Predicting Anemia Using Tree-Based Classification Algorithms and Feature Selection

Ihsan Salman[1], Arshed A. Ahmad[2], Sahar Jasim Mohammed[3], Mohammed S. Mohammed[3] and Qusay Kanaan Kadhim[4]

[1]*Department of Computer Science, College of Basic Education, University of Diyala, 32001 Baqubah, Iraq*
[2]*University of Diyala, 32001 Baqubah, Iraq*
[3]*Department of Computer Science, College of Education for Pure Science, University of Diyala, 32001 Baqubah, Iraq*
[4]*Department of Computer Science, University of Diyala, 32001 Baqubah, Iraq*
dr.arshed.adham@uodiyala.edu.iq, drihsan@uodiyala.edu.iq, m.sc.sahar.jasim.m@uodiyala.edu.iq, dr.mohammed.sami@uodiyala.edu.iq, dr.qusay.kanaan@uodiyala.edu.iq

Keywords: Anemia, Logistic Model Tree, Hoeffding Tree, Random Tree, Prediction.

Abstract: The Paper discussed the performance of 3 classification techniques (Hoeffding Tree, Logistic Model Tree, and Random Tree) by applying it on a dataset of 539 samples. This was done using 11 features related to blood tests regarding anemia classes by applying 10-fold cross-validation to each technique with time consideration as well. The process was done firstly without feature minimization and showed that LMT achieved the highest accuracy with about 85.53%, then followed by the Random Tree and the Hoeffding Tree with approximately 1 and 2 differences respectively. Classification time also evaluated for this process with leading amplitude for the Random Tree about 0.02 seconds, then followed by the Hoeffding Tree and the LMT with higher running time due to the long mathematical calculations of LMT and Hoeffding Tree. In the second step, feature selection was applied and the elimination mechanism was done for features that considered as less impact on the prediction results such as Gender, Age, and WB. These combinations of feature selection provided better results for three utilized techniques that reach 96.27% for LMT, 94.42 for the Hoeffding Tree's, and 91.26% for th Random Tree. The results highlighted the impact features on Anemia dataset that should be assembled for such a clinical dataset to provide accuracy improvement and lower processing time.

1 INTRODUCTION

Medical data diagnoses and classification attract researchers of the data mining field and data analyses in the last few years. Serious diseases which threaten human lives were the main targets in these studies to predict disease behavior depending on previous different patients' data recorded in healthcare centers like hospitals, research centers, or else. Many different data-mining algorithms in literature are used to classify several types of diseases, into specific types based on the data-mining algorithms [1]. Anemia is considered as one of the complicated disorders due to unpredictable symptoms or not clear to the affected patients. This is one of several reasons that people may expose their health to danger without knowing the cause. The duty of these prediction systems that programmers designed will save their life or to reduce the seriousness of late prediction of this disease. Researchers are aware of applying and studying such disorders such in [2]. Iron deficiency was introduced and calculated for children in this paper with applying fuzzy system, it was employed for only two features and gave a review comparison between its results and the previous results at the same field and methods. Some of these studies get the classical way and normal tree classifications or by using an optimization technique like sequential. Experiments were the main layout that authors applied for anemia and depends on patient's attributes, which either collected from hospitals or healthcare. Some of these studies deal with Anemia disease as a binary problem issue such in [3], when authors applied six techniques as a multiple class detection for the 3 required prediction classes based on data collected from an infected zone in some places in India. Balancing data with key selections to the most voting attributes, perception was provided

with the best results with 10 validations among datasets. Researchers compared between some algorithms to classify and clear the way of prediction system according to the collected data. Generating and assembling data from patients is important but not adequate without processing to import the knowledge from these data. Techniques are applied to different datasets regardless of knowing the matching ability between techniques and datasets. In addition, prediction process has a time scale, feature affection and accuracy to examine its efficiency. In [4], a review study was done by some authors to list the points between some techniques to test and examine its according to anemia information. Worthy studies that considering anemia dataset as a model to lead the way of anemia detection and to get time racing using multiple classification methodologies was explained and studied by years such in [5]-[8]. Also, various studies on blood disease have been conducted by various methods as in the literature [9], [10]. This study considers a problem statement to classify targeted anemia diseases dataset with three classification methods to build prediction system with high accuracy and ideal time in two phases, first with all given attributes, second, use feature selection to reduce attributes. Logistic Model Tree (LMT) with feature selection provides significant results among other suggested models. Based on the different parameters of the blood samples have to be examined to carefully detect the required classes at the early period of this disease.

The major gap that made authors deal with this type of field is the instability with limited performances in Anemia prediction especially for large dataset with varying features [11]. In addition, articles such in [12] deal with Anemia as binary classification for the presence or the absence of disease, but not several types of Anemia. The detection process in this article was done for different Anemia classes as well with plenty type of techniques without providing more combination methods for efficiency improvement such like feature specifications such in [13], when authors utilized classical Decision Tree for this purpose. In addition, several studies as well do not specify the challenge of feature selection adequately, often using all attributes without considering their related to the prediction process, which can lead to unnecessary noise and longer processing times as studied in [14]. The choice of specific techniques in this article like Hoeffding Tree, LMT, and Random Tree due to several reasons, for example the Hoeffding Tree is suitable for large dataset with varying features and limited memory space as in [15]. This characteristic is important especially when dealing with healthcare datasets that usually changed or varied by time and samples. In the other hand, the LMT combines decision tree logic with logistic regression, that make this technique the probability to deal with categorized and continues dataset as in [16]. Also, the selection of the third techniques Random Tree, due to the process of creating multiple tree with a combination scheme to minimize overfitting that occurred in this type of dataset. The flexibility property of Random Tree makes it the best selection for complex classification tasks such for healthcare detection process [17]. While previous studies on anemia prediction have made strides in applying decision tree models (e.g., [12], [15]), they often lack a comprehensive evaluation of the trade-offs between model complexity and computational efficiency. Moreover, there is limited research comparing different tree-based algorithms on the same dataset with a focus on minimizing both prediction error and processing time. By integrating feature selection, this study aims to fill this gap by improving both the accuracy and the efficiency of anemia classification, offering a practical solution for real-world healthcare applications. Furthermore, the selection of these tree-based algorithms allows for a more nuanced understanding of how different features interact and influence anemia diagnosis, thus contributing to more accurate and timely clinical decision-making.

2 STUDY SAMPLES OF THE MEDICAL DATASET

The information used in this study was gathered through anemia reports which included (6–56) years of age ranges for (539 patients, 211 normal individuals, and 328 sick subjects) that were reported in the literature [18]. Acute or chronic diseases have an associated progress related to parameters and elements which referred to be risk factors such as [19], where pointed to several diseases like COVID-19 to define the class results and outcome based on these attributes and properties. The investigation of risk factors was also adopted by authors in [20], [21] to study it according to cardiovascular disease and diabetic patients. Researchers list a few blood disorders, including spherocytosis, iron deficiency anemia, vitamin B12 deficiency, thalassemia, and sickle cell disease (5). Samples that uploaded to this system have some attributes related to each patient who is suffering or not suffering from this disorder. These attributes or

features are independent to each other, in addition, it was read patient by patient after converting these data to a proper form to deal with it. The sequence of these data was not important, especially when it was considered as an independent variable, but each parameter has some effect on the disease differs than other parameters which make it hard to predict in normal ways. Tests were made for each patient, like the Hemoglobin test that consists of some anemia related parameters such as (RBC and MCH). Ten parameters were collected for each sample including the age and the gender type. However, this prediction system which was designed in this paper is addressed for 6 classes of Anemia (one for Anemia absence and the other are related to Anemia types) as explained more in [22]. The dataset is available online in [18], included a details explanation of biomedical features for the classification of anemia types. This dataset consisted of ten features as mentioned in the previous section which included Hemoglobin, the count of Red Blood Cell Count, the count of White Blood Cell. Also, this dataset consists of Platelet Count, Mean Corpuscular Volume, Mean Corpuscular Hemoglobin, Serum Ferritin, Transferrin Saturation, Serum Iron in addition to Age. These features were selected based on their related to anemia and the availability with 6 classes. The dataset classes are Iron Deficiency, Hemoglobin production, Vitamin B12 Deficiency, Folate Deficiency, Aplastic, Hemolytic in addition to Sickle Cell. These classes specified unique classes with a suitable distribution.

3 THE PROPOSED METHODS

The variety in classification models or algorithms that were used for decision making or prediction design for medical assembled data provides multiple and different criteria for the specific problem according to the adopted features. The capability of each model depends on many parameters such as the main model parameters and its values to be close to the better solutions and depends as well on the uploaded data, which is also differs when data changes. Studying these models carefully to be fitting with the proposed model is needed, in addition, the selection of parameters may change in many steps until reaching the required model that provided the best decision to the researcher or according to the related disorder. Therefore, before starting any classification models for the specific problem authors and researchers ought to reach the best path leading the way to the best solution. Tree classification and its modification get researchers interested in the medical, several studies appear to improve tree model in wide and verity problem domains [23]. In this paper, three classification techniques employed to predict a deficiency of important parameter in the body iron or called anemia: by applying tree types such as Hoeffding Tree or combining logistic regression with normal type of decision tree such in Logistic Model Tree or using recursive partitioning such in Random Tree. The proposed study conducted in two phases. The first phase determines the best of the three abovementioned methods in anemia prediction. A feature selection strategy is used in the second phase to discover the effect of attributes reduction in the prediction operation of anemia. The prediction techniques are listed in the following three subsections.

3.1 Hoeffding Tree

It has employed for a fixed dataset which is stable at time, which is applied as well for big data stream as an incremental type of the classical design of the decision tree rules [24]. As shown in Figures 1 and 2, Hoeffding rules are applied incrementally as a decision tree working with multiple steps and parts. After the previous steps have a sufficient calculation of clue that lead to the end of the first (previous) branch, then moving forward to expanding rules and branches one by one and according to the same rules of Decision Tree.

Its Operated by incrementally learning from data regardless storing the entire dataset by working with each incoming data point one at a time. The split process was done when the node is full of data using the Hoeffding Bound. When the bound is reached, the node was split according to the best features with threshold. The split process will be repeated when ever changing was occurred in the streaming dataset. The process is efficient, allowing the tree to learn in real time while handling large, continuous data streams.

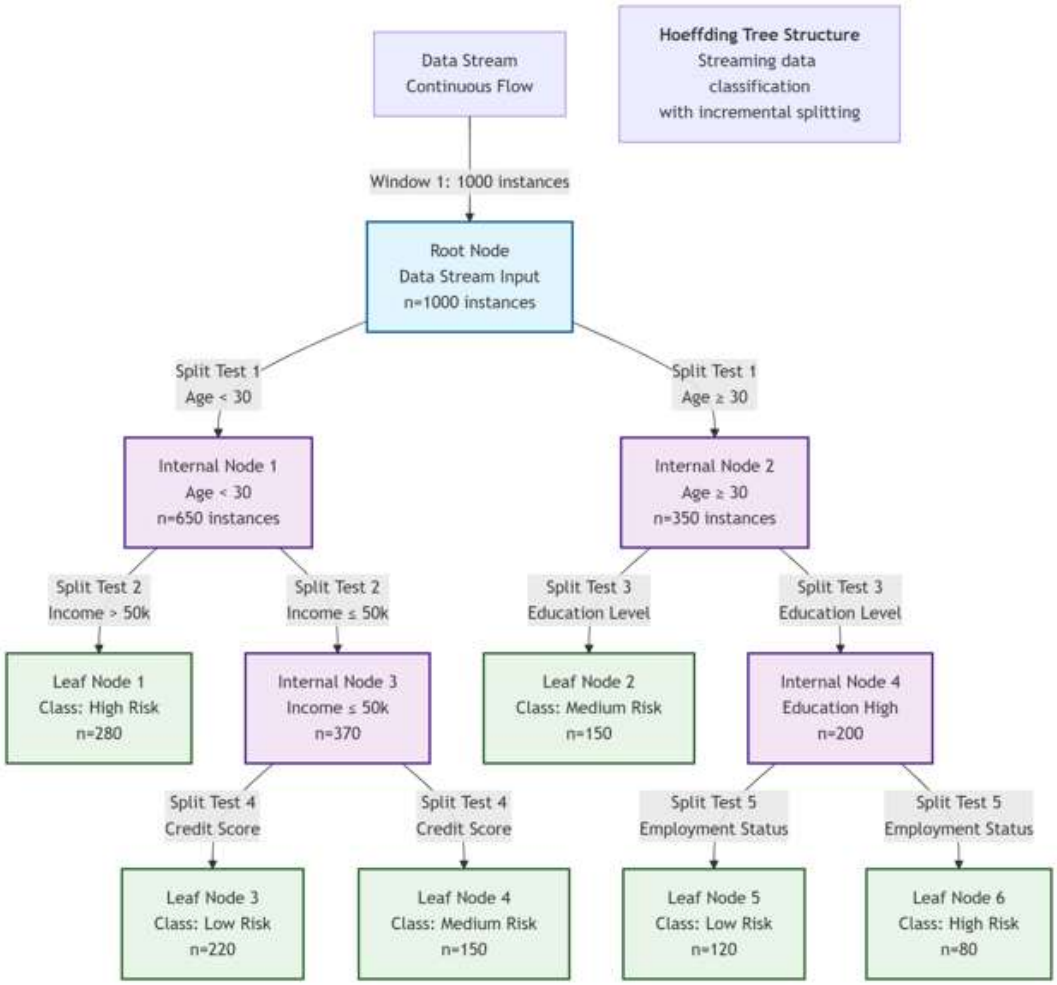


Figure 1: Hoeffding Tree first branches moving and splitting steps.

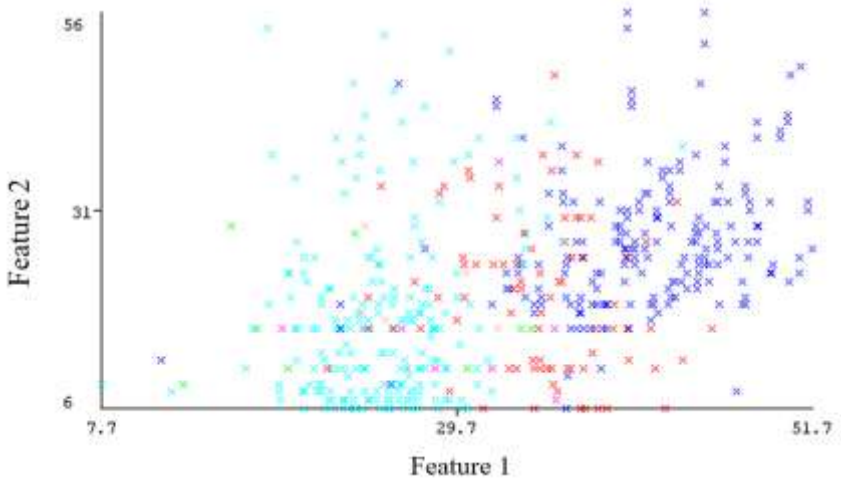


Figure 2: Hoeffding Tree based on two features (RBC and HB), where the colors dots are the different 6 Anemia classes.

3.2 Logistic Model Tree

As mentioned before, the use of linear regression model and applied it with decision tree algorithm which defined as Logistic Tree, such is explained in [25]. Figure 3 shows LMT which follows an idea which mixes between logistic regression and the conventional normal decision tree rules and leads to clearly class estimation rather than decision tree itself. These are the following calculations related to Anemia dataset and based on LMT.
Logistic model tree:

```
------------------
HB <= 11.1
    RBC <= 8.7
        MCH <= 30.2
          HB <= 9.7: LM_1:5/25 (228)
          HB > 9.7
               Age <= 17: LM_2:5/30 (23)
                Age > 17: LM_3:5/30 (15)
          MCH > 30.2: LM_4:5/20 (9)
        RBC > 8.7: LM_5:5/15 (54)
HB > 11.1: LM_6:5/10 (210)
Number of LMT leaves: 6
Size of the LMT: 11.
```

The LMT used a series of decision rules to classify data based on features as taken in this results like HB, RBC, MCH, and Age. The first split was a according to threshold value when HB <= 11.1 or RBC <= 8.7. each split in this case was lead to different branches until reaching the final leaf node which is called LM_1, LM_2 that represent final classification decisions. The case of LM_1:5/25 (228) means that 5 out of 25 instances reaching this leaf were classified correctly, while the total number of processed samples was 228 by this node only. The tree processing continued with new split to refine the prediction based on other features as well until providing the best results with higher performances.

3.3 Random Tree

Random tree is applied for data classifications by adding the regression tree as well. In this model, data is divided into subsets that have the same output class type of anemia (in this paper to 6 subsets or groups of data). Then checking these subsets according to the overall related attributes and define the noisy feature. These subsets are split into two repeatedly, until the constraints or the stopping conditions are satisfied. Training data was used to develop this tree model, then it randomly uses a data part to divide the tree again [26]. These are the following conditions and steps calculations of Random tree according to Anemia dataset as Figure 4 and Figure 5. The tree started by evaluating HB value as threshold condition, then according to the selected value of 6.3 the tree was divided into two branches and continues to split more based on different features, in this case was th AGE and MCV feature, with threshold values for AGE < 26 and MCV >= 66.5. this process will continue till stopping condition like maximum depth or even minimum size are reached. This technique selected different groups of features at each epoch, ensuring that the tree does not overfitting and remains adaptable to new data. The tree ultimately creates 6 subsets, each corresponding to a different class of anemia, with each leaf node representing a final classification based on the feature conditions.

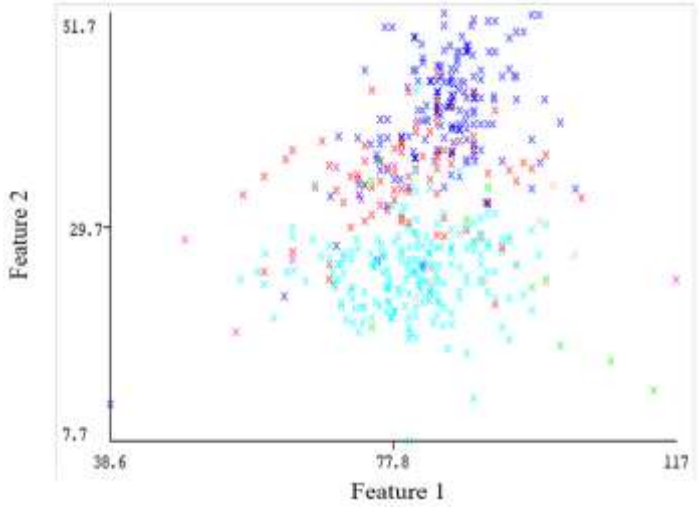


Figure 3: LMT results based on the Age and MCH, where the colors dots are the different 6 Anemia classes. .

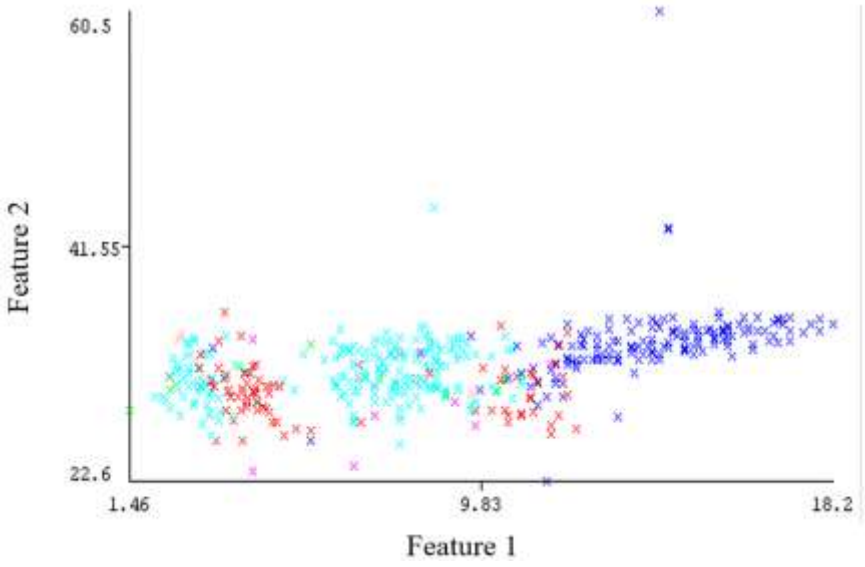


Figure 4: Random Tree results based on MCH and HB where the colors dots are the different 6 Anemia classes.

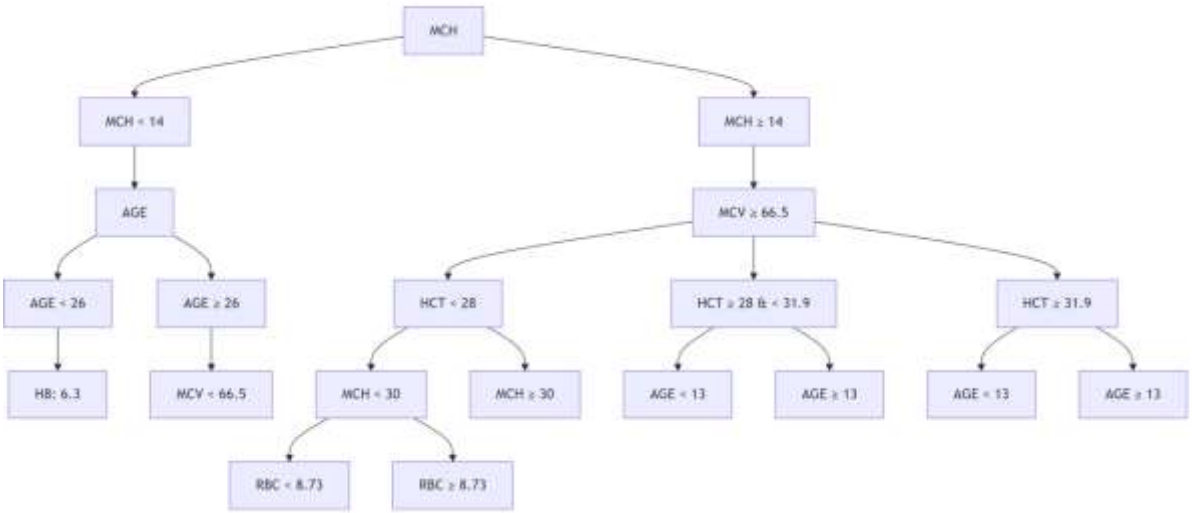


Figure 5: A part of the random tree which overall size is 151.

3.4 Related Measurements of the Methods

In order to satisfy the objective of the study we need to measure the proposed methods using suitable metrics. Most evaluation metric use for classification is accuracy [27]. Accuracy metric calculated and represented the rate between true prediction to overall prediction. Which is related to TPOS: positive classes predicted true, TNEG: negative classes predicted true, FPOS: positive classes predicted false, FNEG: negative classes predicted false. Mean absolute error is another evaluation metric used to measure models' performance, MAE evaluate the absolute error of the model and calculated to provide more decision accuracy. Finally, the process time to accomplish dataset classification for every model.

4 RESULTS AND DISCUSSION

The number of attributes applied before cancelling or deleting unnecessary features/attributes are 11 attributes. Which are listed as the blood tests with its parameters and the last feature are the classes of anemia. Each method applied is measured with 10-

fold cross validation and time taken is also considered as well. Correctly identified cases were done for Hoeffding Tree, LMT and Random Tree, which were done as well for the same 11 features.

Feature selection techniques have been applied like Classifier Attribute Evaluation, Classifier Subset Evaluation, and Gain Ratio to improve the overall system accuracy and deleted any undesired features. Also, Classifier Subset Evaluation to ensures that the only relevant features, when considered together, contribute the most to the classifier's ability to make accurate predictions. In addition, Gain Ratio was utilized as well to measure the efficiency of each attribute which it is effective for identifying features that provide substantial discriminative power. The less impact features that have been approved by these three techniques were Gender, Age, and WBC. As a result, these features have been deleted from the dataset in the second step of article works to enhance the performance. By focusing on the more relevant features, this selection process improved the accuracy of the predictive models, that ensuring only the most influential attributes have been chosen for classification.

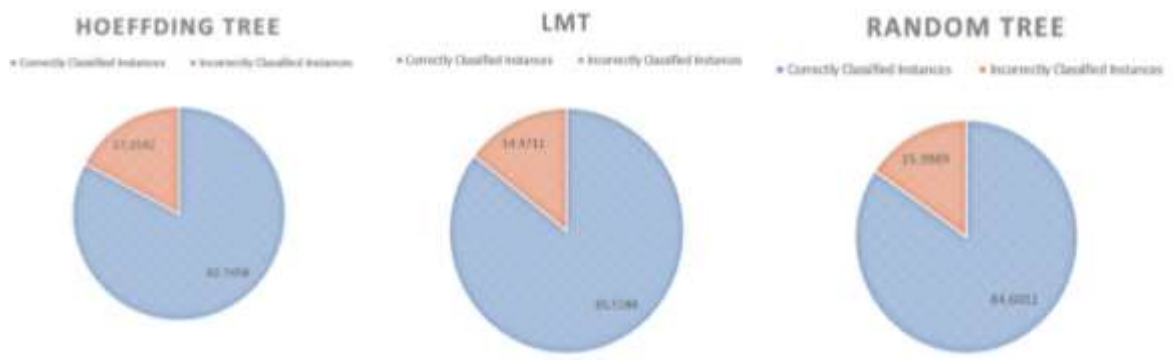
The results of three utilized techniques for feature process was shown in Table 1, demonstrating the stable and rank value for the applied features. The relative important with high rank features that have higher value than 0.5 was kept and other related features with less contribution was omitted from the dataset for the next procedure. The highest rank was for several parameters like HB, RBC and Serum Ferritin, showing their strong contribution in the results of prediction system. Other features have been moderate affection on the prediction performances like PLT, MCV, MCH, Serum Iron, and TS. However, attributes like Gender, Age, and WBC have very low impact on the prediction process especially for decision tree models. This stable and matching between all three selection techniques justifies their removal from the dataset, as these features introduced noise and redundancy which gradually changed the final model results. After deleting low effect features from Anemia dataset, not only one of the models was improved, overall techniques have been enhanced and provided more efficient.

Accuracy for all of the mentioned methods was shown in Table 2, which is done for 539 instant and before minimizing of patient's attributes. As cleared in Table 2, accuracy is measured for the overall models and has a promising result. But the LMT provided the best classification technique's results. Figure 6 shows the correct and incorrectly classified instances. While the time taken was 0.02 seconds for Random Tree technique which was the minimum value compared to elapsed time taken by other techniques. In addition, Hoeffding Tree was close to this value with 0.03 seconds and this due to that, these techniques are dealing with overall characteristics and features of 539 samples. While, the LMT has taken about 0.47 seconds, and this is due to the combination between ordinary tree decision and logistic regression for a noisy or a small amount of data are available.

After applied feature selection on 11 overall attributes, it has been specified that three of these features are not important to the overall measurement. These attributes are Gender, Age and WBC. This study is focus on these unusual collecting features which minimize the overall accuracy and increase time taken by some of applied techniques as shown in Table 3.

Table 1: Feature selection evaluation results for each attribute.

Feature	Rank Score of the Classifier Attribute Evaluation	Contribution of the Classifier Subset Evaluation (Contribution %)	The information gain of Gain Ratio	Impact on Anemia Dataset	Decision
HB	0.962	18.5%	22.7%	High Impact	Kept
RBC	0.945	16.8%	21.2%	High Impact	Kept
WBC	0.452	5.4%	6.1%	Low Impact	Canceled
PLT	0.833	12.1%	10.3%	Moderate Impact	Kept
MCV	0.885	10.7%	11.8%	Moderate Impact	Kept
MCH	0.904	11.5%	12.4%	Moderate Impact	Kept
Serum Ferritin	0.939	12.9%	13.7%	High Impact	Kept
TS	0.911	10.8%	11.6%	Moderate Impact	Kept
Serum Iron	0.928	10.2%	10.9%	Moderate Impact	Kept
Age	0.388	4.6%	5.3%	Low Impact	Canceled
Gender	0.341	3.5%	4.1%	Low Impact	Canceled



(a) (b) (c)

Figure 6: The correctly and incorrectly classified instances for a) hoeffding tree, b) LMT and, c) random tree.

Figure 7: Overall accuracy and mean absolute error for LMT.

Table 2: Overall accuracy of applied techniques without minimizing features.

Method	Accuracy (%)	Mean Absolute Error
Hoeffding Tree	82.7458	0.0651
LMT	85.5288	0.0707
Random Tree	84.6011	0.0513

Table 3: Overall accuracy of applied techniques by minimizing features.

Method	Accuracy (%)	Mean Absolute Error
Hoeffding Tree	94.4156	0.0624
LMT	96.2709	0.0699
Random Tree	91.2616	0.0625

In Table 3, results show that Hoeffding Tree and LMT have the most impact with features minimizations. In addition, the mean absolute error has been minimized according to attributes number decreasing as shown in Figure 7. These attributes have been proved that there is no need to collect more data than required for each patient. Data should be calculated and measured before collecting from patients, which provided minimum time for the overall process and makes applied techniques more efficient with minimum data. The overall time taken by these applied techniques is about 0.01 seconds, while 0.03 seconds was for LMT technique. This led the researchers to make such a study which minimizes cost and time for the overall detection and classification process. While random tree shows that a larger amount of data is required to identify each sample and patient.

5 CONCLUSIONS

In this study, the proposed tree-based classification models achieved with a high efficiency the objective of anemia type classification based on the utilized dataset of 539 samples and 10 features. Three decision tree algorithms were applied Hoeffding Tree, Logistic Model Tree, and the Random Tree, that divides the dataset into sets for classification. First the techniques have been applied without deleting any attributes with performances evaluations. The same techniques have been applied after feature selection

was done and then some of the features were deleted. The deleted attributes were Gender, Age, and WBC have less effected to model performance and due to this effectiveness, these features were deleted. Through the utilized model, LMT achieved the highest accuracy which has been increased with 9% after deleting the less effected features from 85.5% to 96.27%. These results specified that careful preprocessing of such a clinical dataset enhance the performances of Machine Learning. The study proved that the importance of identifying relevant features to support accurate automated classification was one of the major thoughts for such a work. In this paper, the obtained accuracy for the proposed model met the objective of anemia classification for the six types based on tree classification models. According to the calculations, these parameters add some noises to the uploaded dataset that made the predication path wrong based on the applied methods. This study approves that selecting the suitable dataset with a suitable attribute is important as selecting the main parameters of the method.

The feature selection also presented the results that revealed the critical ones for predicting anemia depending on their three-evaluation metrics. Among the utilized features, the HB provided the most impact with highest rank score of about 0.962. while, the RBC count also showed a high impact with a rank score of about 0.945. Serum Ferritin follows closely, reflecting its role in assessing iron stores, particularly for iron-deficiency anemia, with a rank score of about 0.939. Other features, like PLT, MCV, MCH, TS, and Serum Iron, also contribute moderately to the prediction of anemia, but were not like the HB or RBC in their overall effect. These moderate impact features provided valuable amplitude which used to smooth the refine classification especially for six types of anemia classifications. In the other hands, features such as WBC count, Age, and Gender were provided low impact and omitted from the model. WBC's role in anemia was typically indirect, and age and gender alone had not strongly predicted anemia status in this usage which lead to their removal. This proposed approach enhanced model efficiency by focusing on the most relevant variables.

REFERENCES

[1] S. Kilicarslan, M. Celik, and Ş. Sahin, "Hybrid models based on genetic algorithm and deep learning algorithms for nutritional anemia disease classification," Biomed. Signal Process. Control, vol. 63, p. 102231, 2021.

[2] R. B., B. Kabir, D. Mamta, D. Namrata, R. Sarita, K. Ajay, and R. Y. K., "Analysis and investigation of fuzzy expert system for predicting the child anemia," Mater. Today: Proc., vol. 56, pt. 1, pp. 231-236, 2022.

[3] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," PLoS One, vol. 17, no. 7, 2022.

[4] P. V. and V. C., "Machine learning algorithms for anemia disease prediction - a review," Int. Res. J. Mod. Eng. Technol. Sci., vol. 4, no. 4, 2022.

[5] S. Mohammed, A. Abbas, A. Ahmad, M. Mohammed, M. Sarı, and H. Uslu Tuna, "Data mining technique's parameters definition and its prediction effect's based on iron deficiency dataset," Sigma J. Eng. Nat. Sci., vol. 43, no. 2, 2025.

[6] C. Li and D. C. Coster, "Improved particle swarm optimization algorithms for optimal designs with various decision criteria," Mathematics, vol. 10, no. 13, p. 2310, 2022.

[7] N. Q. Sultan and M. S. Siti, "Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis," Appl. Soft Comput., vol. 11, no. 1, pp. 1427-1438, 2011.

[8] N. Sharma, V. Khullar, and A. Luhach, "Comparative study of back-propagation and PSO based back-propagation for anemia diagnosis in pregnant ladies," Int. J. Sci. Eng. Comput. Technol., vol. 7, no. 1, pp. 1-5, 2017.

[9] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J. M. Ginoux, "Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm," Biocybern. Biomed. Eng., vol. 38, no. 2, pp. 362-372, 2018.

[10] V. Laengsri, W. Shoombuatong, W. Adirojananon, C. Nantasenamart, V. Prachayasittikul, and P. Nuchnoi, "ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, p. 212, 2019.

[11] T. Qadah and A. Munshi, "Synthesis and prediction of anemia from multi-data attribute co-existence," IEEE Access, 2024.

[12] A. M. El-Boghdady, S. Kishk, M. M. Ashour, and E. AbdElhalim, "Machine-learning based stacked ensemble model for accurate multi classification of CBC anemia," Mansoura Eng. J., vol. 49, no. 3, p. 4, 2023.

[13] M. Ramzan, J. Sheng, M. U. Saeed, B. Wang, and F. Z. Duraihem, "Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms," Vis. Comput. Ind. Biomed. Art, vol. 7, no. 1, p. 18, 2024.

[14] L. J. Marcos-Zambrano et al., "Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment," Front. Microbiol., vol. 12, p. 634511, 2021.

[15] E. Elbasi and A. I. Zreikat, "Heart disease classification for early diagnosis based on adaptive Hoeffding tree algorithm in IoMT data," Int. Arab J. Inf. Technol., vol. 20, no. 1, pp. 38-48, 2023.

[16] S. A. Fayaz, M. Zaman, and M. A. Butt, "An application of logistic model tree (LMT) algorithm to ameliorate prediction accuracy of meteorological data," Int. J. Adv. Technol. Eng. Explor., vol. 8, no. 84, pp. 1424-1440, 2021.

[17] V. Babenko, I. Nastenko, V. Pavlov, O. Horodetska, I. Dykan, B. Tarasiuk, and V. Lazoryshinets, "Classification of pathologies on medical images using the algorithm of random forest of optimal-complexity trees," Cybern. Syst. Anal., vol. 59, no. 2, pp. 346-358, 2023.

[18] S. J. M. Sahar, A. A. Arshed, and S. M. Mohammed, "Anemia prediction based on rule classification," in Proc. 13th Int. Conf. Developments in eSystems Engineering (DeSE), Liverpool, United Kingdom, pp. 427-431, 2020, doi: 10.1109/DeSE51703.2020.9450234.

[19] C. Wu et al., "Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease pneumonia in Wuhan, China," JAMA Intern. Med., vol. 180, no. 7, pp. 934-943, 2020.

[20] N. Friis-Møller et al., "Cardiovascular disease risk factors in HIV patients–association with antiretroviral therapy. Results from the DAD study," AIDS, vol. 17, no. 8, pp. 1179-1193, 2003.

[21] K. M. West et al., "The role of circulating glucose and triglyceride concentrations and their interactions with other risk factors as determinants of arterial disease in nine diabetic population samples from the WHO multinational study," Diabetes Care, vol. 6, no. 4, pp. 361-369, 1983.

[22] D. Kocev, M. Ceci, and T. Stepišnik, "Ensembles of extremely randomized predictive clustering trees for predicting structured outputs," Mach. Learn., vol. 109, pp. 2213-2241, 2020.

[23] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," Water, vol. 11, no. 5, p. 910, 2019.

[24] A. Kumar, P. Kaur, and P. Sharma, "A survey on Hoeffding tree stream data classification algorithms," CPUH-Res. J., vol. 1, no. 2, pp. 28-32, 2015.

[25] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," Mach. Learn., vol. 95, no. 1-2, pp. 161-205, 2015.

[26] A. K. S. and L. Jaya, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy," Prot. Control Mod. Power Syst., vol. 3, p. 29, 2018.

[27] S. J. M. Sahar and S. M. Mohammed, "COVID-19 risk factors specification using decision tree based on the degree of redundancy between features," in Proc. IEEE 3rd Global Conf. Advancement in Technology (GCAT), Bangalore, India, pp. 1-11, 2022, doi: 10.1109/GCAT55367.2022.9971950.