Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan Vol. 24 No. 4-2019
ISSN:1813-1786 (Print) 2313-7770 (Online)

Hematological Diseases Diagnosing Using Machine Learning

J.Hanif[1], M.M. Iqbal[2]

[1, 2]*Department of Computer Science, University of Engineering and Technology Taxila, Pakistan*

[1]javeria389@gmail.com

Abstract- The amount of information related to Hematological diseases and patients has expanded significantly. Hematological diseases should be diagnosed on time. Pathologists use conventional diagnostic tests to diagnose hematological diseases which are often low in cost and give inaccurate diagnostic results. Complete Blood Count (CBC) is used to finding the existence of diseases. Machine learning helps to elucidate acquired data and in predicting hematological diseases. Classification is used to diagnose diseases that categorize features into respective four target classes e.g., Anemia, Leukemia, Thalassemia, and healthy patients. Five machine learning algorithms were used with all features and reduced features in this investigation. The most efficient algorithm found is Random Forest with the highest accuracy at 98.59% with the lowest error rate of 0.06%. Findings show that the first indicator for blood disease is Hemoglobin.

Keywords- Complete Blood Count, Hematocrit, Hemoglobin, Packed Cell Volume, Red blood cell, White Blood Cell.

I. Introduction

Data related to the medical field is quite vast and complex. Medical data contain hidden patterns that are comprised of information related to diseases. If hidden diseases are diagnosed correctly, then the mortality rate can be reduced.

Machine learning algorithms can be used for prediction and classification purposes in the medical field[1]. Machine learning algorithms also aid doctors in diagnosing hematological diseases in cheap rates and provide more accurate results to patients. A worldwide survey for hemophilia was conducted in 2016, and statistic shows that there were about 184,723 people diagnosed with hemophilia diseases. Accounting of 24.8% of people is suffering from anemia, while 5-13% of people are affected by thalassemia. Hematological issues are leading to Down syndrome abnormalities in 1 out of 700 population[2]. Accounting of 9% of all cancer, Hematological Blood cancer is the fourth most commonly diagnosed cancer in developed countries according to Globocan 2012 Cancer incidence[3].

Hematological diseases occur due to malfunctioning of red blood cells, white blood cells, platelets, blood vessels, bone marrow, spleen, and the proteins involved in bleeding and clotting. Hematological diseases are classified according to abnormalities in the blood which causes blood disorder. Children who are suffering from Iron deficiency also suffer from poor verbal skills[4]. Hematological diseases are classified into two types inherited and acquired. For instance, hemophilia, thalassemia, blood clotting and sickle cell are inherited disorders, whereas anemia and blood clots are acquired diseases[5]. Inherited blood diseases commonly occur during childhood, and Thalassemia is a genetic disease that can be controlled by stopping marriages between both people who have carried the same genes[6-9].Hematological diseases can be benign and malignant. Benign hematological diseases can be cured. While Hematological Malignancies badly affect immune sensitivity and hinder treatment progress[10].At any age, Hematological malignancies can be diagnosed, and survival patterns differ between children and adults[11]. From a clinical point of view, complete blood count is the most effective diagnostic profile given the widespread use and interpretation of hemophilia[12]. Diagnosing Hematological diseases is challenging due to similarities in appearance. Classification is a way to diagnose a type of disease. Labeling can be used to improve diagnosis[13].

In this paper, three types of hematological diseases are discussed: Anemia, Leukemia, and Thalassemia and patients have been discriminated from healthy people.

To achieve the desired results, we have proposed an automated technique to diagnose and classify diseases. Complete blood count tests are used in order to get a better understanding of underlying blood disorders. Our main purpose of this research is to improve the classification and diagnostic results of hematological

disorders. All experiments are done only on three types of blood diseases.

The remainder of this paper is structured as follows. Section 2 discusses the related work. Section 3 describes our proposed method. Section 4, Dataset, and preprocessing. In Section 5, experimental results and discussions are addressed. In Section 6, the conclusion and future work are given.

II. Related Work

In recent years, Worldwide Healthcare Organizations have faced many challenges to diagnose diseases. Several types of researches are made to give and improve the ways to diagnose diseases. There are different ways given to detect and diagnose diseases using a different technique.

In this paper, data mining techniques of classification were used to address the diagnostic problems of hematological diseases. By using hematological data, the Random forest classifier proved to be best with the highest accuracy of 96.47% to diagnose a disease [14]. The author investigated the presence of thalassemia by using data mining classifiers depending on complete blood count. Three classifiers were used to differentiate thalassemia traits patients. Experiments were done on the dataset with full features and reduced features. Findings show that mean corpuscular volume (MCV) is the main feature to find the existence of blood disease. But more investigations are required to determine results for reduced classes [15]. A survey has been conducted to define path physiology of acute myeloid leukemia. Almost 5234 mutations were identified across 76 genes with 2 or more driver mutations that were detected in 86% of patients. Diseases prediction required to be validated in prospective clinical trials [16]. In this research, the author proposed a model based on a hybrid data mining approach which aided pathologists in diagnosing β-thalassemia. It mainly consists of two steps. One step is Synthetic Minority Over-sampling Technique (SMOTE) and second one is problem handling through different models. Results showed that SMOTE effectively identified patients suffering from β-thalassemia. Naïve Bayes showed the best performance in differentiating between normal patients and β-thalassemia [17]. Another study highlighted an investigation of different types of anemia diseases. Complete blood count tests were used to diagnose types of anemia. Five classification algorithms were used and compared with the voting algorithm on the basis of higher accuracy and lower error rate. Results show that the Hybrid Algorithm would be best to diagnose any type of anemia diseases with a higher accuracy rate instead of a single algorithm [18].

The author investigated the use of the artificial neural network for the classification of thalassemia. All hematology parameters were taken from hemachromocytometric analysis. An automated system was proposed for thalassemia diagnosing. Different artificial neural networks were used which discriminates normal patients from thalassemia patients with 94% classification accuracy, 92% sensitivity and 95% specificity [19]. An analysis of the prediction and classification of anemia patients has been presented. The classification was applied to different complete blood count samples. Decision Tree showed the best result for the classification of anemia disease [20]. By using a wireless multimedia medical sensor network, a Context-sensitive Seamless Identity Provisioning(CSIP) mutual authentication network model has been proposed. It showed the improvement in the performance of the health care system and also increased utilization of bandwidth [21].It reduces the uncertainty factor in cloud-based IoT, a framework has been introduced. This framework comprises Artificial Intelligence (AI) techniques and it improved the system performance by 60% overall as compared to the centralized version [22]. By using the Internet of Medical (IoM), a secure biometric authentication scheme has been proposed for electronic healthcare applications. It prevents secret information from any leakage and also performs efficiency analysis to make certain information security [23].

III. Proposed Framework

The primary purpose of this research is to diagnose hematological diseases. Machine learning techniques are used to differentiate blood diseases. In order to suggest a Naïve method with higher accuracy and lower error rate, machine learning algorithms are used. The proposed technique consists of four blocks to solve a problem: 1) Dataset 2) Preprocessing 3) Classification 4) Performance measure and evaluation. Figure 3.1. shows the proposed model.

3.1 Dataset Collection and Preprocessing

The hematological dataset is collected from different blood test laboratories. The dataset comprises 531 samples, of which 230 samples are of males and 301 samples are of female patients. Overall, 349 suffering from anemia, 92 leukemia, 45 thalassemia and 45 people are considered healthy. Some features are dropped out due to privacy concerns. Dataset has four different labels: Anemia, Leukemia, Thalassemia and No. In the preprocessing of the dataset, all irrelevant attributes are eliminated; all missing values are refilled and outliers and extreme values are removed.

3.2 Classification

Five machine learning algorithms are used to perform classification: Naïve Bayes, Multilayer Perceptron, Vote, J48 and Radom Forest. The classification has been performed in two groups: One group with all features and the second group with reduced features.

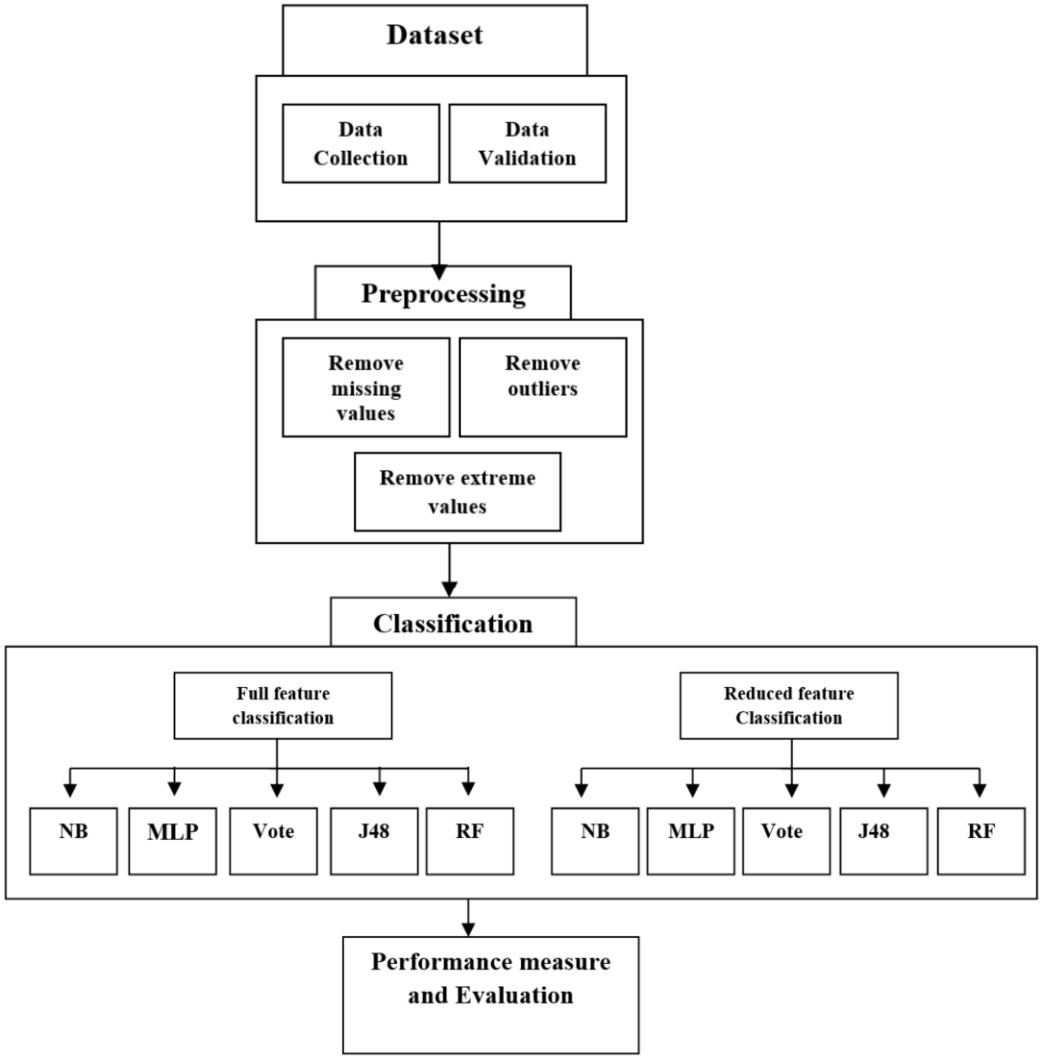

Figure 3.1. Methodology for hematological diseases classification

It classifies features into four labels: Anemia, Thalassemia, Leukemia, and No. All features are taken from the CBC test. 100 Batch size is set for all experiments.

IV. Results and Discussions

The results of machine learning algorithms are shown on the basis of accuracy, precision and error rate. Analysis has been performed for both classifications with full features and reduced features. To evaluate the accuracy of diagnoses, 80% of data is used for training, 20% for testing. The results of the classification are shown in this section.

4.1. Classification with Full Features:

The accuracy, precision, and error of each classifier are evaluated by using the WEKA tool. Results of each algorithm are assessed and an algorithm with the best result is used to diagnose a disease. In this research, five algorithms are used including: Naïve Bayes, Multilayer Perceptron, Vote, J48 and Random Forest. In Table 1 Comparison of classifiers with all features is shown:

Table 1: Comparison of classifiers with all features

Index	Classifier	Accuracy	Precision	Error Rate
1	Naïve Bayes	87.32	0.889	0.119
2	Multilayer Perceptron	66.17	0.67	0.173
3	Vote	60.563	0.605	0.290
4	J48	60.563	0.605	0.290
5	Random Forest	98.59	0.989	0.072

According to Table 1 accuracy could be regarded as the basis to compare all classifiers and diagnose diseases. This table clearly shows that the Random forest gives the best results for disease diagnosing. In Figure 1. Comparison of accuracy, precision and mean error rate is shown:

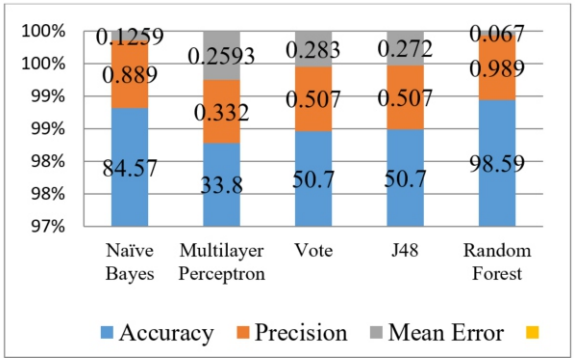

Figure 1: Comparison of accuracy, precision and mean error value with full features

4.2. Analysis with Reduced features

In the present research, reduced features by the same algorithms are determined and the result of each algorithm is compared. The algorithm with the best accuracy is considered to diagnose disease. A comparison of classifiers with reduced features is shown in Table 2.

Table 2: Comparison of classifiers with reduced features

Index	Classifier	Accuracy	Precision	Error Rate
1	Naïve Bayes	83.09	0.868	0.125
2	Multilayer Perceptron	53.52	0.821	0.211
3	Vote	50.70	0.507	0.27
4	J48	50.70	0.507	0.27
5	Random Forest	95.77	0.96	0.076

As Table 2 shows that an algorithm with the highest accuracy is considered the most efficient model to diagnose diseases. Among all algorithms, the Random forest is the most efficient model with the highest accuracy and lowest error rate. In Figure 2. Comparison of accuracy, precision and mean error rate is shown.

4.3. Rules Induction

Diagnosed diseases depend on some rules. If any value is not lying in range, then it indicates the presence of some hematological issue. Some rules are listed in Table 3.

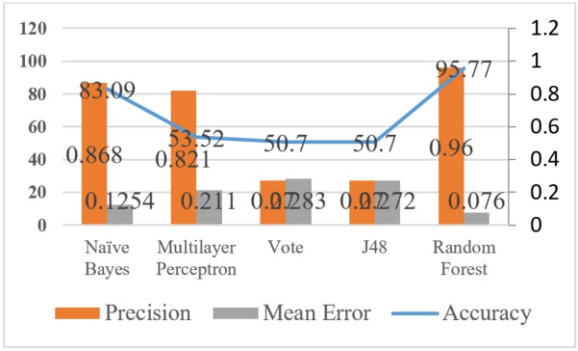

Figure 2: Comparison of accuracy, precision and mean error value with reduced features

Sr.#	Rules	Disease
1	if(HB <12) and (HCT <37) and (RBC<4)then	Anemia
2	else if (HB<12) and (HCT<37)and (RDW<55) then	Anemia
3	else if (HB<12) and (HCT<37) and (MCV<80) then	Anemia
4	else if (HB<12) and (MCH<32) and (MCV<80) then	Thalassemia
5	else if (HB<12) and (RBC<4) and (MCV<80) then	Thalassemia
6	else if (MCHC<3) and (HCT<37) and (HGB<12) then	Anemia
7	else if (RBC<3)and (PLT<400,000) and (WBC>13000)	Leukemia

The objective of this research is to diagnose hematological diseases using CBC and WEKA software. A technique is provided to diagnose hematological diseases Anemia, Thalassemia, leukemia from normal subjects.

Classifications with all CBC samples are divided into two groups: one with all features and second with reduced features. Some of the best algorithms were used to analyze hematological data. The performance of all algorithms has been compared. Analyze all features and reduced features, it is found that Random Forest is the best classifier with an accuracy of 98.59% and 95.77%. Findings showed Random Forest classifier would help pathologists to diagnose diseases using the CBC sample.

Conclusion and Future Work

In this research, a solution has been proposed for empowering doctors and patients to diagnose diseases. This paper assessed five machine learning algorithms based on WEKA software. By using the hematological data, Random forest proved to be best for disease diagnosing with the accuracy of 98.59% for full features and 95.77% for reduced features. Random forest took less time of 0.16s as compare to other

algorithms. Vote and J48 showed the least accuracy of 60.563% for the full feature and 50.70% with the reduced feature. The obtained results from sample data are overall satisfactory in identifying the patient's disease. In the future, more diseases will be diagnosed using the proposed technique. The accuracy and performance enhancements, advanced technologies will be used.

REFERENCE

[1] M. S. Borah, B. P. Bhuyan, M. S. Pathak, and P. K. Bhattacharya, "Machine Learning in Predicting Hemoglobin Variants," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 2, pp. 140–143, 2018.

[2] M. Mahamood *et al.*, "Hematological Studies in Young Individuals with Down Syndrome," no. December, pp. 8–10, 2014.

[3] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.," *Int. J. cancer*, vol. 136, no. 5, pp. E359-86, 2015.

[4] P. East, E. Delker, E. Blanco, P. Encina, B. Lozoff, and S. Gahagan, "Effect of Infant Iron Deficiency on Children ' s Verbal Abilities : The Roles of Child Affect and Parent Unresponsiveness," *Matern. Child Health J.*, no. 0123456789, 2019.

[5] B. A. Khair-allah, "Studying Correlation Between Genotype and Beta Thalassemia Major Severity Factors," no. 1, pp. 7–10, 2016.

[6] D. Setsirichok, T. Piroonratana, W. Wongseree, and T. Usavanarong, "Biomedical Signal Processing and Control Classification of complete blood count and haemoglobin typing data by a C4 . 5 decision tree , a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening," *Biomed. Signal Process. Control*, vol. 7, no. 2, pp. 202–212, 2012.

[7] A. Akay, A. Dragomir, A. Yardimci, D. Canatan, A. Yesilipek, and B. W. Pogue, "A data-mining approach for investigating social and economic geographical dynamics of β-thalassemia's spread," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 774–780, 2009.

[8] P. Paokanta and N. Hampomchai, "Risk Analysis of Thalassemia Using Knowledge Representation Model : Diagnostic Bayesian," vol. 25, no. Bhi, pp. 155–158, 2012.

[9] D. Weatherall, "2003 WILLIAM ALLAN AWARD ADDRESS The Thalassemias : The Role of Molecular Genetics in an Evolving Global," pp. 385–392, 2004.

[10] P. Armand and P. Armand, "Immune Checkpoint Blockade in Hematologic Malignancies Short title : Checkpoint Blockade in Hematologic Malignancies Corresponding author : Copyright © 2015 American Society of Hematology," pp. 617–632, 2015.

[11] J. Li, A. Smith, S. Crouch, S. Oliver, E. Roman, and A. Smith, "Estimating the prevalence of hematological malignancies and precursor conditions using data from Haematological Malignancy Research Network (HMRN)," *Cancer Causes Control*, 2016.

[12] S. Piplani, M. Madan, R. Mannan, and M. Manjari, "Original Article Evaluation of Various Discrimination Indices in Differentiating Iron Deficiency Anemia and Beta Thalassemia Trait : A Practical Low Cost Solution."

[13] A. Haldar, G. P. Raj, and S. V. S. S. Lakshmi, "Comparison of Different Classification Techniques Using WEKA for Diabetic," pp. 509–516, 2018.

[14] F. Akter, M. A. Hossin, G. M. Daiyan, and M. M. Hossain, "Classification of Hematological Data Using Data Mining Technique to Predict Diseases," *J. Comput. Commun.*, vol. 06, no. 04, pp. 76–83, 2018.

[15] E. H. Elshami and A. M. Alhalees, "Automated Diagnosis of Thalassemia Based on DataMining Classifiers."

[16] P. Paschka *et al.*, "Genomic Classification and prognosis in acute Myeloid Leukemia," *N. Engl. J. Med.*, vol. VOl.374, 2016.

[17] A. S. Alagha, H. Faris, B. H. Hammo, and A. M. Al-zoubi, "Arti fi cial Intelligence In Medicine Identifying β -thalassemia carriers using a data mining approach : The case of the Gaza Strip , Palestine," *Artif. Intell. Med.*, no. December 2017, pp. 0–1, 2018.

[18] M. Hasani and A. Hanani, "Automated Diagnosis of Iron Deficiency Anemia and Thalassemia by Data Mining Techniques," vol. 17, no. 4, pp. 326–331, 2017.

[19] S. R. Amendolia *et al.*, "A Real-Time Classification System of Thalassemic Pathologies Based on Artificial Neural Networks," no. September, 2001.

[20] S. A. Sanap, M. Nagori, and V. Kshirsagar, "Classification of Anemia Using Data Mining Techniques," pp. 113–114, 2011.

[21] F. Al-turjman and S. Alturjman, "Context-sensitive Access in Industrial Internet of Things (IIoT) Healthcare Applications," vol. 3203, no. c, 2018.

[22] F. Al-turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying Uncertainty in Internet of Medical Things and Big-Data Services Using Intelligence and Deep Learning," *IEEE*.

[23] B. D. Deebak, F. Al-turjman, M. Aloqaily, and O. Alfandi, "An Authentic-Based Privacy Preservation Protocol for Smart e-Healthcare Systems in IoT," *IEEE Access*, vol. PP, p.1,2019.