and data analysis together with graphical user interfaces to get easy access to its functionalities. WEKA tool is platform-independent, easily available, and open source (Singhal & Jena, 2013).

In this research, WEKA explorer and experimenter were used to explore data and perform experiments and statistical tests. The implementation of the proposed method began by using four classifiers on WEKA on training and testing datasets to validate results with the highest accuracy and least mean square error (Jovic *et al.*, 2014). After classification on WEKA explorer, generated results were evaluated in the WEKA experimenter. Figure 2 below is the research flowchart which explains the proposed method.

Figure 2 above presents the steps followed for the execution of the proposed method. The process begins with collecting data, identifying the attributes, uploading the dataset on WEKA in CSV or ARFF format, applying various classification algorithms, predicting results, and evaluating the performance of each classifier is identified from the accuracy, precision, Recall, true positive rate, etc.

The Algorithms Used for Classification

J48 Decision Tree

A decision tree is a supervised classification algorithm that verifies a problem and determines the dependent variable values based on independent values via powerful approaches to data mining and knowledge discovery (Drazin & Montag, 2012; Sahu & Mehtre, 2015). J48 algorithm uses a divide and conquer approach to grow a tree, checking the cases belonging to the same class and labelling a leaf with that class (Drazin & Montag, 2012). Furthermore, Figure 3 presents the flowchart of the J48 decision tree algorithm, presenting the main purpose of the algorithm as a data mining classifier. It is a predictive machine learning model that decides conditions based on the training dataset's attributes. Further, it provides a pruned decision tree useful for addressing overfitting and classifying instances correctly. It splits the values based on the threshold specifying what is upper than, equal to, or less than that value, especially when dealing with continuous attributes (Sahu & Mehtre, 2015).

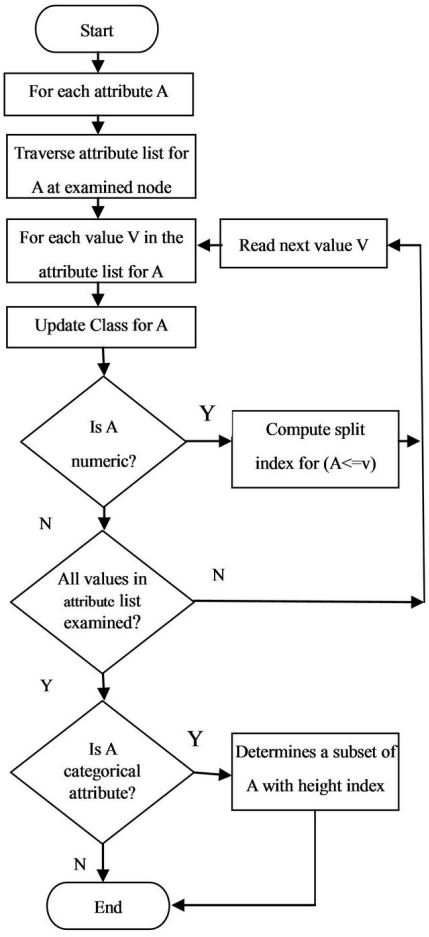

Figure 3: Flowchart for J48 Decision Tree Algorithm

Naïve Bayes Classifier

Naïve Bayes scans the training data and estimates all the probabilities with high learning efficiency (An *et al.*, 2017). The classification in this algorithm is based on the assumptions for identifying the object that is likely to be classified in the identified category, showing a direct or inverse relationship between one conditional probability (An *et al.*, 2017; Xu, 2018).

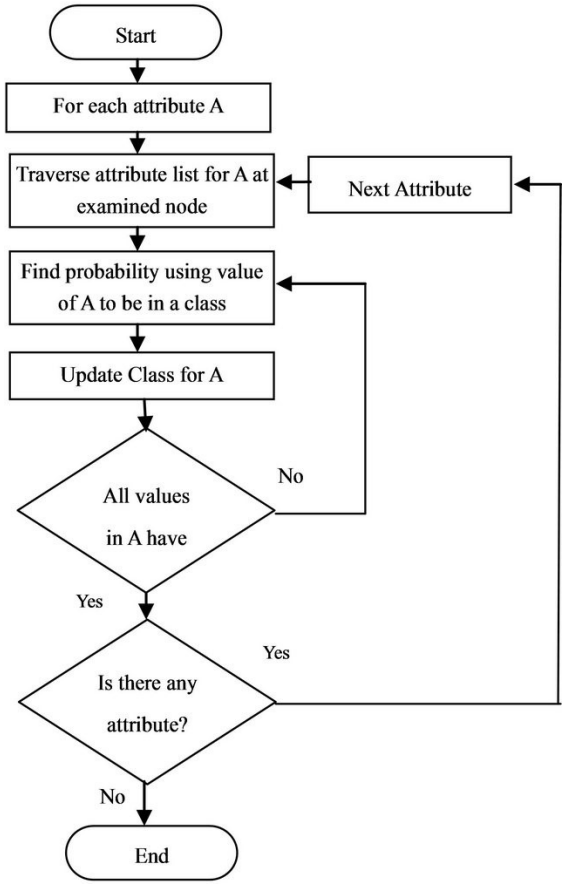

Figure 4: Flowchart for Naïve Bayesian Algorithm

Figure 4 illustrates the generative process of the Bayesian classifier, where each arrow represents conditional dependency among variables (Xu, 2018). It computes a probability after calculating the rate of values and their combinations in a specific dataset. Given the

probability of a single event, the Naïve Bayes classifier identifies another event that has already happened (Granik & Mesyura, 2017). The algorithm uses kernel density estimators dealing with numeric attributes using supervised discretization (An *et al.*, 2017; Jabbar & Samreen, 2016).

SMO Algorithm

SMO is an outstanding SVM algorithm in memory and efficiency requirements abbreviated as Sequential Minimal Optimization. It requires cross-validation to avoid overfitting and optimize parameters in the mathematical model (Luo *et al.*, 2016). SMO in WEKA is a supervised Support Vector Machine algorithm that analyses data and recognizes patterns. Furthermore, the SVM algorithm also possesses the same functional form of neural networks and functions as a radial basis (Zhang *et al.*, 2018). The working of the SMO algorithm is shown in Figure 5.

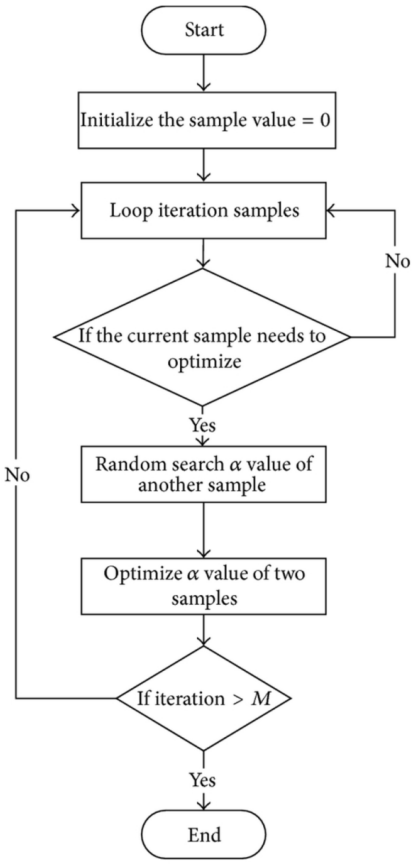

Figure 5: Flowchart for SMO Algorithm

SMO is generally made for a two-class classification problem to analyse the greatest separation between two classes, discovering the maximum distance to the nearby point. It has also been studied that SVM algorithms are based on the advances of the theory of machine learning in a high dimensional feature space that uses a hypothesis space of linear function and implements a learning bias derived from statistical learning theory (Luo *et al.*, 2016; Zengin *et al.*, 2017). In this research, a set of training data belonging to two classes is given with associated class labels, and the algorithm of SMO was applied using the WEKA tool to train the model.

Multilayer Perceptron

One of the significant models in artificial neural network (ANN) is the multilayer perceptron, which contains a single input layer, one or more than one hidden layer and a single output layer; the neurons are organised in these layers which are not connected in the same layer (Singh *et al.*, 2018). The values from the input node pass to the first hidden layer, and the same process continues with all the hidden layers until outputs are produced. Furthermore, the neurons in the input layer are the same as measurement variables for pattern problems, whereas the number of classes equals the number of neurons in the output layer (Ramchoun *et al.*, 2016; Singh *et al.*, 2018). The corresponding input generates the desired output in ANN, and the results can be viewed by computing the difference between the training set's desired output and the network's output on test data (Kwon *et al.*, 2017; Singh *et al.*, 2018). The multilayer perceptron is the most utilised form of neural network used as a back-propagation training algorithm. Excess connections can create a problem of overfitting; however, a lack of connections in the neural network may face the problem of insufficient parameters (Ramchoun *et al.*, 2016). In this research, the multilayer perceptron model was used, a feed-forward neural network model that maps the input data into desired or suitable outputs. The working of the multilayer perceptron algorithm is shown in Figure 6.

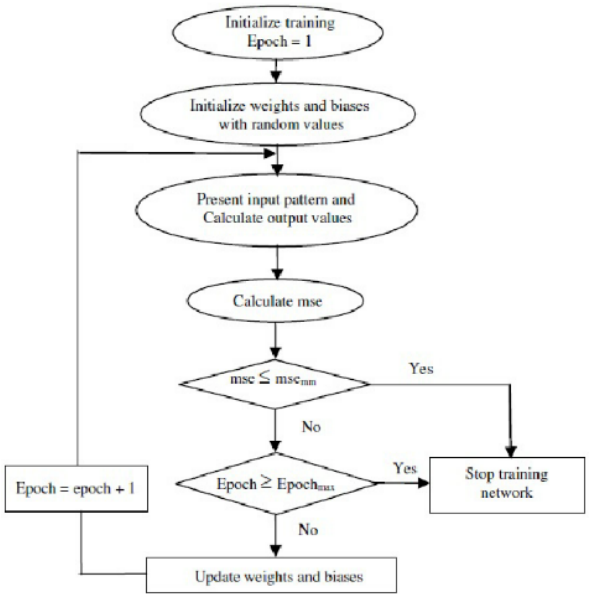

Figure 5: Flowchart for SMO Algorithm

Results & Discussion

J48 Decision Tree

The training data set was uploaded on WEKA Explorer, and after applying the J48 decision tree classification algorithm, the results derived are shown in Figure 7.

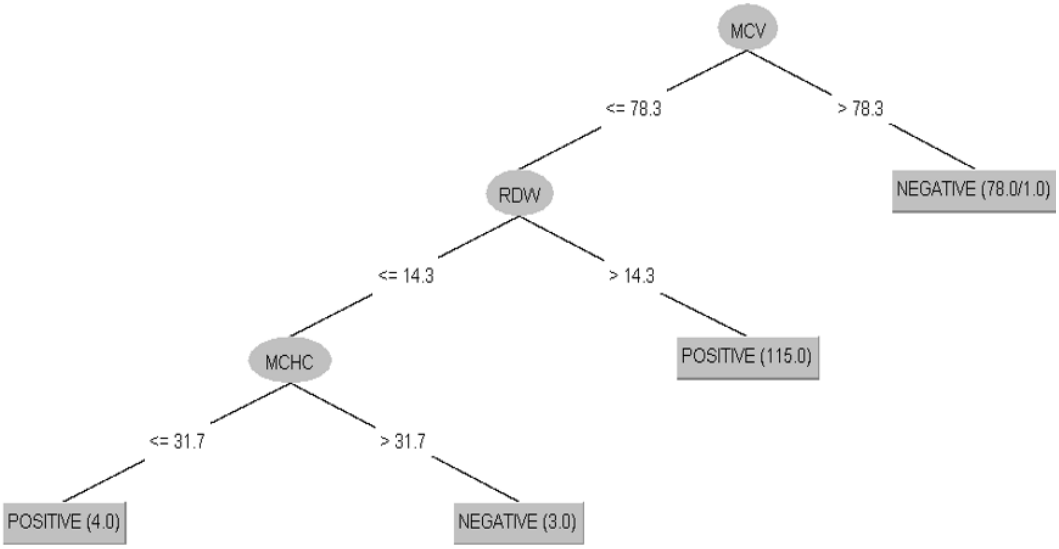

Figure 7: J48 Decision Tree

The decision tree in Figure 7 above reflects the rules that predicted the positive or negative risk of having thalassemia traits in the patient. The tree size is 7, whereas the number of leaves is 4. The WEKA tool constructed the pruned tree under the conditions stated in Figure 8.
Based on the J48 decision tree results per the trained model, rules and decisions are listed below in Table 3. The results were derived by showing the status of CBC test results as POSITIVE (Persons having traits of thalassemia) or NEGATIVE (Normal).

```
J48 pruned tree
------------------

MCV <= 78.3
|   RDW <= 14.3
|   |   MCHC <= 31.7: POSITIVE (4.0)
|   |   MCHC > 31.7: NEGATIVE (3.0)
|   RDW > 14.3: POSITIVE (115.0)
MCV > 78.3: NEGATIVE (78.0/1.0)

Number of Leaves  :     4

Size of the tree :      7
```

Figure 8: Decision Tree Rules

Table 3: Thalassemia Classification Rules

Rules	Decision
IF (MCV > 78.3) then,	NEGATIVE
Else if (MCV <= 78.3 and AND RDW > 14.3) then,	POSITIVE
Else if (MCV <= 78.3 and AND RDW <= 14.3 AND MCHC > 31.7) then,	NEGATIVE
Else if (MCV <= 78.3 and AND RDW <= 14.3 AND MCHC <= 31.7) then,	POSITIVE

There are different parameters based on which the accuracy of the results is defined stated below:
Relative Absolute Error: It is the way that provides the measure of the performance of a trained or predictive model used in data mining and machine learning. It is a general measure of accuracy or precision expressed as a ratio as a result of comparing a mean error to errors produced by the naïve model (Iyer *et al.*, 2015).

Confusion Matrix

The performance of a classification model is described by the confusion matrix on test data by trained data for which true values are known (Singh *et al.*, 2018). Besides, to measure the effectiveness of the classification model confusion matrix presents different combinations of actual or predicted values. The four entries of the confusion matrix are defined below:
TP Rate: The number of entries/ records classified as true and true in actuality represents the true positive rate (Singh *et al.*, 2018).
FP Rate: The number of entries/ records classified as true and false in actual represents the false positive rate (Singh *et al.*, 2018).
FN Rate: The number of entries/ records classified as false and where they were true in actual represents the false-negative rate (Singh *et al.*, 2018).
TN Rate: The number of entries/records classified as false and were false in actual represents the true negative rate (Singh *et al.*, 2018).
Accuracy: It is the ratio of the number of correct predictions to the total number of predictions (Wang & Li, 2019).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: It analyses how many are actually positive out of all the correctly predicted positive classes (Alam *et al.*, 2022; Wang & Li, 2019).

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is also known as the probability of detection, sensitivity, or TP rate and detect how much were correctly predicted out of all positive classes (Alam *et al.*, 2022; Wang & Li, 2019).

$$Recall = \frac{TP}{TP + FN}$$

F-measure: The weighted average of Recall and precision is known as F-measure (Alam *et al.*, 2022; Wang & Li, 2019).

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 9 presents the results of the J48 model developed using WEKA software. The resampling procedure for evaluating data mining tools is cross-validation, referring to the number of groups for the given dataset to be splitted. For the evaluation of different data mining algorithms in this research, the dataset was divided into 10-fold cross-validation. The rate of correctly classified instances was 97%, whereas the rate for incorrectly classified instances was 6%. The relative absolute error of the model was computed as 7.4438%. The detailed accuracy of the model by class is depicted in the form of TP Rate, FP Rate, Precision, Recall and F-measure. The confusion matrix in Figure 10 represents that the instances classified as POSITIVE are 117 (TP), whereas 77 (TN) were classified as NEGATIVE. The TP Rate, Precision, Recall and F-measure were derived as 0.975 for Class POSITIVE and 0.963 for Class NEGATIVE.

Furthermore, the test data was supplied on the J48-trained model to predict/diagnose thalassemia traits and visualize results. The results were transformed into a CSV file, and Table 4 shows the sample of 17 records of test data out of 101 evaluated using the trained model showing the 'predicted margin' and 'predicting status' for risk prediction of thalassemia.

```
Correctly Classified Instances         194               97      %
Incorrectly Classified Instances         6                3      %
Kappa statistic                          0.9375
Mean absolute error                      0.0357
Root mean squared error                  0.172
Relative absolute error                  7.4438 %
Root relative squared error             35.1074 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975     0.038     0.975       0.975   0.975      0.938    0.975     0.982     POSITIVE
                 0.963     0.025     0.963       0.963   0.963      0.938    0.975     0.942     NEGATIVE
Weighted Avg.    0.970     0.033     0.970       0.970   0.970      0.938    0.975     0.966

=== Confusion Matrix ===

   a   b   <-- classified as
 117   3 |   a = POSITIVE
   3  77 |   b = NEGATIVE
```

Figure 9: J48 Results

Table 4: J48 Classifier Output Results using WEKA Tool

MCV	HCT	HGB	MCHC	RDW	Gender	'prediction margin'	'predicted Status'
66.4	4.9	22.3	29.5	31.5	2	1	POSITIVE
74.6	4.9	19.3	22.6	33.7	1	1	POSITIVE
69	9.3	21.7	26.7	32.6	1	1	POSITIVE
73.1	7.4	25.8	32	26.4	2	1	POSITIVE
77.3	6.8	34.2	31.3	33.6	1	1	POSITIVE
71.2	6.8	21.5	30.1	36.7	1	1	POSITIVE
80.1	10.5	11.6	35.2	42.1	1	-0.974359	NEGATIVE
75.3	6.7	17.6	27.5	31.5	2	1	POSITIVE
76.4	6.9	27.7	25.9	34.1	1	1	POSITIVE
69	9.4	21.9	27.4	29.6	1	1	POSITIVE
68.9	9.9	19.9	31.2	26.9	1	1	POSITIVE
73.1	9.1	21.4	30.5	32.1	2	1	POSITIVE
83.5	10.1	11.7	33.6	42.4	1	-0.974359	NEGATIVE

82.4	10.3	12.7	33.6	44	1	-0.974359	NEGATIVE
57.9	5.9	16.9	30.1	33.5	1	1	POSITIVE
82.6	11.9	14	33.9	43.2	1	-0.974359	NEGATIVE
80.9	12.4	13.6	33.5	41.2	1	-0.974359	NEGATIVE

Naïve Bayes Classifier

The results for the Naïve Bayes classifier are shown in the Figure 10 below.

After applying the Naïve Bayes algorithm on the WEKA tool, the correctly classified instances based on the given data were 99%; however, only 1% of data was incorrectly classified. Furthermore, as shown in Figure 5.6, the relative absolute error of Naïve Bayes results was 1.445%. After applying the Naïve Bayesian algorithm on the training dataset, results showed that there is only one entry for FP and FN, suggesting that the model predicting thalassemia with a 99.0 % accuracy rate; therefore, the TP rate for Class POSITIVE was identified as 99.2% and 98.8% for Class NEGATIVE suggesting that 119 CBC test results for thalassemia are positive where 79 are negative based on the given data.

```
Correctly Classified Instances         198               99      %
Incorrectly Classified Instances         2                1      %
Kappa statistic                          0.9792
Mean absolute error                      0.0069
Root mean squared error                  0.0684
Relative absolute error                  1.445  %
Root relative squared error             13.9575 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.013    0.992      0.992    0.992      0.979    1.000     1.000     POSITIVE
                 0.988    0.008    0.988      0.988    0.988      0.979    1.000     1.000     NEGATIVE
Weighted Avg.    0.990    0.011    0.990      0.990    0.990      0.979    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 119   1 |   a = POSITIVE
   1  79 |   b = NEGATIVE
```

Figure 10: Naïve Bayes Results

SMO Algorithm

The results for the SMO Algorithm are represented in Figure 11, analysed using the WEKA tool.

The results derived from the SMO model on the given dataset provided results with 3.1236% relative absolute error, where 98.5% were correctly classified instances, whereas 1.5% were incorrectly classified instances. Moreover, as shown in figure 5.7, after applying the

```
Correctly Classified Instances         197               98.5    %
Incorrectly Classified Instances         3                1.5    %
Kappa statistic                          0.9689
Mean absolute error                      0.015
Root mean squared error                  0.1225
Relative absolute error                  3.1236 %
Root relative squared error             24.9999 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.000    1.000      0.975    0.987      0.969    0.988     0.990     POSITIVE
                 1.000    0.025    0.964      1.000    0.982      0.969    0.988     0.964     NEGATIVE
Weighted Avg.    0.985    0.010    0.986      0.985    0.985      0.969    0.988     0.980

=== Confusion Matrix ===

   a   b   <-- classified as
 117   3 |   a = POSITIVE
   0  80 |   b = NEGATIVE
```

Figure 11: SMO Algorithm Results

SMO algorithm on the training dataset, results showed that there are three entries for FP and no entry for FN, suggesting that the model predicting thalassemia with a 98.5 % accuracy rate; therefore the TP rate for Class POSITIVE was identified as 97.5% and 100% for Class NEGATIVE suggesting that 117 CBC test results for thalassemia are positive where 80 are negative based on the given data.

Multilayer Perceptron

After applying a Multilayer Perceptron Neural Network on the training dataset, the results shown in the Figure 12 below were derived.

```
Correctly Classified Instances         196               98      %
Incorrectly Classified Instances         4                2      %
Kappa statistic                          0.9585
Mean absolute error                      0.0169
Root mean squared error                  0.1024
Relative absolute error                  3.5101 %
Root relative squared error             20.903  %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.013    0.992      0.975    0.983      0.959    0.998     0.999     POSITIVE
                 0.988    0.025    0.963      0.988    0.975      0.959    0.998     0.998     NEGATIVE
Weighted Avg.    0.980    0.018    0.980      0.980    0.980      0.959    0.998     0.998

=== Confusion Matrix ===

   a   b   <-- classified as
 117   3 |   a = POSITIVE
   1  79 |   b = NEGATIVE
```

Figure 12: MLP Results

Figure 5.8 presents the results for training data using the Multilayer Perceptron classifier on the WEKA tool. The algorithm's rate of correctly classified instances is 98%, whereas the rate for incorrectly classified instances is 2%. As shown in Figure 5.8, the relative absolute error for the trained model was derived to be 3.5101%. Furthermore, it shows that there are three entries for FP and only one entry for FN, suggesting that the MLP model predicts thalassemia with a 98.0 % accuracy rate; therefore, the TP rate for Class POSITIVE was identified as 97.5% whereas 98.8% for Class NEGATIVE suggesting that 117 CBC test results for thalassemia are positive and 79 are negative based on the given data.

Comparison of Algorithms' Results through WEKA Experimenter

This section of the report compares all four classifiers using the WEKA experimenter based on precision, Recall and F-measure. A paired t-test was performed to identify the difference between the algorithms used. It was used in this research using the WEKA experimenter to compare the results among all four algorithms. The data mining measures that are useful to analyse the performance of each algorithm are precision, Recall and F-measure.

Figures 13, 14 and 15 illustrate the WEKA experimenter's snapshots using precision, Recall, and F-measure. Paired t-test was applied in these experiments to identify the algorithm with the highest accuracy for the given dataset.

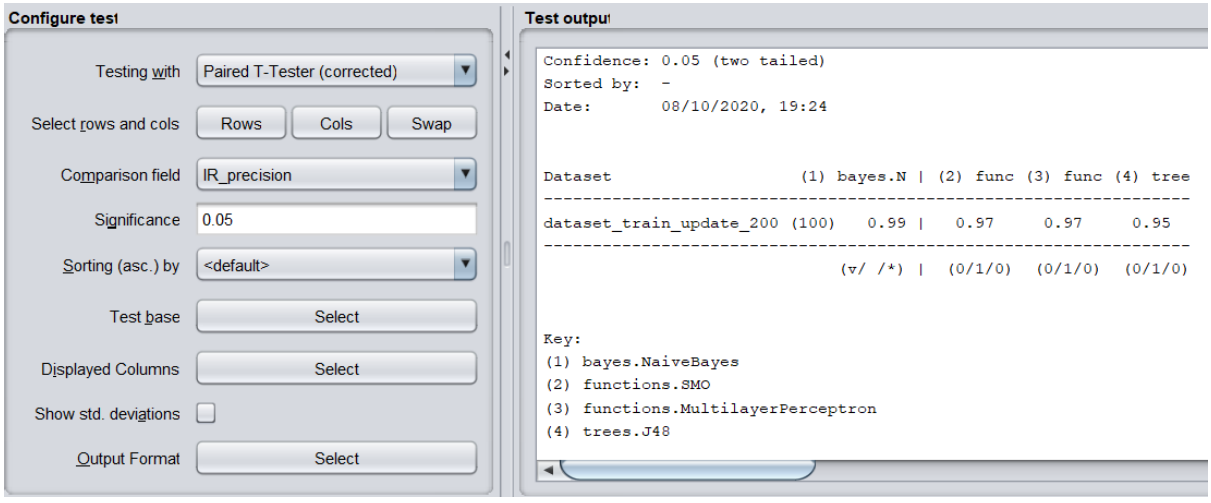

Figure 13: Comparison of J48, NB, SMO, and MLP algorithms using WEKA experimenter using Precision

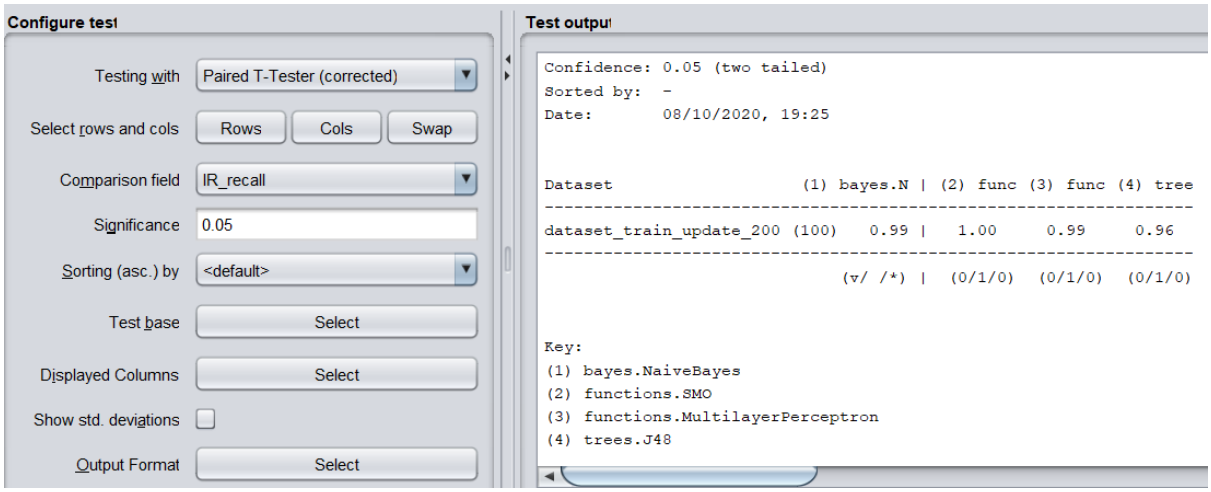

Figure 14: Comparison of J48, NB, SMO, and MLP algorithms using WEKA experimenter using Recall

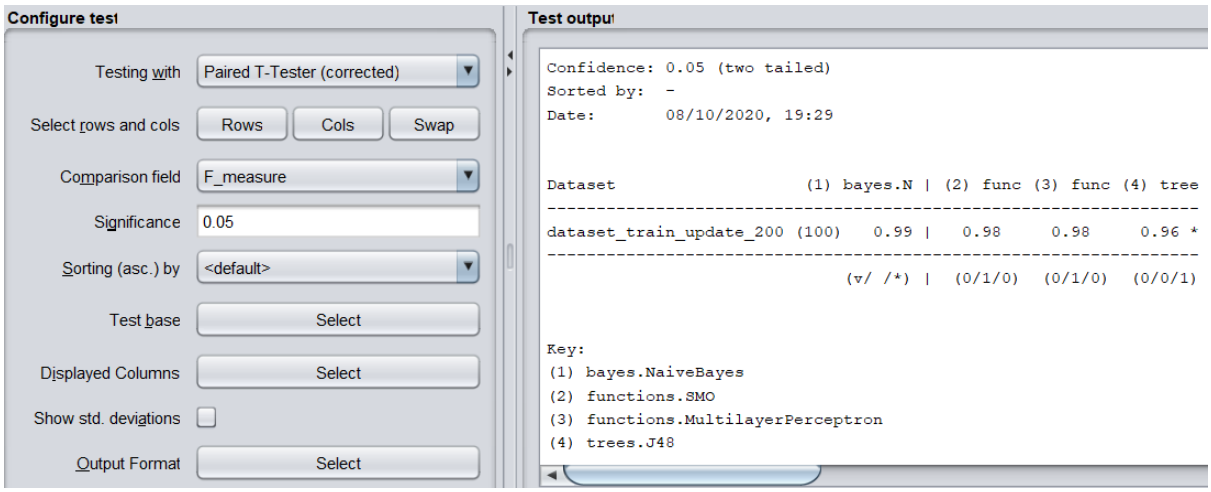

Figure 15: Comparison of J48, NB, SMO, and MLP algorithms using WEKA experimenter using F-measure

It was evaluated that Naïve Bayes performed the risk prediction of thalassemia with the highest performance among all four classifiers with precision, Recall and F-measure of 99.0%. The comparative performance for all four classifiers is also visualised based on relative absolute error, as shown in the figure below.

It can be viewed in Figure 16 that the least root mean square error of 1.445% was for the Naïve Bayes model in comparison with models of J48, SMO and MLP. Hence, it proved that based on the given dataset, Naïve Bayes provided results with the highest accuracy and least mean square error for the risk prediction of thalassemia.

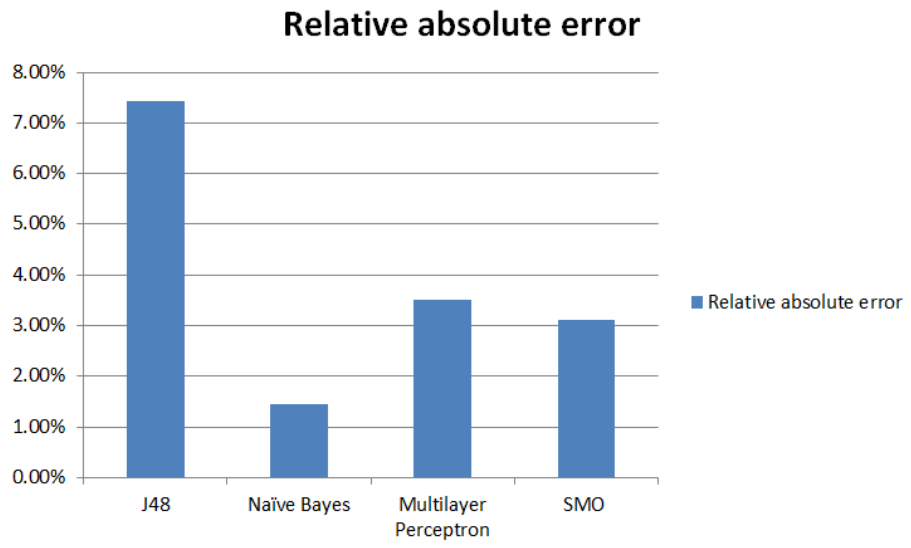

Figure 16: Comparison of J48, NB, SMO, and MLP algorithms using WEKA experimenter using F-measure

CONCLUSION

A rapid increase is observed in inherited haemoglobin disorders, and despite the efforts to control the spread of these diseases, the number of major cases leading to death is increasing. Thalassemia is one of these inherited blood disorders that has received excessive attention over the years in medical data mining. Diagnosis of thalassemia is based on some characteristics derived after performing the CBC test. Therefore, various data mining techniques were utilised in this research to identify the hidden patterns in the given dataset of CBC test results for thalassemia. The research was conducted for the diagnosis or risk prediction of thalassemia using data mining classification algorithms. The selected classification algorithms for this study were Naïve Bayes, Multilayer Perceptron, J48 and SMO using the WEKA data-mining tool.

Further, it also analysed the best potential classification algorithm for thalassemia by determining accuracy, precision, Recall, and F-measure using WEKA Experimenter. The data of 301 CBC results provided by NIBD was divided into training and test data, and it was pre-processed by applying a resample filter on the WEKA tool. For the evaluation of different data mining algorithms in this research, the dataset was categorised into 10-fold cross-validation. The results were transformed into a CSV file showing the 'predicted margin' and 'predicting status' for risk prediction of thalassemia. Results revealed that after applying four different algorithms on the training dataset, the Naïve Bayes model predicted thalassemia with the highest accuracy of 99.0 % accuracy rate suggesting that 119 CBC test results for thalassemia are positive, where 79 are negative based on the given data.

Future Work

This research was conducted by applying different classification algorithms to get the best prediction or diagnosis of thalassemia based on the dataset of 301 constructed of CBC results. In the future, the research work can be expanded by using more types of data mining algorithms, such as clustering or association rules, to identify the best-performing algorithm based on the given dataset. Association rules can be used to identify important relations among different attributes of the given dataset to detect the type of blood disease. Furthermore, in the future, the aim of the analysis could be slightly different, such as identifying the survival rate of thalassemia patients, classifying between anaemia and thalassemia, or developing a classification model for the different types of thalassemia. Moreover, a big dataset can be selected or requested from NIBD or any other organization in Pakistan to present more genuine and authentic results, training the model with the highest possible accuracy. In addition, the big dataset can be divided into training and test data by setting up different split-up percentages for training data to analyse which dataset is capable of the highest performance results. Future work can be conducted by comparing the CBC report's other parameters and identifying the core relationship with the variables that can cause different blood diseases. The data mining techniques are very powerful, but they should be used with great care in the field of medicine, and therefore, there is a need to discover the best mining algorithm for the specified medical area in the future.

Acknowledgements

We express our deepest gratitude to our independent research project supervisor, Dr. Raheela Asif (Department of Software Engineering, NED University of Engineering & Technology, University Road, Karachi 75270, Pakistan), for her relentless assistance throughout the research. We are also grateful to NIBD (National Institute of Blood Diseases), Karachi City, Sindh, Pakistan, for providing data for this research.

REFERENCES

Abdullah, M., & Al-Asmari, S. (2016). Anemia types prediction based on data mining classification algorithms. In Communication, management and information technology (pp. 629-636). CRC Press.

Alaa, M., & Shurrab, A. H. (2017). Blood tumor prediction using data mining techniques. *Health Informatics—An International Journal, 6*, 23-30.

AlAgha, A. S., Faris, H., Hammo, B. H., & Ala'M, A.-Z. (2018). Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine. *Artificial intelligence in medicine, 88*, 70-83.

Alam, B. R., Khatun, M. S., Taslim, M., & Hossain, M. A. (2022). Handling Class Imbalance in Credit Card Fraud Using Various Sampling Techniques. *American Journal of Multidisciplinary Research and Innovation, 1*(4), 160-168.

Amin, M. N., & Habib, M. A. (2015). Comparison of different classification techniques using WEKA for hematological data. *American Journal of Engineering Research, 4*(3), 55-61.

An, Y., Sun, S., & Wang, S. (2017). Naive Bayes classifiers for music emotion classification based on lyrics. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS),*

Asere, G. F., & Botson, D. E. (2021). Data Mining technology as a tool for supporting analytical decision making process in Health Information Management System (HIMS). *American Journal of Agricultural Science, Engineering, and Technology, 5*(2), 139-147.

Drazin, S., & Montag, M. (2012). Decision tree analysis using weka. Machine Learning-Project II, University of Miami, 1-3.

Egejuru, N. C., Olusanya, S. O., Asinobi, A. O., Adeyemi, O. J., Adebayo, V. O., & Idowu, P. A. (2019). Using data mining algorithms for thalassemia risk prediction. *International Journal of Biomedical Science and Engineering, 7*(2), 33-44.

Elshami, E. H., & Alhalees, A. M. (2012). Automated diagnosis of thalassemia based on datamining classifiers. The *international conference on informatics and applications (ICIA2012)*

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*

Hasani, M., & Hanani, A. (2017). Automated diagnosis of iron deficiency anemia and thalassemia by data mining techniques. *International Journal of Computer Science and Network Security (IJCSNS), 17*(4), 326.

Herbert, L., Muncie, J., & Campbell, J. (2009). Alpha and beta thalassemia. *Am Fam Physician, 80*(4), 339-344.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

Jabbar, M., & Samreen, S. (2016). Heart disease prediction system based on hidden naïve bayes classifier. *2016 international conference on circuits, controls, communications and computing (I4C).*

Jameel, T., Baig, M., Ahmed, I., Hussain, M. B., & bin Doghaim Alkhamaly, M. (2017). Differentiation of beta thalassemia trait from iron deficiency anemia by hematological indices. *Pakistan journal of medical sciences, 33*(3), 665.

Jatoi, S., Panhwar, M. A., Memon, M. S., Baloch, J. A., & Saddar, S. (2018). Mining complete blood count reports for disease discovery. *International Journal of Computer Science and Network Security, 18*(1), 121-127.

Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. *Procedia computer science, 72*, 306-313.

Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining. *2014 37th International convention on information and communication technology, electronics and microelectronics (MIPRO)*

Kamil, S., Kousar, S., Rafique, S., Qadir, H., Farooqui, W., Tauheed, M., Kamil, N., & Liaquat, A. (2021). Frequency of carrier state of thalassemia and various hemoglobinopathies in tertiary care hospital of Pakistan. *IJEHSR-International Journal of Endorsing Health Science Research, 9*(2), 195-200.

Khaliq, S. (2022). Thalassemia in Pakistan. *Hemoglobin, 46*(1), 12-14.

Kwon, K., Kim, D., & Park, H. (2017). A parallel MR imaging method using multilayer perceptron. *Medical physics, 44*(12), 6209-6224.

Luo, Y., Xiong, Z., Xia, S., Tan, H., & Gou, J. (2016). Classification noise detection based SMO algorithm. *Optik, 127*(17), 7021-7029.

Mdaghri, Z. A., El Yadari, M., Benyoussef, A., & El Kenz, A. (2016). Study and analysis of data mining for healthcare. *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt),*

Meena, K., Tayal, D. K., Gupta, V., & Fatima, A. (2019). Using classification techniques for statistical analysis of Anemia. *Artificial intelligence in medicine, 94*, 138-152.

Mekić, M. S., Pedišić, I., Šobat, H., Boras, V. V., Kirac, I., Štefančić, L., Šekerija, M., Vrdoljak, B., & Vrdoljak, D. V. (2018). The role of complete blood count parameters in patients with colorectal cancer. *Acta Clinica Croatica, 57*(4), 624.

Ogasawara, A., Matsushita, H., Tanaka, Y., Shirasugi, Y., Ando, K., Asai, S., & Miyachi, H. (2019). A simple screening method for the diagnosis of chronic myeloid leukemia using the parameters of a complete blood count and differentials. *Clinica Chimica Acta, 489*, 249-253.

Origa, R. (2017). β-Thalassemia. *Genetics in Medicine, 19*(6), 609-619.

Ramchoun, H., Ghanou, Y., Ettaouil, M., & Janati Idrissi, M. A. (2016). Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence, 4,*, 26-30. https://doi.org/http://doi.org/10.9781/ijimai.2016.415

Sahu, S., & Mehtre, B. M. (2015). Network intrusion detection system using J48 Decision Tree. 2015 *International Conference on Advances in Computing, Communications and Informatics (ICACCI),*

Saichanma, S., Chulsomlee, S., Thangrua, N., Pongsuchart, P., & Sanmun, D. (2014). The observation report of red blood cell morphology in Thailand teenager by using data mining technique. Advances in hematology, 2014.

Sharma, N., Bajpai, A., & Litoriya, M. R. (2012). Comparison the various clustering algorithms of weka tools. *facilities, 4*(7), 78-80.

Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine, 13*(sup1), 121-124.

Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee), 2*(6), 250-253.

Sultana, M., Haider, A., & Uddin, M. S. (2016). Analysis of data mining techniques for heart disease prediction. *2016 3rd international conference on electrical engineering and information communication technology (ICEEICT),*

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology, 5*(5), 241-266.

Wang, R., & Li, J. (2019). Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,*

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science, 44*(1), 48-59.

Yin, Y., Zhang, Y., Wang, D., Han, X., Chu, X., Shen, M., & Zeng, X. (2020). Complete blood count reflecting the disease status of giant cell arteritis: A retrospective study of Chinese patients. *Medicine, 99*(39).

Zengin, K., Güngör, C., & Eşgi, N. (2017). Heart Rate Signal Classification By Smo Algorithm. *International Research Journal of Mathematics, Engineering and IT , 4*(12).

Zhang, Q., Wang, J., Lu, A., Wang, S., & Ma, J. (2018). An improved SMO algorithm for financial credit risk assessment–evidence from China's banking. *Neurocomputing, 272*, 314-325.