Determining Local Hematology Reference Ranges: A Data-driven Approach

K.A.Hasara Semini, H.A.Caldera
University of Colombo School of Computing
University of Colombo, Colombo
Sri Lanka

***Abstract*—Hematology is the study of blood, blood-forming organs, and blood diseases. Hematological tests such as Full Blood Count (FBC) can be used to diagnose a wide range of infections and diseases by comparing their results with the standard hematology reference (SHR) ranges. These ranges were established many years ago by considering the Caucasian population and all countries have used them until recent times to measure the healthiness of the people. But these reference ranges can be varied according to various reasons such as dietary habits, geographical location, climate, environmental factors, etc., and the use of them by all countries may not be correct. Many researchers have started research in finding Local Hematology Reference (LHR) ranges. Most of them used statistical analyses which have their limitations. Machine learning is a solution to overcome those limitations. Finding an approach to determine the LHR range based on machine learning techniques is the goal of this research. The dataset was generated using FBC test reports in Sri Lanka. The LHR range of WBC count of healthy adults in Sri Lanka is only addressed in this research. A difference between the SHR range of WBC and the LHR range of WBC is observed.**

Keywords*—Hematology science; standard hematology reference range; domestic hematology reference ranges; local hematology reference range; machine learning; white blood cell count

I. Introduction

Medical Science helps to maintain and restore health. It is a combination of diagnosis, prognosis, treatment, and prevention of disease. Medical science can be divided into several sub-sections such as Cardiology, Anesthesiology, Dentistry, Hematology, and Physiology. Hematology consists of four major components as plasma, red blood cells, white blood cells, and platelets. Each of these four components consists of other sub-components. For example, white blood cells contain lymphocytes, monocytes, eosinophils, basophils, and neutrophils. Each of these components has a reference range that is considered to measure the healthiness of a person in Health-related fields.

When considering these hematology reference ranges, they are important to monitoring pathophysiological changes after getting infected with a disease. It can be used to detect diseases such as Dengue fever, HIV, cancer, etc., and track the effects of the given drugs or vaccines for clinical observations. The set of values for hematology reference ranges that are accepted worldwide is called "Standard hematology reference ranges" which were determined many years ago by doing some researches for the Caucasian populations. A Caucasian population is a group of people who are originated from Europe and are also commonly known as "white" or "white-skinned" people.

Gradually, people realize that the Standard hematology reference range can vary due to many reasons such as age, gender, genetics, attitudes, lifestyle, ethnic origin, dietary habits, geographical location, climate, environmental factors, etc.[1]–[3]. Hence the Clinical and Laboratory Standards Institute (CLSI) has recommended that a domestic hematology reference range should be established for each region [2], [4]. As a result, hematology reference ranges per country are the focus of this paper.

When considering Sri Lanka, the population is not a Caucasian population due to their significant variation on factors for which the standard hematology reference ranges were determined.. For example, suppose the minimum hemoglobin reference value of Sri Lankan healthy adults and the minimum value of the standard hemoglobin reference range are 'x' and 'y' respectively and 'x' has become less than 'y' due to the nutrition factors of Sri Lankan population. The doctors give medicines for hemoglobin deficiency to people whose hemoglobin values in between 'x' and 'y' until the local reference ranges are established. But these people may be healthy and do not need to take any medicine as far as the local conditions are concerned. These kinds of situations can happen to any other hematology attribute too. Taking medicine without any illness may cause dangerous side effects too. Therefore, a local hematology reference range is a mandatory thing for every country which is not under the Caucasian population category. As the local hematology reference ranges for Sri Lanka have not yet been determined, the research conducted in the paper is an attempt to fill this gap. Hence, the dataset used in this research was taken by considering the Sri Lankan population.

The paper is organized as follows. We first focus on related works in Section II investigating the impact of standard reference ranges on different populations and different methods in establishing the domestic reference ranges. Section III describes the proposed method in meeting the requirement. Results and Evaluations are described in Section IV. Finally, Section V presents the conclusion of this study.

II. Background Study

Eastern India, China, Morocco, Ethiopia, Sudan, Malawi, Nigeria, and many other African countries have successfully done many kinds of research and established the domestic hematology reference ranges for their countries which contain different values rather than the standard hematology reference range values [1], [2], [5]–[7].

Eastern India researchers had found lower hemoglobin (HB) and platelet (PLT) values when compared to the standard hematology reference ranges. The difference was statistically significant only for the platelet count [2]. The lower hematology reference range values for White Blood cells (WBC), Hemoglobin (HB), Hematocrit (HCT), Mean Corpuscular Volume (MCV), and platelet (PLT) counts which were compared to the standard hematology reference ranges had been found by researchers in Malawi. The research was driven by categorizing the dataset both gender-wise and age-wise. These researchers also compared the received male hematology reference ranges with domestic hematology values of other African countries and found that the male Malawians have lower HB and HCT values than others [5]. A lower reference range for WBC than the standard WBC reference range was found by the Sudanese research team who researched to find a local WBC reference range. The dataset which was used to take this result belonged to a particular city in Sudan. The research team identified that the received result was different by studying two other researches which were also conducted to find domestic WBC reference value considering another two cities in Sudan [1]. A Nigerian research team had analyzed blood and urine samples of males, pregnant females, and non-pregnant females to establish local reference ranges for their country. The different local reference ranges for males and females as well as pregnant females and non-pregnant females were received as the result of the research. They compared their result with reference ranges in the USA. Glucose levels, Urea levels, enzyme levels in the Nigerian population were higher than the USA reference ranges [8]. A local hematology reference range was established by considering both male and female healthy adults in Togo. The blood samples for the research have been taken from 1349 donors who were discovered as healthy. The received ranges have differed from other African countries [9]. In another research, hematology reference ranges were established based on the population of old people in rural southwest Uganda. The received ranges were compared with age groups which are categorized as adults, old people (age between 50 and 65), and very old people (65+). The ranges were changed between age groups [10]. Another research team has established Hematology Reference ranges among Healthy Adults in Bamenda, North West Region of Cameroon. The statistical analysis was used to determine the ranges [11].

Most of these researches were driven by considering age, gender, the country-wise or different area in the same country and were found different reference ranges if the considered dataset did not belong to the Gaussian country. Hence the researchers have focused to establish a local reference range to their countries. The blood samples were taken from the donors who were selected carefully by considering various conditions such as BMI value, medical history, and so on to generate a dataset. The number of data records that were used to establish the local hematology reference range in those countries was less than 1000. It may not enough to decide a local reference range for the whole country. Most of them used statistical analysis such as the Mann-Whitney U test, Chi-square test, t-test, and so on to establish those local hematology reference ranges. Hence, the dataset is statistically analyzed and described by using statistical theories. Most statistical theories define limitations to take a good result. In general, if the dataset contains huge data records then it will help to make a better output. But, because of the data limitation consideration, a huge dataset may not even give efficient results in statistical analysis. Every statistical theory may not apply to a certain dataset. And statistical analysis cannot be efficiently used to make predictions or find hidden patterns.

In general, most data-driven studies were handled by using data mining and machine learning concepts to avoid the above-mentioned problems. Many kinds of research in the medical sector have already used data mining and machine learning concepts to get efficient outcomes[12]–[14].

III. Methodology

Sri Lanka is a South Asian country. It is a tropical island with hot and humid weather all over the year. The population used in this research is Sri Lankan adults with age over 21 years and the data set used for the study is extracted from the Full Blood Count (FBC) test reports from a sample of the population. The sample contains both healthy and unhealthy people. Ethical approval is mandatory for this research as the research works with human medical data.

When considering the hematology components, some of them (dependent components) have an interrelationship with other components (independent components). The normal values of dependent hematology components may change as the independent component changes. For example, White blood cells (WBC) have two categories as granulocytes and non-granulocytes. Each of these categories has five types of white blood cells in human blood as Lymphocytes, Monocytes, Eosinophils, Basophils, and Neutrophils. The task of each of these white blood cells is given below.

- Lymphocyte: Generate antibodies, fight with inflection and viral cells. Also, it has B-lymphocyte, T-lymphocyte, and natural killer cells.
- Monocyte: Attacks chronic infections if present.
- Basophil: Sensitive when occurring allegories.
- Eosinophil: Working with the immune system's responses.
- Neutrophil: Helps to remove fungi and bacteria from the body.

All of these white blood cell types help to be healthy and highly affect the hematology components. When you take the FBC report, it shows leukocyte value (total WBC value) as well as values for five types of WBC cells. In this research, only the leukocyte value was considered. The normal value for leukocytes is taken when all WBC cell types follow normal values.

Data mining techniques and Machine Learning concepts are used to define the new approach which will be introduced via this research to determine local hematology reference ranges.

The general steps involved in the new approach are given in Fig. 1. A local hematology reference range for WBC count is only checked using the approach and a similar procedure can be used to determine local reference ranges of other hematology components.

A. Data Preparation

Various types of tests are used in hematology science such as the Full Blood Count (FBC) test, C-reactive protein (CRP) test, liver function tests, thyroid function tests, etc. Only the FBC test reports are considered to extract the required hematology reading as it is the most common test and shows all the main components of hematology.

As the initial data set obtained contained various types of blood tests, for example, WBC, on multiple rows, it was processed and restored to reflect all the individual's data on a single row as shown in Fig. 2. The values of Lymphocyte, Monocyte, Basophil, Eosinophil, and Neutrophil are stored as count values instead of percentages. A bivariable "Health State" is generated by using the standard hematology reference ranges of other hematology tests without considering the total WBC (leukocyte) value as a local hematology reference range for WBC count has experimented in this research. The Health State takes two values; "H" to identify the healthy persons and "UH" to unhealthy.

The 604 records are taken from the initial dataset for this research.

- 477 people in the dataset whose Lymphocyte value, Monocyte value, Basophil value, Eosinophil value, and Neutrophil value lie between the standard hematology reference range.
- 110 people in the dataset whose health state is equal to "H" generated by considering the other hematology reference ranges without considering the total WBC hematology reference range.
- 90 people in the dataset are healthy by considering all components in standard hematology reference ranges including the total WBC hematology reference range. This group of people will be called "SH" in further works in this research.

Preprocessing is one of the major steps in the Data Mining process and the dataset is explored for this purpose. The following preprocessing tasks were exercised to prepare the dataset for the data mining task.

1) *Removed unique attributes from the dataset.*
2) *Removed unwanted fields from the dataset.*
3) *Converted all data types to nominal.*
4) *Grouped data as necessary.*
5) *Set class attribute as "health state".*
6) *Visualized data:* This helps to take a proper image of the selected dataset. The visualization of the data is very important to make pre-decisions about every attributes before the analysis.

B. Classification Model

Building a classification data model is the next step. Classification can be applied to both structured and unstructured datasets. The classification technique categorizes new data instances into classes by learning through an experienced/training dataset where the classes of each instance of it are already given. When applying a classification algorithm for the training dataset, it separates each data point into given classes. This is the model that is created using a classification algorithm. The accuracy of the classifier is assessed on a separate data set called testing data set by checking the correctly classified records. There are two types of classification algorithms that are mostly used for data analysis. In Binary Classification, the class attribute has only two categories, but it has more than two categories in Multi-Class Classification. Each classification assumes that one data point belongs to only one class and non-class attributes of each data point must be independent of each other and discrete.

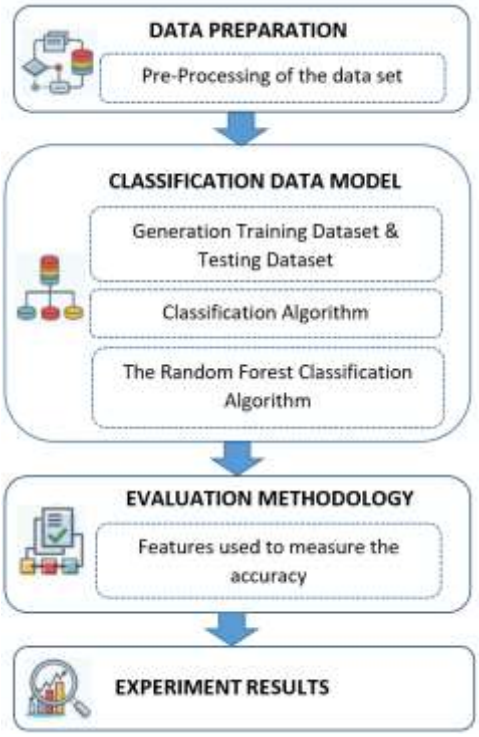


Fig. 1. Methodology Diagram.

Fig. 2. Dataset.

It is needed to find which classification algorithm is suited for analyzing the dataset and achieve the final goal. There are many classification algorithms available in literature but the tree-like algorithms are the most commonly used because of their ease of implementation and easier to understand compared to other classification algorithms [15]–[17]. Random Forest, Random Tree, REP Tree, LMT, J48, and Hoeffding Tree algorithms are such algorithms.

1) The Random Forest Tree (RFT) classification algorithm: Random Forest Tree (RFT) algorithm is a supervised learning algorithm categorizing under the decision trees. The word “Forest" is included in the name of the algorithm hence it makes a bunch of trees. It makes 500 trees in default. The bootstrapping techniques are used to make 500 samples from the dataset to make 500 trees. The algorithm uses ensemble methods to produce the final solution. The random forest algorithm uses horizontal filtering (make samples by considering the different variations of the dataset) as well as vertical filtering (make samples by considering ranks of the attributes) techniques. The main advantage of using a Random Forest algorithm is controlling the overfitting of the data set with predictions [18], [19].

2) Random Tree (RT) classification algorithm: Random Tree (RT) algorithm is also a supervised learning algorithm. The algorithm uses ensemble methods bagging technique to generate a random set of data from which to build a decision tree. The random tree uses k number of attributes at each node of the decision tree but no control over the overfitting of the data set [20].

3) REP Tree classification algorithm: The REP (Reduced Error Pruning) Tree is the simple and most comprehensible decision tree which is used to reduce error pruning strategy. It's a convenient decision tree learner that creates a decision or regression tree with feature selection using information gain based on class variable and prunes it with reduced error pruning. The tree traversal was performed using the REP algorithm from bottom to top, and then each internal node was checked and replaced with the most frequent class with the most concern about the tree accuracy, which must not be reduced. This technique will be repeated until no further pruning reduces the accuracy [20].

4) LMT classification algorithm: The basic structure of a Logistic Model Tree (LMT) is a regular decision tree structure with logistic regression functions at the leaves. A tree structure is made up of a set of non-terminal nodes and a set of terminal nodes. The LMT approach deals with both numeric and nominal attributes, and missing values, as well as binary and multiclass target variables. Induction trees and logistic regression are combined in LMT. Cost-complexity pruning is used in LMT. The speed of this method is much slower than the others [21].

5) J48 classification algorithm: Quinlan's C4.5 algorithm is used to create J48. It builds the tree by selecting the best attributes at each node using the gain ratio. Each feature of the dataset divides it into small partitions to rank the attributes based on criteria that each partition is more consistent with respect to the class outcome. The attribute with the highest rank is used to split the data set. The algorithm will ideally terminate when all the instance in each partition belongs to one class. J48 creates a decision node based on the class's predicted estimations. The J48 decision tree can handle specific characteristics, data with lost or incomplete attribute estimations, and variable attribute prices. Pruning can improve accuracy in this situation [17], [21].

6) Hoeffding Tree (HT) classification algorithm: The Hoeffding Tree (HT) algorithm is one of the most basic algorithms for both stream data and static data classification. In the case of stream data, it can learn from huge data streams incrementally and at any time, given that the distribution of producing examples does not change over time. It generates decision trees in the same way as the standard batch learning approach does. Mathematically, Hoeffding trees and decision trees are connected. The HT technique is based on the basic principle that minimum sample size can frequently be sufficient for determining the best splitting attribute [22].

C. Evaluation of the Methodology

A trained classifier performs the function of assigning new data items in a given 255 collection to a target category or class by using two approaches.

1) Holdout method: as shown in Fig. 3, in training a classifier, the data set is separated into the training data set and test data set. In general, the training dataset contains 2/3 of data from the dataset, and the rest of the data is categorized as the testing dataset[23], [24].

2) 10–folds Cross-Validation (10-folds CV): k-folds Cross-Validation (CV) is a commonly used training control technique in learning a classifier. It applies the resampling technique to evaluate classification algorithms. The dataset is split into k number of groups. It randomly selects 1 data group as the testing dataset and the remaining groups (k-1) as the training dataset. Then build the data model by using the training dataset and evaluates it by using the testing dataset. Here each group has a chance to select as the testing dataset at a time and select as the training dataset at k-1 times. In general, 10-fold CV is used [23], [24].

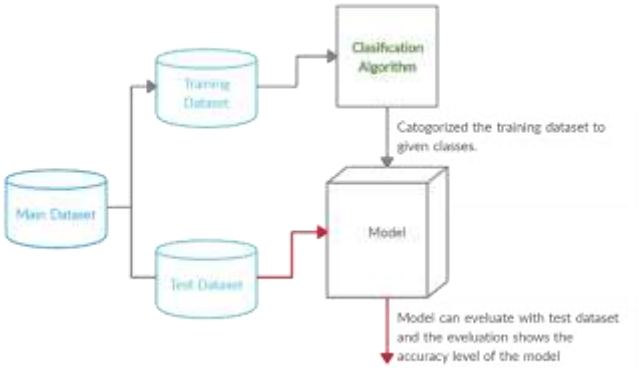


Fig. 3. General Classification Model.

There are several metrics to measure the accuracy of the classifier built.

1) Correctly classified instances and incorrectly classified Instances: This shows how many instances are correctly classified by the model and how many instances are incorrectly classified by the model. it gives numerical value as well as the percentage. Therefore, if it shows a higher value for the Incorrectly Classified Instances than Correctly Classified Instances then the built model is not good. And also, if it shows Correctly Classified Instances as 100% then it is also not a good outcome as the data over-fitting to the model may more validation error. The value between 80 to 100 for Correctly Classified Instances is usually accepted.

2) Confusion matrix: Confusion matrix also generates according to the Correctly Classified Instances and Incorrectly Classified Instances. The matrix size depends on the number of options in the output class. For example, if the output class has only two options then it generates a 2 by 2 matrix.

$option\_1$	$option\_2$	
a	b	$option\_1$
c	d	$option\_2$

In here,

a + d = value of Correctly Classified Instances

b + c = value of Incorrectly Classified Instances

3) Kappa statistic: This is also a good measurement to check the accuracy of the model. Simply it shows the accuracy of classifying into the correct class when considering any random data point. This value is generated by matching expected accuracy with observed accuracy. The following formula uses to calculate the kappa statistic.

$$\text{kappa statistic} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}} \quad (1)$$

This statistic is used for the model evaluation as follows:

- Kappa statistic < 0 means there is no agreement with accuracy.
- 0 < kappa statistic < 0.20 means that the accuracy is slight.
- 0.21< kappa statistic < 0.40 means that the accuracy is fair.
- 0.41< kappa statistic < 0.60 means that the accuracy is moderate.
- 0.61< kappa statistic < 0.80 means that the accuracy is substantial.
- 0.81< kappa statistic <1 means that the accuracy is perfect.

4) TP rate: This means True Positive rate and it is also a numerical value. It gives how many positive instances are correctly classified into the classes. The following formula uses to calculate the TP rate.

$$\text{TP rate} = \frac{\text{True positive instances}}{\text{Total number of positive instances}} \quad (2)$$

5) FP rate: Opposite of the TP rate. This means a False Positive rate and it is also a numerical value. It gives how many negative instances are incorrectly classified into the classes. The following formula uses to calculate the FP rate.

$$\text{FP rate} = \frac{\text{False positive instances}}{\text{Total number of negative instances}} \quad (3)$$

6) Precision: This talks about how many selected items are relevant to the given class. It shows a proportion of instances that are truly inside the class. The value can take by using the following formula.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (4)$$

7) Recall: This talks about how many relevant items are selected in the given class. It shows the proportion of instances that are classified inside the class. The value is derived by using the following formula.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False negative}} \quad (5)$$

8) F-measure: This is a value that is taken by considering precision and recall. It shows the connection between the low false positives and the low false negatives. Therefore, it is better to get a value near 1 for this and if you take a value near 0 then the model is quite bad. The following formula determines F-measure.

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

9) ROC Curve/AUC: The Receiver Operator Characteristic curves are commonly used to graphically show the relation between TP rate and FP rates for every possible cut-off test or a combination of tests. An important parameter associated with ROC curves is AUC that stands for Area Under Curve. The classifiers usually take an AUC value between 0.5 and 1.0 where 0.5(random guessing) is considered as worst and 1.0 is the best.

D. Generation of Training and Testing Dataset

Dividing the dataset into training and testing data can be done as shown in Fig. 4. The training dataset takes 2/3 of the total dataset and the testing dataset takes 1/3. All of these two datasets contain “SH” type people as well as people whose health state=“H” and health state = “UH”. Here the health state = “H” group contains both “SH” type people and the people who have become healthy without considering the total WBC’s standard hematology reference value. These notations were introduced and described under Data Preparation in Section III.

E. Experiment Results

Classification algorithms are applied to the dataset with training controls, holdout (using given training/testing dataset), and 10-fold cross-validation to find the best outcome. Table I shows the parameter measures needed to check the accuracy of the model for the training dataset.

The model built by the Random Forest algorithm using the training dataset and the model built by the Random Tree using the training dataset showed the highest number of correctly classified instances and the lowest number of incorrectly classified instances.

When considering the Kappa Statistic, the model built by the Random Forest algorithm's kappa statistic value is higher than the model built by the Random Tree algorithm's kappa statistic value. Therefore, the Random Forest Tree algorithm outperforms all the other algorithms.

The detailed experiment results of the RFT algorithm is given in Fig. 5. As depicted in Table I and Fig. 5, the model built by the Random Forest algorithm with the training dataset has the highest TP rate for both healthy and unhealthy classes. The weighted average of the TP rate and FP rate of this model are 0.915 and 0.279 respectively. The weighted average value for the precision and recall are 0.911 and 0.815, respectively.

F-measure value is an important value when considering the accuracy of the model. The F-measure value for healthy and unhealthy instances is 0.738 and 0.949, respectively. The weighted average is 0.911 which is close to 1.

As shown in Fig. 5, the Confusion Matrix of Random Forest summarizes that 48 healthy instances are classified as healthy, 24 healthy instances are classified as unhealthy, 318 unhealthy instances are classified as unhealthy and 24 unhealthy instances are classified as healthy.

The AUC curves drawn to the Random Forest models for both unhealthy and healthy instances are shown in Fig. 6 and Fig. 7, respectively. AUC values of it for both healthy instances and unhealthy instances give the same value of 0.971. Both curves have deviated much far from the diagonal line close to 1.0.

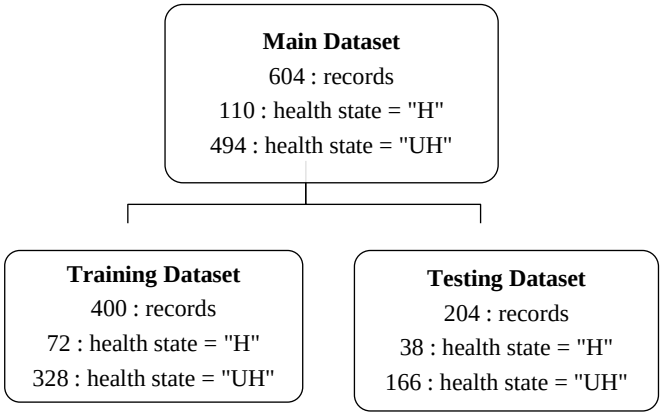


Fig. 4. Training Dataset and Testing Dataset.

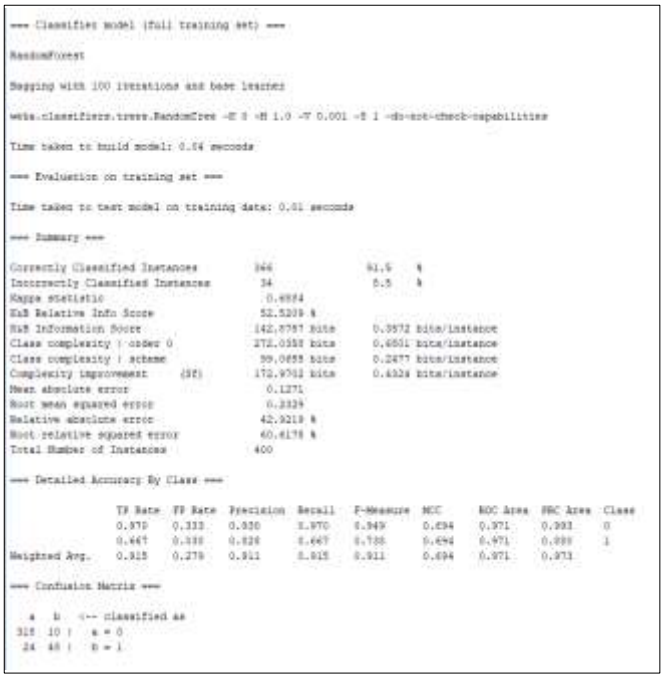


Fig. 5. The Experiment Results of the RFT Algorithm.

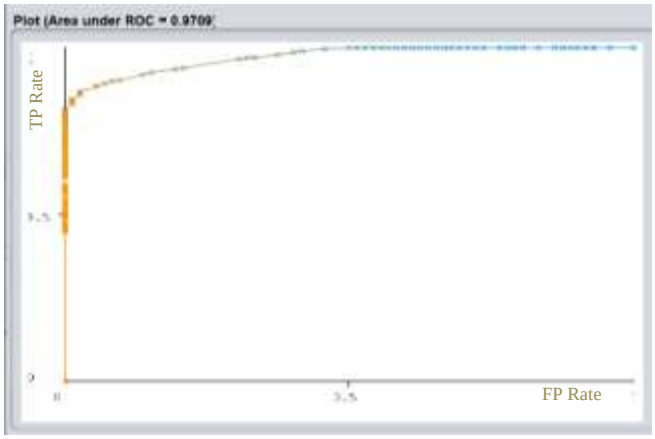


Fig. 6. ROC Curve for unhealthy Instances in the Model Built by using the Random Forest Algorithm.

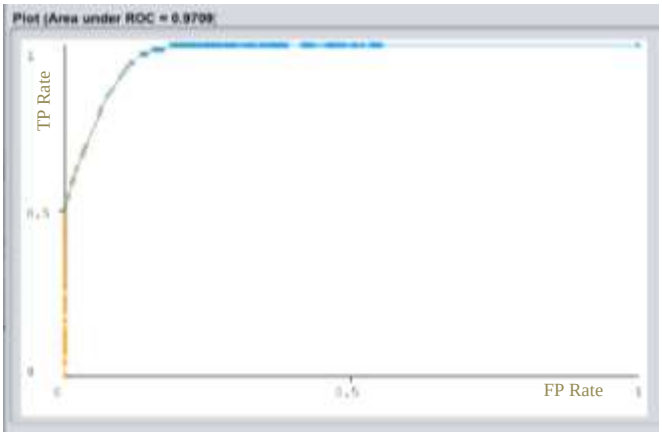


Fig. 7. ROC Curve for Healthy Instances in the Model Built by using the Random Forest Algorithm.

TABLE I. ACCURACY FEATURES

Algorithm	Correctly classified	Incorrectly classified	Kappa Statistic	Class attribute	TP Rate	FP Rate	Precision	Recall	F-measure	AUC
RFT - training dataset	366	34	0.6884	H	0.667	0.030	0.828	0.667	0.738	0.971
				UH	0.970	0.333	0.930	0.970	0.949	0.971
RFT – 10 folds CV	331	69	0.3563	H	0.403	0.079	0.527	0.403	0.457	0.854
				UH	0.921	0.597	0.875	0.921	0.897	0.854
RT - training dataset	366	34	0.6648	H	0.583	0.012	0.913	0.583	0.712	0.972
				UH	0.988	0.417	0.915	0.988	0.915	0.972
RT – 10 folds CV	326	74	0.2705	H	0.306	0.073	0.478	0.306	0.373	0.739
				UH	0.927	0.694	0.859	0.927	0.891	0.739
REP - training dataset	351	49	0.5139	H	0.472	0.034	0.756	0.472	0.581	0.917
				UH	0.966	0.528	0.893	0.966	0.928	0.917
REP - 10 folds CV	329	71	0.2160	H	0.208	0.043	0.517	0.208	0.297	0.844
				UH	0.957	0.792	0.846	0.957	0.898	0.844
LMT- training dataset	347	53	0.5437	H	0.611	0.076	0.638	0.611	0.624	0.925
				UH	0.924	0.389	0.915	0.924	0.920	0.925
LMT - 10 folds CV	332	68	0.3541	H	0.389	0.073	0.538	0.389	0.452	0.885
				UH	0.927	0.611	0.874	0.927	0.899	0.885
J48 - training dataset	357	43	0.5840	H	0.542	0.030	0.796	0.542	0.645	0.931
				UH	0.970	0.458	0.906	0.970	0.937	0.931
J48 - 10 folds CV	330	70	0.3099	H	0.333	0.067	0.522	0.333	0.407	0.851
				UH	0.933	0.667	0.864	0.933	0.897	0.851
HT - 10 folds CV	327	73	0.1456	H	0.139	0.034	0.476	0.139	0.215	0.670
				UH	0.966	0.861	0.836	0.966	0.897	0.670

IV. RESULTS AND EVALUATION

The model built by using the Random Forest algorithm was used to evaluate the test dataset and the results of it are shown in Fig. 8. As similar to the results for the training dataset described in Fig. 5, the results for the test dataset have also shown the excellent performance of RFT with the low variance.

As depicted in Fig. 4, there are 38 healthy instances and 166 unhealthy instances. The confusion matrix shown in Fig. 8 shows that the deviation of TP and TN values from the actual is very much low.

An estimation of the local reference range of WBC is determined based on the values of WBC of the healthy instances predicted by the RFT algorithm. The manual inspection of those healthy instances shows that WBC values lie within the range of $4100mm^3$ – $12800mm^3$. When compared with the standard reference range of WBC shown in Fig. 9, it can be observed that the local reference range of WBC is not the same as the standard reference range.

Some local WBC reference ranges which were established in other countries are as follows.

- $4420mm^3$ – $11100mm^3$ in Estern India [2].
- $2900mm^3$ – $9600mm^3$ in Sudan[1].
- $1900mm^3$ – $10100mm^3$ in Togo [9].
- $2800mm^3$- $7200mm^3$ in Malawi [5].

The above ranges have differed from the standard WBC reference range as well as our result. The nutrition levels in African countries may cause to return very lower boundary value for WBC. Sri Lanka is in a better state when compared with the African countries regarding nutrition levels. Hence the received range for WBC is reasonable.

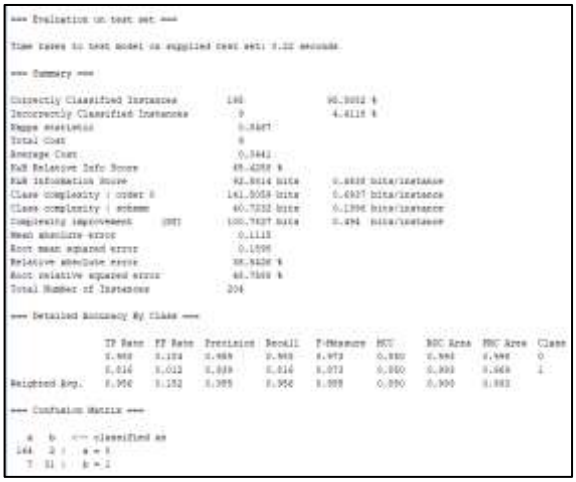
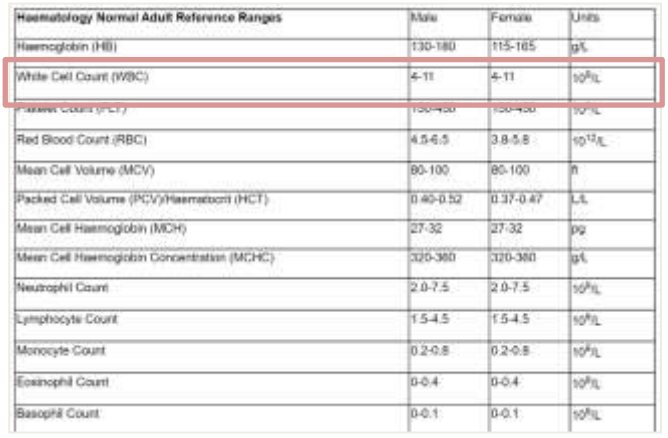


Fig. 8. ROC Outcome of the Test Dataset-RFT.



Haematology Normal Adult Reference Ranges	Male	Female	Units
Haemoglobin (Hb)	130-180	115-165	g/L
White Cell Count (WBC)	4-11	4-11	10^9/L
Platelet Count (PLT)	150-450	150-450	10^9/L
Red Blood Count (RBC)	4.5-6.5	3.8-5.8	10^{12}/L
Mean Cell Volume (MCV)	80-100	80-100	fl
Packed Cell Volume (PCV)/Haematocrit (HCT)	0.40-0.52	0.37-0.47	L/L
Mean Cell Haemoglobin (MCH)	27-32	27-32	pg
Mean Cell Haemoglobin Concentration (MCHC)	320-360	320-360	g/L
Neutrophil Count	2.0-7.5	2.0-7.5	10^9/L
Lymphocyte Count	1.5-4.5	1.5-4.5	10^9/L
Monocyte Count	0.2-0.8	0.2-0.8	10^9/L
Eosinophil Count	0-0.4	0-0.4	10^9/L
Basophil Count	0-0.1	0-0.1	10^9/L

Fig. 9. Standard Hematology Reference Ranges (https://www.royalwolverhampton.nhs.uk/services/service-directory-a-z/pathology-services/departments/haematology/haematology-normal-adult-reference-ranges/).

V. Conclusion

Medical Science is an immense area. Day by Day the area is updated. This research relates to Medical Science which spans across a vast area covering many disciplines and gets updated daily. The whole research was run based on certain disciplines such as Hematology, healthiness, unhealthiness, etc. The concept "healthy" is very complicated to define in medical science. It does not have a simple idea. It can be defined by using various subcategories. Hematology is such a subcategory. Medical officers may identify a person as healthy by looking at his blood reports. But it may not be correct as there will be a person with a good blood report but with disabilities on eye, ear, etc., inappropriate living styles like food habits, exercise habits, alcohol addictions, etc., diseases that cannot be detected from blood reports, or else everything is good but not healthy in mentally. Therefore, healthiness may not be determined always from the blood reports.

However, assuming that if a person is suffering from a disease, medical officers often check the blood reports to detect the disease, Hematology science is very helpful in such kinds of scenarios. This small area was addressed only through this research which assumes the healthiness depends only according to the blood reports.

The outcome of this research is to determine a local hematology reference range using data mining and machine learning techniques and the selected dataset was belongs to the adults in Sri Lanka. The research focused on finding a local reference range for total WBC value only. Hence a local reference range for the total WBC value was determined. The standard WBC (leukocyte) referential range is $4000mm^3$ – $11000mm^3$ and the local reference range which was empirically determined as the result of this research was $4100mm^3$ – $12800mm^3$. By repeating the same experiment for several different datasets and taking the average of the results, the accuracy of the local reference range can be further improved. And the proposed framework for WBC can be used to determine the local hematology reference ranges for other Hematology tests as well.

When considering the similar works discussed in Section II, most of the researchers who tried to establish local hematology reference ranges for their countries have used statistical approaches. The drawbacks of statistical approaches have been discussed in Section II. The proposed approach using data mining and machine learning concepts is the best solution to address all mentioned drowbacks. It does not depend on data distributions or the number of data records. Also, data mining and machine learning concepts perform well with the fair representation of the sample dataset.

There are several limitations in this research. The data used for the research belongs to people over 21 years old. The dataset has been taken from hospitals, laboratories in a particular area and hence does not cover the whole area of Sri Lanka. Further, the accuracy of the dataset plays a vital role in this nature of the research. In this regard, a carefully set up survey can be conducted to obtain blood samples from selected donors. Despite the set setup is not straightforward and involves high cost, the following suggestions are recommended in selecting the donors.

- Should cover different geographical areas, climate zones, cultural backgrounds, etc. in Sri Lanka.
- Use the medical history of the donors to categorize them as healthy or not.
- Incorporate the food habits, living styles, disabilities, BMI index, smoking habits, alcohol consumption, sexual diseases, etc. too to determine the healthiness of the donors.

References

[1] Taha et al., "Reference Ranges of White Blood Cells Count among Sudanese Healthy Adults," pp. 554–559, Oct. 2018, doi: 10.21276/sjm.2018.3.10.2.

[2] D. D. Dey, "CLSI-Derived Hematology Reference Intervals for Healthy Males in Eastern India," vol. 2, no. 2, 2013, Accessed: Jul. 16, 2021. [Online]. Available: https://www.academia.edu/8462096/CLSI_Derived_Hematology_Reference_Intervals_for_Healthy_Males_in_Eastern_India

[3] N. Tekkesin, H. Bekoz, and F. Tükenmez, "The largest reference range study for hematological parameters from Turkey: A case control study," J. Clin. Exp. Investig., vol. 5, pp. 548–552, Dec. 2014, doi: 10.5799/ahinjs.01.2014.04.0455.

[4] T. Huma and U. Waheed, "The Need to Establish Reference Ranges." Journal of Public Health and Biological Sciences, Jun. 08, 2013.

[5] W. L. Mandala, E. N. Gondwe, J. M. MacLennan, M. E. Molyneux, and C. A. MacLennan, "Age- and sex-related changes in hematological parameters in healthy Malawians," J. Blood Med., vol. 8, pp. 123–130, Aug. 2017, doi: 10.2147/JBM.S142189.

[6] L. G. Bimerew et al., "Reference intervals for hematology test parameters from apparently healthy individuals in southwest Ethiopia," SAGE Open Med., vol. 6, pp. 1–10, Oct. 2018, doi: 10.1177/2050312118807626.

[7] O. El Graoui et al., "Hematology reference intervals in Moroccan population," Clin. Lab., vol. 60, no. 3, pp. 407–411, 2014, doi: 10.7754/clin.lab.2013.130117.

[8] T. Miri-Dashe et al., "Comprehensive Reference Ranges for Hematology and Clinical Chemistry Laboratory Parameters Derived from Normal Nigerian Adults," PloS One, vol. 9, p. e93919, May 2014, doi: 10.1371/journal.pone.0093919.

[9] I. M. Kueviakoe, A. Y. Segbena, H. Jouault, A. Vovor, and M. Imbert, "Hematological Reference Values for Healthy Adults in Togo," ISRN Hematol., vol. 2011, p. e736062, Nov. 2010, doi: 10.5402/2011/736062.

[10] J. O. Mugisha, J. Seeley, and H. Kuper, "Population based haematology reference ranges for old people in rural South-West Uganda," BMC Res. Notes, vol. 9, no. 1, p. 433, Dec. 2016, doi: 10.1186/s13104-016-2217-x.

[11] N. Omarine Nlinwe, Y. Larissa Kumenyuy, and C. Precious Funwi, "Establishment of Hematological Reference Values among Healthy Adults in Bamenda, North West Region of Cameroon," Anemia, vol. 2021, pp. 1–7, Feb. 2021, doi: 10.1155/2021/6690926.

[12] G. Gunčar et al., "An application of machine learning to haematological diagnosis," Sci. Rep., vol. 8, Jan. 2018, doi: 10.1038/s41598-017-18564-8.

[13] S. Sivapalaratnam, "Artificial intelligence and machine learning in haematology," Br. J. Haematol., vol. 185, no. 2, pp. 207–208, 2019, doi: 10.1111/bjh.15774.

[14] F. K. Alsheref and W. Hassan, "Blood Diseases Detection using Classical Machine Learning Algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, 2019, doi: 10.14569/IJACSA.2019.0100712.

[15] M. N. Anyanwu and S. G. Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms."

[16] B. R. J. Vadhanam, S. Mohan, and V. V. R. and V. Sugumaran, "Performance Comparison of Various Decision Tree Algorithms for Classification of Advertisement and Non Advertisement Videos," Indian J. Sci. Technol., vol. 9, no. 48, pp. 1–10, May 2016, doi: 10.17485/ijst/2016/v9i48/102098.

[17] Periyar University, N. S. anaN, and V. G. thri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)," Int. J. Comput. Trends Technol., vol. 59, no. 2, pp. 73–80, May 2018, doi: 10.14445/22312803/IJCTT-V59P112.

[18] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," Int. J. Comput. Sci. IssuesIJCSI, vol. 9, Sep. 2012.

[19] S. Sivapalaratnam, "Artificial intelligence and machine learning in haematology," Br. J. Haematol., vol. 185, no. 2, pp. 207–208, 2019, doi: 10.1111/bjh.15774.

[20] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," vol. 2, no. 2, p. 9.

[21] M. Maulana and M. Defriani, "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period," PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log., vol. 8, pp. 39–48, Mar. 2020, doi: 10.33558/piksel.v8i1.2018.

[22] P. K. Srimani and M. M. Patil, "Performance analysis of Hoeffding trees in data streams by using massive online analysis framework," Int. J. Data Min. Model. Manag., vol. 7, no. 4, p. 293, 2015, doi: 10.1504/IJDMMM.2015.073865.

[23] M. Halkidi and M. Vazirgiannis, "Quality Assessment Approaches in Data Mining," 2010, pp. 613–639. doi: 10.1007/978-0-387-09823-4_31.

[24] F. Maleki, N. Muthukrishnan, K. Ovens, C. Md, and R. Forghani, "Machine Learning Algorithm Validation," Neuroimaging Clin. N. Am., vol. 30, pp. 433–445, Nov. 2020, doi: 10.1016/j.nic.2020.08.004.