IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2021, accepted March 9, 2021, date of publication March 17, 2021, date of current version March 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066782

Classification of β-Thalassemia Carriers From Red Blood Cell Indices Using Ensemble Classifier

SAIMA SADIQ[1], MUHAMMAD USMAN KHALID[1], MUI-ZZUD-DIN[1], SALEEM ULLAH[1], WAQAR ASLAM[2], ARIF MEHMOOD[2], GYU SANG CHOI[3], AND BYUNG-WON ON[4]

[1]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38542, South Korea
[4]Department of Software Convergence Engineering, Kunsan National University, Gunsan 54150, South Korea

Corresponding authors: Arif Mehmood (arifnhmp@gmail.com) and Gyu Sang Choi (castchoi@ynu.ac.kr)

ABSTRACT Thalassemia is viewed as a prevalent inherited blood disease that has gotten exorbitant consideration in the field of medical research around the world. Inherited diseases have a high risk that children will get these diseases from their parents. If both the parents are β-Thalassemia carriers then there are 25% chances that each child will have β-Thalassemia intermediate or β-Thalassemia major, which in most of its cases leads to death. Prenatal screening after counseling of couples is an effective way to control β-Thalassemia. Generally, identification of the Thalassemia carriers is performed by some quantifiable blood traits determined effectively by high-performance-liquid-chromatography (HPLC) test, which is costly, time-consuming, and requires specialized equipment. However, cost-effective and rapid screening techniques need to be devised for this problem. This study aims to detect β-Thalassemia carriers by evaluating red blood cell indices from the complete-blood-count test. The present study included Punjab Thalassemia Prevention Project Lab Reports dataset. The proposed SGR-VC is an ensemble of three machine learning algorithms: Support Vector Machine, Gradient Boosting Machine, and Random Forest. Comparative analysis proved that the proposed ensemble model using all indices of red blood cells is very effective in β-Thalassemia carrier screening with 93% accuracy.

INDEX TERMS Thalassemia, prenatal screening, complete blood count, support vector machine, gradient boosting machine, SGR-VC.

I. INTRODUCTION

Thalassemia is mainly a combination of two Greek words, ''Thalassa'' meaning sea and ''Hema'' means blood [1]. Thalassemia is an inherited blood disorder that is commonly found in different parts of the world especially in South Asia. Inherited disease means that it is passed from parents to their children [2]. In Thalassemia, haemoglobin level decreases from normal limit which causes reduction in the count of productive red blood cells, which may lead to severe anemia. Red blood cells (RBCs) mainly consist of a protein containing a great deal of iron named as haemoglobin which form main concentration of RBCs.

Haemoglobin normally consists of four protein chains, 2-alpha globin and 2- beta globin. Any change in a single gene from all these four chains may lead to anemia. Genetic and acquired malfunctions may affect the proper structure of haemoglobin resulting in many disorders. Acquired malfunctions are not as severe as genetic which results in mutation in genes and their proper functioning [3]. Both qualitative and quantitative abnormalities in its structure can lead to severe disorders leading to gene mutations. These disorders are globally threatening. Production rate of haemoglobin is affected by the quantitative malfunctions while variations in the qualitative aspects have influence the rate with which protein chains are produced in tetramer of haemoglobin resulting in Thalassemia. Initially it was thought that Thalassemia is a clinical disease [4]. In the late twentieth century first case of Thalassemia was reported and not surprisingly in the Mediterranean region. Major symptoms owned by such patients were huge spleen, defects in bones, and severe anemia. After this many such cases were reported especially in the

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

Mediterranean region. β-Thalassemia is the abnormality of genes that controls haemoglobin, and is most common Thalassemia disease [5].

Severity of anemia depends upon the number of missing genes. Moderate to severe anemia may occur if more than two genes are missing [6]. It is estimated that 9.8 million people are β-Thalassemia carriers in Pakistan which is the 5-7% of total population. Therefore one of the major tasks for medical experts is to identify the Thalassemia carriers among the normal persons. Generally identification of the Thalassemia carriers is established by some quantifiable blood traits changes to the count, shape and size of red blood cells. Cousin marriage culture in countries like Pakistan increases the chances of Thalassemia patients. As if both the parents are Thalassemia carriers there is 25% of chance that their every child will be Thalassemia patient. As Thalassemia is an inherited blood disorder so if the family is affected by this disorder then there is a huge possibility that a β-Thalassemia carrier will get married with another β-Thalassemia carrier which causes a high risk of Thalassemia patient. In most cases, β-Thalassemia carriers are unaware of this disorder. If an early and correct diagnosis is made, a lot of lives can be saved from this life-threatening disorder [7]. β-Thalassemia carriers are asymptomatic and completely healthy, a Complete Blood Count (CBC) test which is fast and inexpensive, helps in diagnosis of Thalassemia carriers. Other tests used in diagnosis of Thalassemia carriers are expensive and time consuming. Serum iron, HbA2, ferritin and iron binding capacity are normally used for diagnosis, but these tests are not performed commonly at each center.

Several researches have been performed in the literature to diagnose β-Thalassemia from other disorders using CBC. Some works used indices [8], [9] while other applied machine learning models to differentiate β-Thalassemia from Iron Deficiency Anemia (IDA) [10], [11]. In [12] authors detected β-Thalassemia carriers using Support Vector Machine (SVM). Aszhari *et al.* [13] classified Thalassemia disease using Random Forest (RF). A hybrid model was used for Thalassemia carrier detection in [10]. They deployed Naive Bayes (NB), K-nearest neighbour (K-NN), Decision Tree (DT) and a neural network model multilayer perceptron (MLP). In [14] authors achieved 91% accuracy by using MLP neural network model, with Principal Component Analysis (PCA), to classify β-Thalassemia from IDA. Pattern based input selection with Artificial Neural Network (ANN) was used for disease classification in [15].

Screening of Thalassemia carriers is of major importance to avoid and control this disease. Process for accurate detection of carriers is very costly, prolonged and requires expertise and specialized equipment which adds to its severity especially for those people living in backward areas. In this study, a dataset containing nine red blood cells indices_RBC, HB, HCT, MCV, MCH, MCHC, RDW, PLT and WBC_ of 5066 patients is used to train machine learning models. At first, real values of blood tests are normalized into six divisions (0-5). Experiments have been performed on Tree-based machine learning models (Random Forest (RF) and Gradient Boosting machine (GBM)) and probability-based machine learning models (Support Vector Machine (SVM)). This research aims to propose a Voting Classifier (VC) based on the ensemble of SVM, GBM and RF that we called SGR-VC. Our proposed SGR-VC can automatically diagnose Thalassemia carriers by using CBC test which is a cost effective and fast solution. Major contributions of this study are summarized below:

- SGR-VC, that is an ensemble of SVM. GBM and RF is designed to isolate and analyze β-Thalassemia carriers from β-Thalassemia non-carriers.
- Attributes of dataset are normalized to improve effectiveness of classifiers.
- Performance of SGR-VC is compared with RF, SVM and GBM individually for β-Thalassemia carrier identification.
- The proposed voting classifier outperforms other models and provides a rapid and cost-effective solution for β-Thalassemia carriers screening.

The rest of the paper is organized as: Section II presents related work, section III discusses material and methods, section IV elaborates results and discussion. Finally section V concludes the work.

II. RELATED WORK

Most effective way to diagnose β-Thalassemia carrier is to use medical knowledge based on patient data. Manual diagnosis is difficult and can cause delay in proper control. However there is a need for an automatic prediction system for Thalassemia carrier detection to control its transfer to the next generation. Its treatment is very expensive and impossible to be afforded by most of the citizens of developing countries. Performance of machine learning models differs when applied to different domains [16]. Many researchers applied different data mining methods such as RF, NB, DT, KNN, SVM and ANN for Thalassemia diagnosis.

Genotype of β-Thalassemia patients was classified using the PCA method in [17]. Authors analysed all the basic components of the blood. They used their model for the screening of the patients and then they compared the performance of the PCA model with other classifiers like multinomial Logistic Regression (LR), Bayesian Network (BN), MLP, NB and k-NN etc. Dataset used consists of 127 β -Thalassemia patients records. Common KADs suite was used for explaining the variables and for selection of the best attributes. After selection of features, they transferred and minimized them by using their proposed PCA model. Then the results were run on the above stated algorithms. It was observed that the more efficient algorithm was MLP with accuracy of about 87% while the accuracy of MLP, BN, NB and k-NN was found to be 83%, 85.0394%, 83.0394% and 86% respectively.

Upadhyay [18] classified the β-Thalassemia patients and screened the major and minor patients by his proposed

algorithm model. Author used an artificial neural method for the screening of the Thalassemia patients. The model was basically used by employing a feed-forward and backward-propagation channel algorithm. The input dataset consisted of clinical records of about 100 patients and data was run on the neural network to check the performance of the model. Results obtained during the performance sounded good with about 77% accuracy.

Sandanayake *et al.* [19] designed an automatic system for complete diagnosis of the Thalassemia patients. Thalassemia, as a genetic disease, was analyzed manually before their work. The manual process was time consuming also it required almost 90 days, that is why they designed an automated mechanism to facilitate the patients. Classification features and image processing were mainly used for the designing of the automated model. The analysis of red blood cells of the patients was the main analysis standard. Three features from the red blood cells like erythrocytes count, their shape and their color were studied to check whether they are in normal limit or not. Any small change in their diameter, physical appearance and color may lead to serious disorders. Authors mainly designed a system to diagnose Thalassemia patients from the shape, color and size of their erythrocytes. They also designed a web system for patient and doctor interaction.

A new model was designed for the screening of Thalassemia minor and IDA patients by [20]. For this purpose they used artificial neural networks with some new patterns. Authors built a MLP model with one hidden layer. A hundred neurons and four inputs were used in the designed model. They collected the clinical records of about 400 persons by using the simple lab test like the CBC tests. The purpose of the model was to classify the patients in three groups i.e. Healthy, Thalassemia minor and IDA.

Shurrab and Maghari [21] used three different classifiers of the data mining technique for the extraction of knowledge about the early diagnosis of the disease. Before this, much information was collected about blood diseases like Thalassemia using the medical records in different areas of the world especially in Gaza strips. Early diagnosis of disease can be beneficial for its cure. If a disease is diagnosed at an early stage then chances of cure are increased to a greater extent. Data was collected by a large number of medical tests which were analyzed properly. The Tree classifiers, DT, NB and Rule Induction (RI) were mainly used for this purpose. The blood tests were collected from Gaza hospital, an area in Gaza strips. Probability of blood diseases in the early stage which can increase the curing factor was analyzed through CBC test results of the collected dataset. The dataset of patients with three main diseases was analyzed. The diseases were adult hematology, childhood hematology and tumor patients. It was observed that NB can predict tumor disease with 56% accuracy. From all three classifiers the accuracy of DT was least in detecting or analyzing the above mentioned three diseases. While RI could analyse adult and childhood hematology with a 58% to 66% accuracy respectively.

Medical images like ultrasound images [22], microscopic images [23]–[25] have been extensively examined by researchers for helping in cure of diseases. El-Halees and Shurrab [26] employed data mining techniques to correlate the link between the blood tests and diseases caused by tumors. For early diagnosis of disease this information was necessary so that it can be cured easily. Before this work a huge amount of the data was collected from the medical labs to screen different diseases. They used three different mining techniques to perform an experiment on the data collected from blood test reports and records. The techniques used were Deep Learning (DL), RI and Association Rules (AR). They performed this experiment to analyze and screen normal blood disease patients from tumor patients. Dataset was collected from Gaza European hospital located in Palestine. They observed that AR is beneficial to analyze the association between blood tumor and the blood diseases. DL based classifier outperformed in this case with 78% accuracy. Blood tumors and normal blood disorders were explained with the help of RI method.

A hybrid model of the data mining for the screening of β-Thalassemia carriers from asymptomatic Thalassemia is proposed by [10]. Main aim of the study was to construct a novel model which gave the best screening results based on the simple CBC reports instead of expensive and time consuming tests. They collected the clinical records of Thalassemia patients from the Palestine Avenir foundation. The experimentation of the model was mainly in two steps. In the first step, random class distribution of the dataset was regularized by a method called SMOTE to balance the imbalance effects in the dataset. In the second step different Machine Learning (ML)classifiers were used. They mainly used K-NN, DT, NB and MLP for screening of the normal and the β-Thalassemia carriers. Efficiency of the model was evaluated by different matrices. It was observed that SMOTE was very helpful for screening of β-Thalassemia ratio from an imbalanced data. Results revealed that NB outperformed other classifiers for the screening of β-Thalassemia carriers and normal persons with 400% oversampling SMOTE ratio.

Egejuru *et al.* [27] designed a predictive model for the analysis of the risk factor of Thalassemia in every age group of the people. Supervised ML algorithms were used to diagnose this risk. Data was collected by arranging talks with the related physicians and public opinion by a questionnaire consisting of several key questions. In this way empirical data was collected. Analysis of the information got through interviews and questionnaires was done through the Waikato environment. The results were compared with that of actual data obtained from the hospital. This data, consisting of 51 patients, explained the causes of Thalassemia. Gender, age, married life, social status, morality were the population related factors which were studied. Some clinical parameters like the urine color, spleen size, diabetic level and family background were studied. This detailed study resulted in 31% high case risk factors while 16% moderate cases, 11% low risk cases and 43% were those with no risk case. The overall

study showed that if the medical centers and the related services centers use MLP for the diagnosis of Thalassemia, results can be improved.

Ismaeel [28] used a bio mining technique using multiple neural networks for the screening of Thalassemia which cause any change in the gene or the protein part of the cell. Before this, many studies were done but techniques were not reliable to predict either it is effective or in the optimal form. This comprehensive study compared the results obtained from the methods used for 64% mutations caused by Thalassemia. They gnomically classified the β-Thalassemia mutations in the INHALANT gene. Results showed that optimal analysis was done by studying BP with about 1000 iteration, which is much better than other methods used before.

Al-Hagger *et al.* [29] analysed the contribution of different factors used for the diagnosis of Thalassemia using multiple criteria decision making techniques. Risk factor of iron transfusion was estimated using the severity index. They used the bioinformatics technology for the construction of these computational methods. They collected the data and stored it for the production and invention of new and novel tools in the medicine. Bioinformatics are less dangerous and were used for the application on the human level. They concluded that by using the personalized prescriptions and laboratory data set, much information can be grabbed about individuals. By using such details, proper medicine can be prescribed to the right person. Authors concluded that the BMI (Biomedical Information) is the best choice for the improvement of the individuals health especially β-Thalassemia patients.

Çil *et al.* [30] proposed a model for the differentiation of β-Thalassemia patients from the IDA patients because the symptoms of both the patients are almost similar. The screening of both diseases is also important because if both are misdiagnosed or confused with one another it would lead to serious complications. If a diagnosis of β-Thalassemia is confused with IDA it will lead to serious complications in the married life and may lead to the outspread of this disease in their offspring. If a Thalassemia patient is declared as IDA person by physician they will ingest extra iron to the person. They proposed a decision support mechanism by using different classifiers like SVM, Regularized Extreme Learning Machine classification algorithms, K-NN, LR and Extreme Learning Machine (ELM). Their dataset consisted of clinical records of 345 persons. Different parameters like Accuracy, Specificity, F1 score and Sensitivity were observed by analyzing the hemoglobin and red blood cells count and CBC along with its all parameters.

The analysis of literature reveals that different approaches have been applied to explore Thalassemia in many aspects but on small size datasets. Several models have been designed for analysis and classification of β-Thalassemia using CBC test. Our study is an effort of this series to provide a cost effective solution with improved results in less amount of time to detect β-Thalassemia carrier using fast and cheap CBC test.

III. MATERIAL & METHODS

This section describes a dataset, preprocessing steps and machine learning framework for β-Thalassemia carrier prediction. Fig. 3 presents the proposed framework used in this research for Thalassemia carrier prediction.

A. DATASET DESCRIPTION

In this research a novel dataset is collected from the database of Punjab Thalassaemia Prevention Programme (PTPP). PTPP is a step taken by Punjab Government Pakistan towards a country free of Thalassemia. PTPP provides healthcare to Thalassemia major patients but its main aim is the screening of β-Thalassemia Carriers. For that purpose, extended family screening is performed. If one member of a family is diagnosed with β-Thalassemia major, the screening is performed on the whole family free of cost. By this method of screening, the maximum number of carriers are identified. PTPP screens more than 300,000 patients every year. These premarital tests and whole family screenings resulted in a decrease in the number of patients in their working areas. In this study, the record of 5066 patients tested in 2019 is taken. In this database out of 5066 records 3051 patients are β-Thalassemia Non-Carriers and 2015 records are β-Thalassemia Carriers as shown in Fig. 1. Gender-wise distribution is 53% and 47% for males and females respectively and the age-wise dataset is 54% and 46% for adults and children respectively.

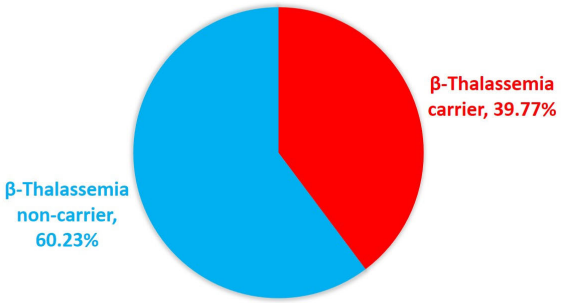


FIGURE 1. Class-wise distribution of dataset.

1) ATTRIBUTES

Attributes used in this work are personal demographic information and patients' CBC parameters. Demographic information includes Age and Sex. CBC Parameters include RBC, HB, HCT, MCV, MCH, MCHC, RDW, PLT, and WBC. Details of these parameters are presented in Table 1.

B. PREPROCESSING

The data collected from PTPP is in the form of reports which are compiled in a pdf data file. After extracting the data from the file it should be arranged in a format so that further processing can be performed. Extracted and arranged data still can have noise, incompleteness and inconsistency which can affect the outcome of the research.

TABLE 1. Attributes of CBC used in dataset.

Attribute	Terms	Description	Normal range of value
RBC	Red Blood cell count	Number of RBCs in haemoglobin is the key to differentiate between anemic and normal persons.	Between 4 to 5 ×1012cells per liter
HB	Hemoglobin	Related to the concentration of the hemoglobin molecules in the blood.	The normal range varies according to age and sex. For adults normal values between 13 to17 grams per deciliter (g/dL) for males and 12 to 15 g/dL for females.
HCT	Hematocrit	Related to the percentage of blood by volume that is composed of red blood cells.	Ranges between 45%–50% for males and 30%–45% for females.
MCV	Mean cell volume	Related to the average size or volume of a typical red blood cell in a blood sample.	Ranges between 80 to 100 femtoliters (femtolitre=10–15 L).
MCH	Mean cell hemoglobin	Related to the average amount of hemoglobin per red blood cell in a blood sample.	The normal value ranges from 27–34 picograms per cell (pictogram=10–12 g)
MCHC	Mean cell hemoglobin concentration	Related to the average hemoglobin concentration in a volume of blood.	Between 32% to 36%.
RDW	RBC distribution width	Related to the variability in the red blood cells size and shape.	Between 11% to 15%.
PLT	Platelets count	Related to the number of platelets in a volume of blood.	Between 150–350×109/L
WBC	White blood cell count	Related to the number of the white blood cells in the blood.	Between 4–10×109/L

TABLE 2. Dataset sample before applying normalization.

Age	Sex	WBC	RBC	HGB	HCT	MCH	MCV	MCHC	RDW	PLT	FINAL FINDING
35	Male	7.1	6.71	13.5	45.6	20.1	68	29.6	16.2	248	Beta Thalassemia Carrier
26	Female	9	4.85	11.5	38	23.7	78.4	30.3	18.3	358	Beta Thalassemia non Carrier
25	Male	7.3	6.65	12.4	40.5	18.6	60.5	30.6	16.4	247	Beta Thalassemia Carrier
25	Female	15.1	4.31	8.9	32.8	20.6	76.1	27.1	16	361	Beta Thalassemia non Carrier
3	Female	12.8	6.23	10.5	40.1	16.9	64.4	26.2	20.1	344	Beta Thalassemia Carrier
4	Male	12.3	6.05	11.3	43.2	18.7	71.4	26.2	15.4	353	Beta Thalassemia Carrier

Preprocessing that is applied to data after extraction and arranging it in this study consists of two steps; data cleaning and normalization.

1) DATA CLEANING

Cleaning of data is a stage in which some operations on the dataset are performed so that impurities like missing values, incorrect inputs, and unrealistic entries can be removed or corrected. This is important as there are classifiers that cannot deal with missing values. Incorrect inputs other than related data can halt the classification process. Unrealistic entries can lead us to unrealistic and unexpected results. The dataset used in this research has passed through the following data cleaning steps.

1) Elimination of incomplete input values: In this step, missing values of CBC tests in the dataset have been searched in the medical record of the hospital and filled in the dataset with the corrected value. If the value is not found in the medical report, the whole record has been removed from the dataset.
2) Elimination of duplicate Data: In this step, we removed all the duplicated data from the dataset which has been performed with the help of patient_id. Patient_id is a unique number that is allocated to all patients.
3) Insignificant attributes removal: In this step, all the insignificant attributes have been removed from the dataset which has no effect on the classification result; like patient_id which is just used to identify patients. Other removed attributes include Test date, Patient name, and Family name.

2) NORMALIZATION

It is very important to apply normalization before applying any classifier to the dataset as attributes in the dataset have a wide range of values that can reduce the classification accuracy. Blood test values are real numbers and scaling real numbers at the same intervals can increase the effectiveness of the classifier. Table 2 shows the dataset before the normalization process and Table 3 shows the dataset after normalization.

In this work, normalization has been performed by taking into account the normal value of every test. Every test has its own normal range and these normal values vary from adults to children and from females to males. In this case, the normalization has been done manually and has different parameters for different attributes depending upon their data type and their normal values. Normal values of every test are different for adults and children so by keeping that in mind, age has been normalized in two values [0, 1] 0 value represents children and 1 value represents adults. In the same way, gender attribute has been normalized in [0, 1] where 0 is for female and 1 is for male. In terms of tests, every test has

TABLE 3. Dataset sample after applying normalization.

Age	Sex	WBC	RBC	HGB	HCT	MCH	MCV	MCHC	RDW	PLT	FINAL FINDING
1	1	2	5	2	2	0	0	0	5	2	1
1	0	3	2	1	1	0	0	0	5	3	0
1	1	2	5	1	1	0	0	0	5	2	1
1	0	5	1	0	0	0	0	0	5	3	0
0	0	5	5	0	1	0	0	0	5	3	1
0	1	5	4	1	2	0	0	0	5	3	1

been normalized in 6 divisions [0, 1, 2, 3, 4, 5] where 0 is below normal range 5 represents higher than normal range and 1, 2, 3, and 4 are four equal divisions within the normal range. The targeted class is also represented by [0, 1] where 0 is for β-Thalassemia non-Carrier and 1 is for β-Thalassemia Carrier. This kind of normalization helps to observe every change in test values rather than binary representation which was used in some previous studies.

C. CLASSIFICATION ALGORITHMS

SVM is a supervised learning model that is used to solve classification and regression problems [31]. It classifies data into various categories of different disciplines. SVM has been used extensively in binary classification tasks for both linear and non-linear classification. SVM draws hyper-planes to divide data into groups and the best hyper-plane divides data with large separation into different classes. SVM is widely being used by many researchers in different applications [32]. Very little parameter adjustment is required in SVM is its advantage over other classifiers. The use of the Gaussian function in each training instance in the classification increases training time and reduces performance on a large dataset is its limitation.

GBM is based on boosting which sequentially creates a base model. Multiple models are developed in a sequence based on training case errors to improve accuracy. Previous base learner errors are considered by the next model during the training phase. Each base model corrects the errors of the previous model. Weak learners are combined into a single strong and more accurate base learner in boosting. GBM adds a base model to minimize loss function in each iteration strategically. It uses strategy to put more focus on the data samples that were previously difficult to estimate. The performance of the model depends upon the number of trees, the number of iterations, and the learning rate. By selecting the optimal combination of these three parameters, the best performance of the model can be achieved. It also performs well on unprocessed data. Among various loss functions, gradient descent is most common.

$$y_i^p = y_i^p - \alpha * \Sigma(y_i - y_i^p) \tag{1}$$

where α represents the learning rate and $\Sigma(y_i - y_i^p)$ represents the sum of residuals.

RF is an ensemble classifier that is used for classification as well as regression. It combines two ML-based techniques bagging and random selection. RF creates decision trees during training and predicts final output with majority voting [33]. Data instance is assigned to the class with more votes which also reduces the chance of overfitting. RF has shown robust results in classification tasks with high accuracy and can deal with noise and outliers in the data. It reduces variance by using a bootstrap dataset and selection of a random subset of data features [34]. Working of RF is presented under.

$$p = mode\{T_1(y), T_2(y), \ldots, T_m(y)\} \tag{2}$$

$$p = mode\left\{\sum_{m=1}^{m} T_m(y)\right\} \tag{3}$$

where p represents the final prediction that is computed by majority votes of trees T_1, T_2 and T_m.

Voting classier is an ML-based classifier that suggests the outcome based on the maximum chances of the selected group as the output class. It is the amalgam of many models on which it works. It works in a simple way. It compiles the results of different classifiers that are passed through these machine classifiers and selects the outcome of every classifier on the basis of their maximum prediction. A voting classifier combines predictions of multiple individual classifiers and improves the performance as compared to individual classifiers [35]. Voting classifier exhibits features of multiple models and in this study bagging(RF) and boosting (GBM) combined with probability-based (SVM) model as a result a significant improvement in classification accuracy has been attained. The training and final prediction criteria of all three classifiers are different. RF makes independent trees while GBM always tries to reduce the error of the previously made decision tree. SVM estimator is used to making final predictions on the basis of prediction support values of each classifier. That's why the proposed model makes better predictions than simple baseline models. It has been explored in many pieces of research that combining prediction of individual classifiers could be more effective in taking decisions as compared to individual ones [36], [37]. However, voting classifiers require a good amount of data to get trained on is a big challenge. The theme behind this is to create a single authentic model that collects the collective findings of each classifier instead of constructing separate classifiers and then compiling the findings of each classifier.

D. PERFORMANCE EVALUATION

The field of data science is growing and helping in solving many problems related to disease diagnosis, business-related decisions, image classification, and many other problems

related to different areas. One of the key aspects of data mining is performance evaluation. It is a hard task to determine the performance of a model in classification. In classification, predictions are made and the reliability of any model can only be analyzed by selecting evaluation measures that exhibit the true performance of a model [38]. Appropriate evaluation parameters are equally important as appropriate machine learning models. The selection of relevant and accurate evaluation parameters helps the research community to understand the strengths and weaknesses of the proposed methodology. This study compared the performance of machine learning models in terms of Accuracy, Precision, Recall, and F1-score.

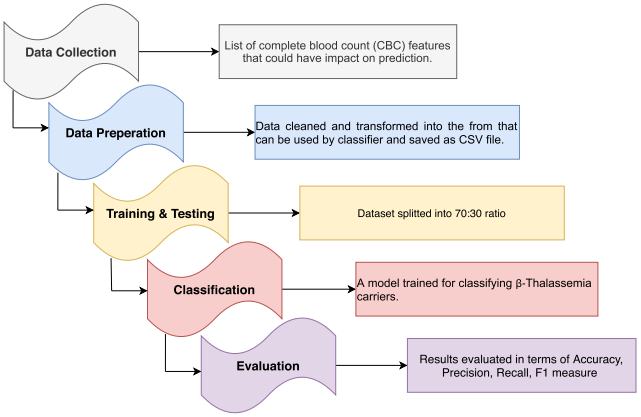


FIGURE 2. Phases of methodology steps.

E. PROPOSED FRAMEWORK

The main goal of this research work is to develop a system that provides a simple and reliable process to identify β-Thalassemia carriers among normal people. Fig. 2 presents the phases of the proposed methodology while Fig. 3 provides a comprehensive view of our model. It starts with a patient's arrival for a test sample at the laboratory. After the sample is collected the CBC tests are performed on these samples and results are stored in the database in the form of reports. Then these reports are fetched from the database. In this work, more than 5066 reports have been taken. Preprocessing involves data extraction from lab reports. It includes data cleaning which consists of elimination of incomplete input values, elimination of duplicate data, insignificant attributes removal are these steps are explained in detail in Section 3.2.1. One of the most important steps in Preprocessing is the normalization of datasets. Normalization is a key point in the work as it has a huge effect on the outcome of the classification and it is performed keeping in view some key aspects of investigations which include normal values and key attributes that affect the normal values like age and gender.

The next step involves splitting the dataset into two parts. One part consists of 70% of the dataset used for the training of models. Training means that models will learn on this dataset and this learning will be used in future prediction of data. The second part is testing, 30% of the total dataset used for testing purposes. Hyper-parameter tuning detail is given in Table 4. The effectiveness of the models is determined by the classification result of the testing dataset. The testing is done by removing the target class of the test dataset and then by feeding the test set to the model. The model predicts the target class of the testing dataset based on the learning of the model from the training dataset.

TABLE 4. Machine learning models and detail of hyper parameters.

ML Algorithms	Hyper parameters
SVM	kernel='linear' random_state=100, degree=2, tol=0.001
GBM	n_estimators=100, random_state=50
RF	n_estimator=200, random_state=12, max_depth=100
Voting Classifier	voting='hard'

For classification, the model used in this work is an ensemble model. The model is based on SVM, GBM and RF, and SGR-VC. The SVM classifier used in this model works well when it is used for a binary class like in our work. Our target class is binary, the patients are β-Thalassemia carriers or β-Thalassemia non-carriers. In the SGR-VC model, the training dataset is fed to all three classifiers. These classifiers are trained on the training dataset. Then the output of the classifiers is fed to the voting classifier. Voting classifier predicts on the basis of the highest number of votes to a class as shown in Fig. 4. After training, the test dataset is used as input, and predictions are made.

The last phase of the proposed model is the evaluation of the models. This is an important phase to check the effectiveness of the models. There are different tools and techniques that are used for evaluation purposes. The indicators in this research are Accuracy, Precision, Recall, and F1-Score. For the sake of comparison, the performance of the proposed model is compared with every classifier used in this model individually.

IV. RESULTS AND DISCUSSION

SVM, RF and GBM were first trained individually on the normalized training dataset, later these models are blended using SGR-VC to give final perdition based on the combination of all of these classifier's perdition by the voting mechanism. The Accuracy results of all the classifiers are shown in table 5 where it can be seen that the SGR-VC achieves the highest accuracy of 93% followed by RF and GBM with 91%, and then SVM with 90%. multirow lineno.

Accuracy alone does not demonstrate the performance of the classifier. SGR-VC achieves the highest precision value with 93% followed by SVM, GBM, and RF with 91% as shown in Table 5. It can be observed that precision of all the

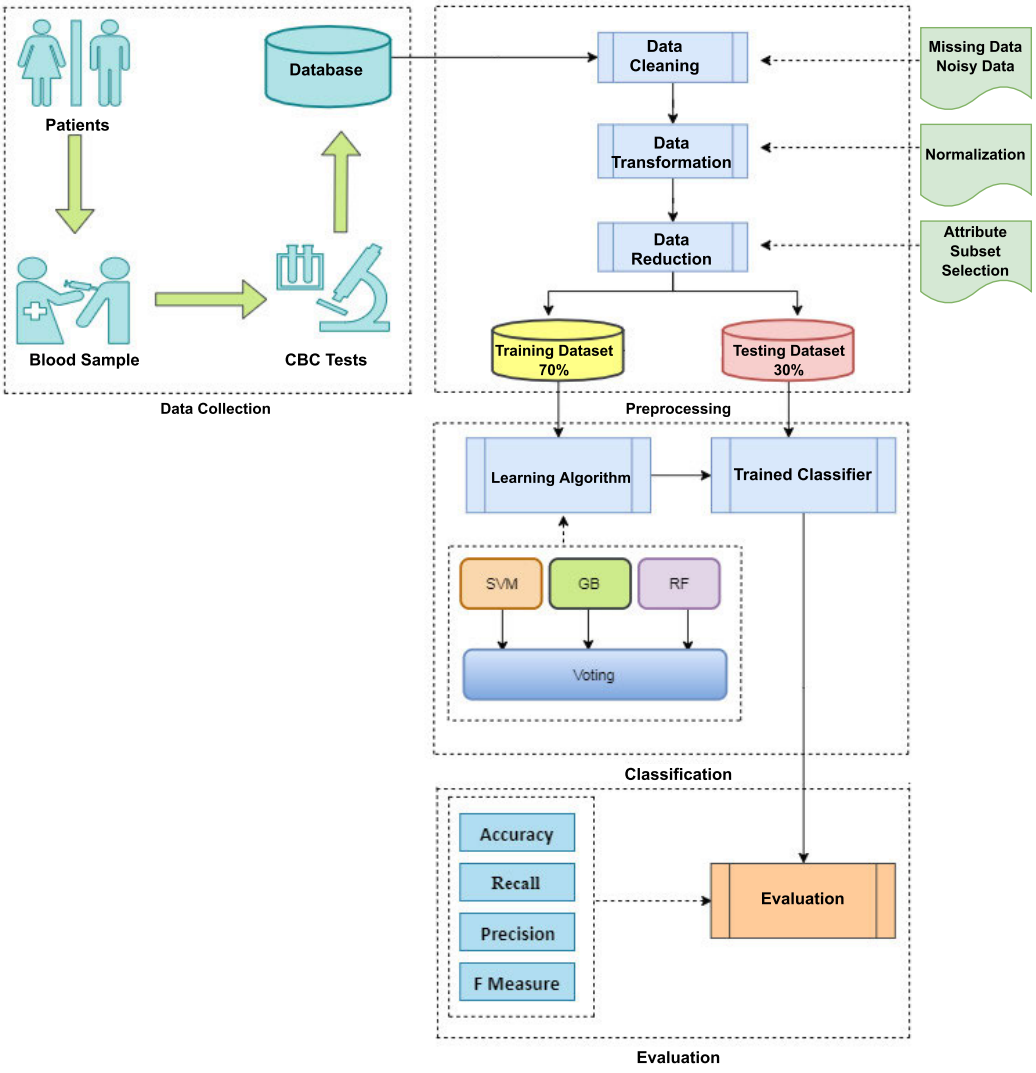


FIGURE 3. Architecture of the proposed framework.

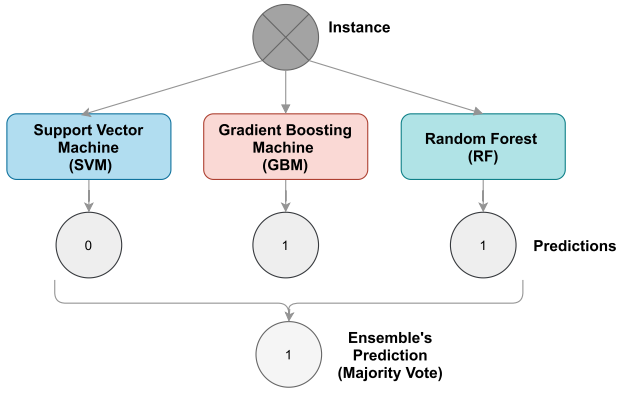


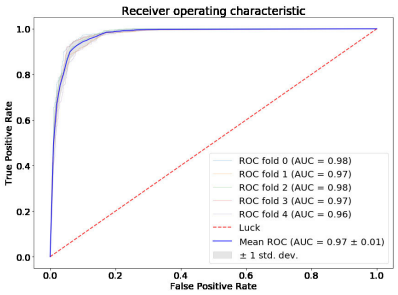
FIGURE 4. Architecture of the proposed voting classifier.

classifier for β-Thalassemia carrier is not high as compared to other evaluation measures which exhibit that classifiers occasionally tends to predict a β-Thalassemia non-carriers as β-Thalassemia carriers patients. After observing the predicted false positive instances it is concluded that these are the patients with very similar features of β-Thalassemia carrier patients but these instances are those patients suffering from anemia. Anemia has very similar symptoms to Thalassemia carrier patients. However, when the predictions of these three classifiers are blended using SGR-VC the value of precision improved and achieved higher precision as compared to individual classifiers.

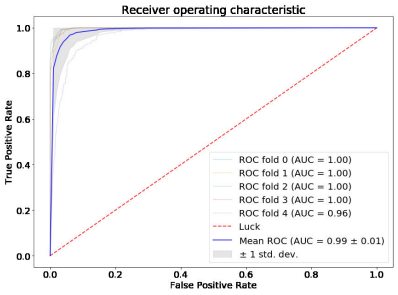
Recall is another important metric in medical diagnosis as it is important to reduce the misclassification of a Thalassemia carrier patient. Recall always has a greater value in medical-related work. It is evident that SGR-VC also attained the highest recall value with 93% closely followed by the individual classifiers with 91%. To better understand the balance between Recall and Precision we have compared the F1-Score of all classifiers. It can be observed that SGR-VC achieves the highest F1-Score with 93% followed by

TABLE 5. Classification results of machine learning models.

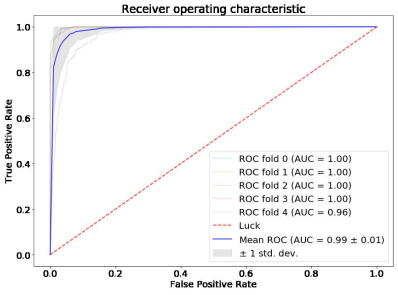
Classifier	Accuracy	β-Thalassemia Carrier			β-Thalassemia non-carrier			Weighted Average		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
RF	91%	90%	89%	90%	92%	93%	93%	91%	91%	91%
SVM	90%	87%	91%	89%	94%	91%	92%	91%	91%	91%
GBM	91%	88%	94%	92%	93%	92%	94%	91%	91%	91%
SGR-VC	93%	89%	89%	90%	96%	93%	93%	93%	93%	93%



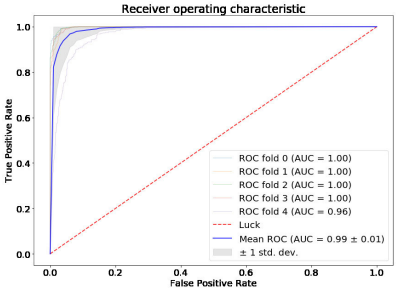
(a) ROC for RF



(b) ROC for SVM



(c) ROC for GBM



(d) ROC for SGR-VC

FIGURE 5. ROC curves.

GBM and RF with the value of 91%. The main purpose of this research work is to classify β-Thalassemia carriers and β-Thalassemia non-carriers.

For a better understanding of how efficiently our classifiers are able to distinguish between β-Thalassemia carriers and β-Thalassemia non-carriers. To prove the robustness of the proposed approach, we performed 5-fold cross-validation and plotted the ROC curve for RF, SVM, GBM, and SGR-VC and presented in Fig. 5. Where the x-axis represents a false positive rate and the y-axis represents a true positive rate. It can be seen clearly that SGR-VC is showing robust results as it is very close to 1.

If we compare the performance of the proposed model with previous works in literature, it can be observed that the majority of works have been performed to identify Thalassemia patients from IDA patients. But very few works are to identify Thalassemia carriers. Screening of Thalassemia carriers is of major importance to avoid and control this disease. Authors in [10] used a highly imbalanced dataset that consists of only 2.5% of Thalassemia carrier patients and which is very small in number to get sufficient features of positive class for proper training of the model. While this research makes use of a dataset consisting of 40% of β-Thalassemia carrier patients which is quite suitable to get enough features of the positive class. Authors performed β-Thalassemia carrier detection by limiting their research on women of fertile age [12]. They completely ignored other age groups and gender in their study. Our proposed model is more generic as it works on the dataset that includes both genders, children and adults.

V. CONCLUSION

As the number of β-Thalassemia patients is increasing in third world countries like Pakistan the exigency of early detection of β-Thalassemia carriers increases. Existing methods of β-Thalassemia carrier detection are expensive and time-consuming therefore screening of β-thalassemia carriers at a larger scale is an uphill task. In this work, an ensemble classifier for β-Thalassemia carrier screening is proposed. The dataset used in this work is compiled on CBC tests of 5066 cases which are collected from the database of the Punjab Thalassemia Prevention Program (PTPP). The proposed SGR-VC model for β-Thalassemia Carrier screening is composed of RF, SVM, and GBM. First RF, SVM, and GBM trained on normalized training data individually afterward voting classifier predict on testing data where the final prediction is made on the basis of majority voting. Accuracy, Recall, Precision, and F1-Score are the parameters used to evaluate the performance of the model.

We compared the performance of the proposed SGR-VC model with RF, SVM, and GBM individually. Experimental results confirm the superiority of the SGR-VC model with

a 93% value of Accuracy, Precision, Recall, and F1-score. Furthermore, Our proposed hybrid approach has the least number of misclassified patients which are 117. After the analysis of results, it can be observed that the proposed SGR-VC is very effective in β-Thalassemia carrier screening and it outperformed all other used models. This study mainly focuses on the performance analysis of tree-based and probability-based models for the classification task. Other regression-based machine learning models and deep learning models will be explored in the future for Thalassemia carrier detection.

REFERENCES

[1] T. Kallenbacha, "Anaesthesia for a patient with beta thalassaemia major," *Southern Afr. J. Anaesthesia Analgesia*, vol. 21, no. 5, pp. 21–24, 2015.

[2] E. P. Vichinsky, "Changing patterns of thalassemia worldwide," *Ann. New York Acad. Sci.*, vol. 1054, no. 1, pp. 18–24, Nov. 2005.

[3] B. G. Forget and H. F. Bunn, "Classification of the disorders of hemoglobin," *Cold Spring Harbor Perspect. Med.*, vol. 3, no. 2, Feb. 2013, Art. no. a011684.

[4] F. Kutlar, "Diagnostic approach to hemoglobinopathies," *Hemoglobin*, vol. 31, no. 2, pp. 243–250, 2007.

[5] A. Walaa, G. Kamal, T. Sallam Mohamed, and A. Soliman, "Blood indices to differentiate between β-thalassemia trait and iron deficiency anemia in adult healthy Egyptian blood donors," *Egyptian J. Haematology*, vol. 39, no. 3, p. 91, 2014.

[6] D. Weatherall, "The thalassemias: The role of molecular genetics in an evolving global health problem," *Amer. J. Hum. Genet.*, vol. 74, no. 3, pp. 385–392, 2004.

[7] A. A. Asadi-Pooya and M. Doroudchi, "Thalassemia major and consanguinity in Shiraz city, Iran," *Turk J. Haematol.*, vol. 21, no. 21, pp. 127–130, 2004.

[8] J. F. Matos, L. M. S. Dusse, K. B. G. Borges, R. L. V. D. Castro, W. Coura-Vital, and M. D. G. Carvalho, "A new index to discriminate between iron deficiency anemia and thalassemia trait," *Revista Brasileira de Hematologia e Hemoterapia*, vol. 38, no. 3, pp. 214–219, Jul. 2016.

[9] S. A. Pessar, "Evaluation of twenty four discriminant indices for differentiating beta-thalassemia trait from iron deficiency anemia in Egyptians," *Iranian J. Pediatric Hematol. Oncol.*, vol. 9, no. 3, pp. 135–146, Jun. 2019.

[10] A. S. Alagha, H. Faris, B. H. Hammo, and A. M. Al-Zoubi, "Identifying β-thalassemia carriers using a data mining approach: The case of the gaza strip, palestine," *Artif. Intell. Med.*, vol. 88, pp. 70–83, Jun. 2018.

[11] L. Kabootarizadeh, A. Jamshidnezhad, Z. Koohmareh, and A. Ghamchili, "Differential diagnosis of iron-deficiency anemia from beta-thalassemia trait using an intelligent model in comparison with discriminant indexes," *Acta Inf. Medica*, vol. 27, no. 2, p. 78, 2019.

[12] I. Lachover Roth, B. Lachover, G. Koren, C. Levin, L. Zalman, and A. Koren, "Detection of β thalassemia carriers by red cell parameters obtained from automatic counters using mathematical formulas," *Medit. J. Hematol. Infectious Diseases*, vol. 10, no. 1, Jan. 2018, Art. no. 2018008.

[13] F. R. Aszhari, Z. Rustam, F. Subroto, and A. S. Semendawai, "Classification of thalassemia data using random forest algorithm," *J. Phys., Conf. Ser.*, vol. 1490, Mar. 2020, Art. no. 012050.

[14] C. Bellinger, A. Amid, N. Japkowicz, and H. Victor, "Multi-label classification of anemia patients," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 825–830.

[15] M. K. Jamei and K. M. Talarposhti, "Discrimination between iron deficiency anaemia (IDA) and β-thalassemia trait (β-TT) based on pattern-based input selection artificial neural network (PBIS-ANN)," *J. Adv. Comput. Res.*, vol. 7, no. 4, pp. 55–66, 2016.

[16] V. M. Khadse, P. N. Mahalle, and G. R. Shinde, "Statistical study of machine learning algorithms using parametric and non-parametric tests: A comparative analysis and recommendations," *Int. J. Ambient Comput. Intell.*, vol. 11, no. 3, pp. 80–105, Jul. 2020.

[17] P. Paokanta and N. Harnpornchai, "Risk analysis of thalassemia using knowledge representation model: Diagnostic Bayesian networks," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat.*, Jan. 2012, pp. 155–158.

[18] A. Upadhyay, "β-thalassemia major and minor classification using artificial neural network," *Int. J. Comput. Appl.*, vol. 13, pp. 14–17, Feb. 2013.

[19] T. Sandanayake, A. Thalewela, H. Thilakesooriya, R. Rathnayake, and S. Wimalasooriya, "Automated thalassemia identifier using image processing," Tech. Rep. 53696127, 2016.

[20] R. HosseiniEshpala, M. Langarizadeh, M. KamkarHaghighi, and T. Banafsheh, "Designing an expert system for differential diagnosis of β-thalassemia minor and iron-deficiency anemia using neural network," *Hormozgan Med. J.*, vol. 20, no. 1, pp. 1–9, 2016.

[21] A. H. Shurrab and A. Y. A. Maghari, "Blood diseases detection using data mining techniques," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 625–631.

[22] K. Sharma and J. Virmani, "A decision support system for classification of normal and medical renal disease using ultrasound images: A decision support system for medical renal diseases," *Int. J. Ambient Comput. Intell.*, vol. 8, no. 2, pp. 52–69, Apr. 2017.

[23] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "A novel stacked CNN for malarial parasite detection in thin blood smear images," *IEEE Access*, vol. 8, pp. 93782–93792, 2020.

[24] N. Dey, A. Ashour, A. Ashour, and A. Singh, "Digital analysis of microscopic images in medicine," *J. Adv. Microsc. Res.*, vol. 10, no. 1, pp. 1–13, Mar. 2015.

[25] S. Acharjee, S. Chakrabartty, M. I. Alam, N. Dey, V. Santhi, and A. S. Ashour, "A semiautomated approach using GUI for the detection of red blood cells," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 525–529.

[26] A. M. El-Halees and A. H. Shurrab, "Blood tumor prediction using data mining techniques," *Blood Tumor Predict. Data Mining Techn.*, vol. 6, no. 2, May 2017.

[27] N. C. Egejuru, S. O. Olusanya, A. O. Asinobi, O. J. Adeyemi, V. O. Adebayo, and P. A. Idowu, "Using data mining algorithms for thalassemia risk prediction," *Sci. Eng.*, vol. 7, no. 2, pp. 33–44, 2019.

[28] A. G. Ismaeel, "Diagnose mutations causes β-thalassemia: Biomining method using an optimal neural learning algorithm," *Int. J. Eng. Technol.*, vol. 8, nos. 1–11, pp. 1–8, 2019.

[29] M. Al-Haggar, M. Rasmy, M. Islam, A. Darwish, and B. Khair-Allah, "Genotype-phenotype correlation and the severity index for β-thalassemia," *Proteomics Bioinf. Current Res*, vol. 1, no. 1, pp. 17–28, 2019.

[30] B. Çil, H. Ayyıldız, and T. Tuncer, "Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system," *Med. Hypotheses*, vol. 138, May 2020, Art. no. 109611.

[31] C. Gold and P. Sollich, "Model selection for support vector machine classification," *Neurocomputing*, vol. 55, nos. 1–2, pp. 221–249, Sep. 2003.

[32] Y. Ma and G. Guo, *Support Vector Machines Applications*, vol. 649. Cham, Switzerland: Springer, 2014.

[33] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Proc. Int. Conf. Inf. Comput. Appl.* Berlin, Germany: Springer-Verlag, 2012, pp. 246–252.

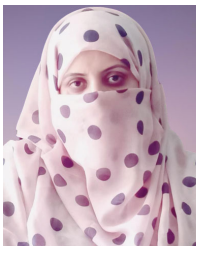
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[35] Y. Zhang, H. Zhang, J. Cai, and B. Yang, "A weighted voting classifier based on differential evolution," *Abstract Appl. Anal.*, vol. 2014, pp. 1–6, May 2014.

[36] M. A. Arbib, *The Handbook of Brain Theory and Neural Networks*, 2nd ed. Cambridge, MA, USA: MIT Press, 2002.

[37] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Emotion recognition by textual tweets classification using voting classifier (LR-SGD)," *IEEE Access*, vol. 9, pp. 6286–6295, 2021.

[38] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, Aug. 2020.



SAIMA SADIQ is currently pursuing the Ph.D. degree in computer science with the Khwaja Fareed University of Engineering and Information Technology (KFUEIT). She is also working as an Assistant Professor with the Department of Computer Science, Government Degree College for Women. Her recent research interests include data mining, machine learning, and deep learning-based text mining.

MUHAMMAD USMAN KHALID received the M.S. degree in computer science from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2020. He is currently serving as a Software Engineer with the Shaikh Zayed Medical College and Hospital, Rahim Yar Khan. His recent research interests include data mining, machine learning and deep learning-based IoT, text mining, and computer vision tasks.

MUI-ZZUD-DIN was born in Multan, Pakistan, in 1980. He received the B.Sc. and M.I.T. degrees in computer science and the M.S. (C.S.). degree from Bahauddin Zakariya University, Multan, in 2001, 2005, and 2014, respectively. From 2007 to 2017, he was a lecturer with different educational institutes. Since February 2017, he has been working as a Lecturer with the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan. He has almost ten years of teaching experience. His current research interests include image processing, the IoTs, cognitive radio networks, and data science.

SALEEM ULLAH was born in Ahmedpur East, Pakistan, in 1983. He received the B.Sc. degree in computer science from The Islamia University Bahawalpur, Pakistan, in 2003, the M.I.T. degree in computer science from Bahauddin Zakariya University, Multan, in 2005, and the Ph.D. degree from Chongqing University, China, in 2012.

From 2006 to 2009, he worked as a network/IT administrator with different companies. From August 2012 to February 2016, he worked as an Assistant Professor with Islamia University Bahawalpur. Since February 2016, he has been working as the Associate Dean with the Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan. He has almost 14 years of industry experience in field of IT. He is also an active Researcher in the field of Adhoc networks, the IoTs, congestion control, data science, and network security.

WAQAR ASLAM received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, and the Ph.D. degree in computer science from the Eindhoven University of Technology, The Netherlands. He is currently an Assistant Professor with the Department Computer Science and IT, The Islamia University of Bahawalpur, Pakistan. His research interests include performance modeling and QoS of wireless/computer networks, performance modeling of (distributed) software architectures, radio resource allocation, the Internet of Things, fog computing, effort/time/cost estimation of software development in (distributed) agile setups, social network data analysis, and DNA/chaos-based information security. He received the Overseas Scholarship from HEC, Pakistan for his Ph.D.

ARIF MEHMOOD received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, South Korea, in November 2017. Since November 2017, he has been a Faculty Member with the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Pakistan. His recent research interest includes data mining, mainly working on AI and deep learning-based text mining and data science management technologies.

GYU SANG CHOI received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University at University Park, PA, USA, in 2005. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been a Faculty Member with the Department of Information and Communication, Yeungnam University, South Korea. His research interests include non-volatile memory and storage systems.

BYUNG-WON ON received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University at University Park, PA, USA, in 2007. For seven years, he worked as a full-time Researcher with The University of British Columbia, the Advanced Digital Sciences Center, and the Advanced Institutes of Convergence Technology. Since 2014, he has been a Faculty Member with the Department of Software Convergence Engineering, Kunsan National University, South Korea. His recent research interests include data mining, especially probability theory and applications, machine learning, and artificial intelligence, mainly working on abstractive summarization, creative computing, and multi-agent reinforcement learning.