Decision Rules Generation Using Decision Tree Classifier and Their Optimization for Anemia Classification

Rajan Vohra, Anil Kumar Dudyala, Jankisharan Pahareeya, and Abir Hussain

Abstract Anemia disease is one of the prevalent diseases observed across women and children in most of the developing countries. This is caused due to the iron deficiency in human body. Detecting this disease at the early stage can help the medical fraternity to prescribe proper medicines and come up with alternate solutions that can prolong the patient's initial stage before it enters into critical stage. Due to the non-availability of historical data of the Anemia patients, there is very sparse literature that addresses the problem of detection of this disease. In this paper, a real-time Anemia dataset pertaining to Indian context is considered and due to the imbalance nature of the dataset, SMOTE is employed for balancing. With the help of decision tree rule-based learning method, rules for detecting the Anemia are derived using two datasets original and SMOTE. The efficacy of the rules is evaluated, and reduced rules are selected based on their individual classifying accuracy. In a quest to give a simple human understandable and optimal rule which can be used by any medical fraternity for detecting the presence of Anemia at different stages, we tried to propose the reduced rules-based method which may come handy. The efficacy of the rules is promising and helps in identifying the presence of Anemia at early stage.

Keywords Decision tree (DT) · Reduced rules · Anemia detection · Data balancing

R. Vohra (✉) · A. Hussain
Department of Computer Science, Liverpool John Moores University, Liverpool, UK
e-mail: R.vohra@ljmu.ac.uk

A. Hussain
e-mail: A.Hussain@ljmu.ac.uk

A. K. Dudyala
Department of Computer Science, National Institute of Technology Patna (NIT Patna), Patna, India
e-mail: ak@nitp.ac.in

J. Pahareeya
Department of Information Technology, Rustamji Institute of Technology, BSF Academy, Tekanpur, Gwalior, India
e-mail: jankisharan@rjit.org

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_53

1 Introduction

Human body is composed of several proteins and amino acids. The sustenance of this body is carried out with the help of minerals and vitamins. Deficiency of any of these essential minerals or vitamins will cause either malfunction of the human body or will lead to disease. One such deficiency of iron in human body will lead to a disease called Anemia. Anemia can be termed as deficiency of hemoglobin caused due to shortage of iron. Anemia is found to be prevalent among the developing countries and most popularly among women and children compared to men of these countries. Globally it is observed as one of the critical health problems.

Identifying the Anemia at the early stage so that it can be prolonged from further deteriorating into advanced stages is one of the most challenging issues. This is due to the non-availability of the data in real-time scenario. It is observed in the literature very few works that have explored this issue of detecting the anemia [1–5]. Hence, we have considered the Indian dataset which had been collected from Eureka diagnostic center, Lucknow, India [6] for the experimental purpose and with the help of Decision Tree model a set of rules has been generated that helps in detecting the anemia at the early stages.

The remaining sections are arranged as follows, Sect. 2 describes related work. Proposed method is discussed in Sect. 3. Section 4 deals with the data description and algorithm used. Results and discussion are elaborated in Sect. 5, while Sect. 6 concludes the work.

2 Related Work

Anemia that is caused by the deficiency of Iron is one of the most critical health problems globally and is a serious public health issue [7]. According to the World Health Organization (WHO), Anemia prevalence of over 40% in a community makes it a public health issue of critical importance [8]. While Anemia prevalence in children can be caused due to genetic reasons or due to deficiencies in nutrition like deficiencies in iron or folate or vitamins A/B12 and copper, iron deficiency is the most important determinant of anemia [9]. Socio demographic characteristics of mothers, households such as region, wealth index, water sources, working status and anemia status along with child features like age, nutritional status, child size at birth are the most critical features influencing anemia in the age group of 6–59 months in children [1]. According to WHO, Anemia prevalence occurs in most of the countries in Africa and South Asia and some countries in East Asia and the Pacific. While the highest prevalence of anemia is found in Africa, the largest numbers of children affected by anemia are found in Asia [2]. Many machine learning models are increasingly used in the analysis and prediction of diseases in the healthcare [10]. Most of studies indicated that machine learning techniques such as support vector machines (SVM), Random Forest and artificial neural networks (ANN) have been applied for

the classification of different diseases such as Diabetes [11–13], Appendicitis [14], and multiple sclerosis [15]. Machine learning techniques to classify anemia in children are still evolving. Along with traditional clinical practices, machine learning techniques can be utilized to predict the risk of anemia prevalence in children. Some key research in this direction has been undertaken as demonstrated in [3, 16], which have constructed prediction models for anemia status in children. The prevalence of anemia among adults was studied by taking complete blood count (CBC) at a referral hospital in southern Ethiopia. Prevalence and severity were related with age and gender and were analyzed [17]. Social factors such as income, wealth, education can affect health markers in people such as blood pressure, body mass index (BMI), and waist size, etc. [18]. Sow et al. used support vector machines (SVM) and demographic health survey data from Senegal to classify malaria and anemia status accurately [4, 19]. Using feature selection, the number of features of both anemia and malaria datasets were reduced. Using variable importance in projection (VIP) scores, the relative importance of social determinants for both anemia and malaria prevalence were computed. Finally, machine learning algorithms were utilized for the classification of both anemia and malaria–Artificial neural networks (ANN), K nearest neighbors (KNN), Random Forests, Naïve Bayes and support vector machines (SVM) were used [20]. Lisboa has demonstrated the utility and potential of Artificial Neural Networks (ANN) in health care interventions [5]. Using CBC samples, a study to classify anemia using Random Forests, C4.5 (Decision tree), and Naïve Bayes (NB) was undertaken. Comparison of the classifying algorithms using mean absolute error (MAE) and classifier accuracy were computed and tabulated in [21]. Some of the research also applied the Naïve Bayes Classifier and entropy classifier for the purpose of classification [6].

Almugren et al. in 2018 conducted a study using the anemia dataset and investigated how Artificial neural networks (ANN), Naïve Bayes (NB), C4.5 and Jrip data mining algorithms can be used to classify instances in the given dataset as being anemic or normal—that is a binary classification problem. In this study, the performance of these algorithms was benchmarked for a comparative analysis, and it was found that ANN and Jrip algorithms were the best performing algorithms in this regard [22]. In a study, Jatoi et al. used data mining methods on complete blood count (CBC) dataset of 400 patients for detecting the presence of anemia. It was found that Naïve Bayes (NB) algorithm had 98% accuracy in predicting the presence of the disease correctly [23]. In the study conducted in 2019, Meena et al. have used Decision tree algorithms to perform classification on an input dataset representing children for the diagnosis of anemia in the given dataset. They also identified the significant features driving the prevalence of anemia in reference to the feeding practices adopted for infant feeding [24]. Ching Chin Chern et al. have used Decision Tree Classifier models to acquire decision rules for classifying eligibility of Taiwanese citizens to be suitable recipients of tele health services. Involvement of a physician, social worker and health care managers was done to ensure a thorough process and J48 algorithm and logistic regression techniques were used to generate the decision trees representing the decision rules generated [25]. A study done by Lakhsmi K.S et al. has used Association rule mining on medical records to

extract decision rules of the type symptom disease. In this computation well-known Association rule mining algorithms like A Priori and FP Growth have been used to derive the decision rules [26]. Song Yan et al. have studied how decision trees can be used to generate decision rules for various medical conditions. In this paper, they have constructed a Decision tree model representing decision rules for the classification and diagnosis of Major Depressive disorder (MDD) [27]. In a study done by Yildiz et al. on a health care dataset obtained from a hospital in Turkey, the authors have used four classification algorithms–artificial neural networks (ANN), support vector machines (SVM), Naïve Bayes (NB) and Ensemble Decision trees to perform classification for various types of anemia and the performance of the algorithms is benchmarked in which Bagged Decision Trees was the best performing algorithm [28]. Heru Mardiansyah et al. have studied the problem of imbalanced datasets and how this can be resolved by using SMOTE techniques to balance the original dataset. In their study, they have selected four datasets from the UCI machine learning repository–German credit cards, Winconsi, Glass and Ecoli to show the application of SMOTE techniques on the given datasets and the resulting datasets arising from this computation [29].

Kilicarslan et al. have constructed two hybrid models using genetic algorithms and Deep learning algorithms of stacked autoencoder (SAE) and convolutional neural networks (CNN) for the prediction and classification of iron deficiency anemia and benchmarked the performance of these two hybrid models in the classification computation for iron deficiency anemia [30].

Although several clinical different machine learning algorithms have been proposed that incorporate several data mining techniques for Anemia prediction, none of them had come up with a set of rules, which come handy in identifying the Anemia existing at different stage. Our proposed methodology is attempting to cover this gap by proposing optimal number of rules.

3 Proposed Method

This work provides an understandable set of rules which can be used for detecting Anemia at early stages using optimized rules extracted from Decision Tree Model using Anemia dataset. As there is unequal ratio of samples of each class, we adopted Synthetic Minority Oversampling Technique (SMOTE) for balancing the minority class. The balanced (SMOTE) dataset is also used with the Decision Tree Classifier for extracting rules. Thus, obtained rules are further optimized and evaluated for their efficacy and strength in classifying the Anemia dataset.

The complete description of the data is given in Sect. 4 and the complete features of the dataset are shown in Table 1.

The block diagram of the proposed model can be seen below in Fig. 1. The original dataset is partitioned into training and testing parts using stratified sampling with the ratio of seventy for training and thirty for testing. The SMOTE is applied only on training dataset to obtain the balanced (SMOTE) dataset. Thus, Decision tree

Table 1 Anemia dataset description

S. No.	Attributes names	Type of attribute	Abbreviation
1	Age	Numerical	Age
2	Gender	Character	Gender
3	Hemoglobin	Numerical	HGB
4	Mean cell volume	Numerical	MCV
5	Mean cell hemoglobin	Numerical	MCH
6	Mean cell hemoglobin concentration	Numerical	MCHC
7	Red cell distribution width	Numerical	RDW
8	Red blood cell count	Numerical	RBC
9	White blood cell count	Numerical	WBC
10	Platelet count	Numerical	PLT
11	Packed cell volume	Numerical	PCV

is trained on these two different training datasets (i.e., Actual and Balanced dataset) separately. The complete description of the proposed model can be seen in the next sub-section.

3.1 Description of the Proposed Model

The Decision Tree Classifier is trained using Actual Training dataset and Balanced (SMOTE) dataset separately. The model obtained after training, is tested with the testing data and the results obtained are noted as testing results. The rules from the Decision Tree model are also extracted separately for each model. It has been observed that 232 rules were generated by the Decision Tree Classifier when it has been trained using Actual Training dataset. On the other hand, it has given only 26 rules when Decision Tree Classifier is trained using SMOTE dataset.

Since 26 rules obtained from SMOTE dataset were giving significantly good results compared to results using Actual dataset we have considered these rules to take further for evaluating their efficacy and strengths and also to reduce the rules further.

These rules were coded individually and evaluated using the Anemia dataset for their relevancy and efficacy in terms of detecting the Anemia and they are sorted in descending order of their accuracies for each class. Next, two top yielding rules for each class are selected as reduced rules. These rules are again coded and are evaluated using the Testing dataset and the final results are obtained in the form of performance metrics defined.

The description of the proposed model can be defined using the following algorithm.

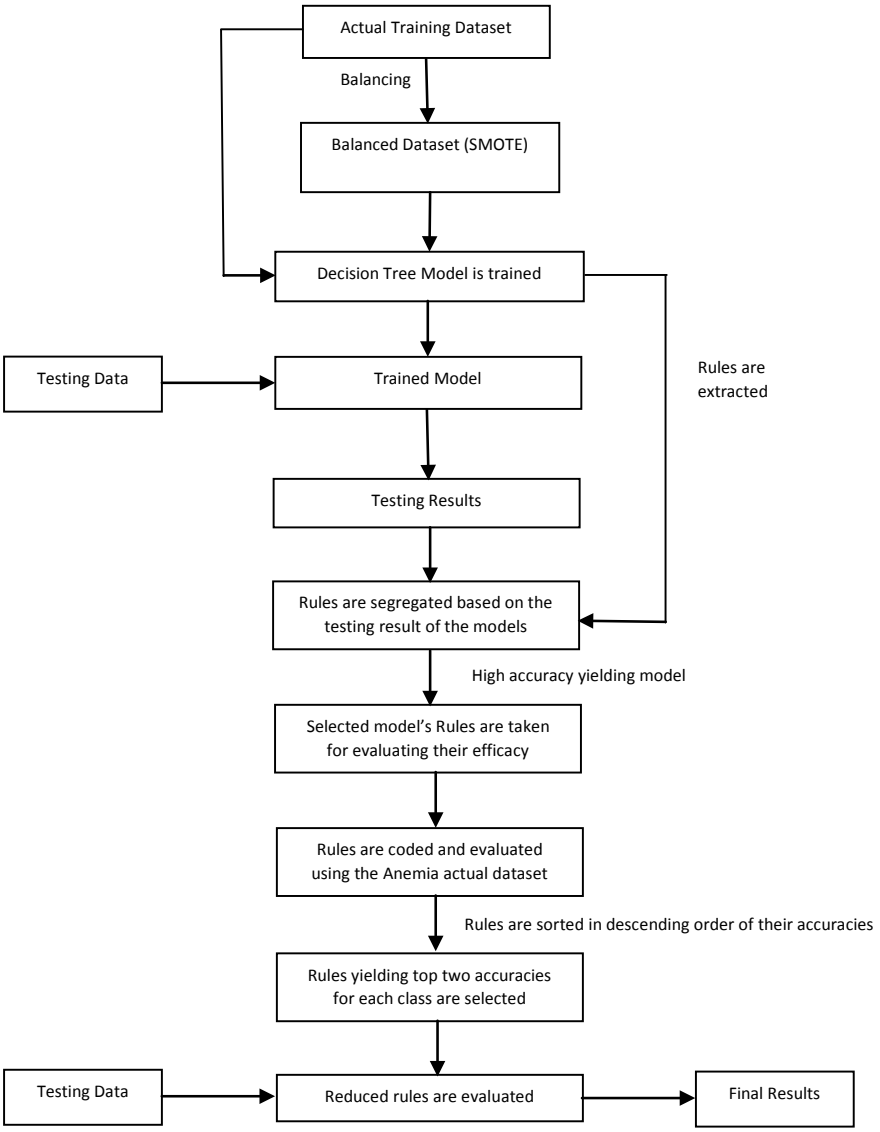

Fig. 1 Block diagram of the proposed model

Input: Train data, Test data
Output: decision rules
Pre-processing steps:
1. Data pre-processing is done (cleaning, removing of redundancy, etc.,)
2. Data is partitioned into training and testing sets with 70:30 ratio
3. Training data is balanced using SMOTE technique.

Algorithm steps:
1) Start
2) For each training dataset
3) Apply Decision Tree classifier
4) Perform hyper parameter tuning, check accuracy
5) If accuracy is high enough (saturated)
6) Extract rules from the model
break.
7) Else repeat step 3
8) End for
9) Test the trained model
10) Obtain the performance metrics (Recall, Precision and Accuracy)
11) Select the model with the best results
12) Code the extracted rules of the selected model
13) Evaluate the efficacy of the rules
14) Select the top two performing rules(reduced rules) for each class
15) Code the reduced rules
16) Evaluate the rules using Actual data
17) Obtain the performance metrics (Recall, Precision and Accuracy)
18) Stop

4 Dataset Description

The data used for the experiment was collected for the period of September 2020 to December 2020in the form of CBC test reports from the Eureka diagnostic center, Lucknow, India [31]. Data was collected with ethical clearance from the diagnostic center and patient consent was obtained.

The Anemia can be classified into three different types based on their severity level. They are

- Mild,
- Moderate,
- Severe.

The Data consists of eleven attributes as illustrated in Table 1 with the size of the dataset being 364 records. The class variable is named as Hemoglobin (HGB) which has three classes, namely **Mild**, **Moderate** and **Severe**. These three classes have been defined using the range of values, where the Mild range lies between 11.0 and 12.0, Moderate range lies between 8.0 and 11.0 and Severe values lies less than 8.0. The distribution of the three classes in the dataset is 70.32% of Mild, 25.28% of Moderate and 4.40% of Severe.

1	Age	Sex	RBC	PCV	MCV	MCH	MCHC	RDW	TLC	PLT /mm³	HGB
2	28	0	5.66	34	60.1	17	28.2	20	11.1	128.3	1
3	41	0	4.78	44.5	93.1	28.9	31	13	7.02	419	0
4	40	1	4.65	41.6	89.5	28.8	32.2	13	8.09	325	0
5	76	0	4.24	36.7	86.6	26.7	30.8	14.9	13.41	264	0
6	20	1	4.14	36.9	89.1	27.8	31.2	13.2	4.75	196	0
7	24	0	4.29	40.1	93.5	29.6	31.7	14.5	13.96	233	0
8	28	1	4.98	42.3	84.9	24.9	29.3	16.2	9.33	213	0
9	14	0	4.97	43.8	88.1	28	31.7	15.2	3.92	229	0
10	16	0	4.16	38.7	93	28.8	31	17.9	5.77	211	0
11	62	0	5.25	45.6	86.9	25.3	29.2	15.6	10.68	151	0
12	42	0	2.17	28.3	93.5	28.1	30	24.6	3.46	92	2
13	28	0	4.81	44.4	92.3	27.9	30.2	14.3	6.22	150	0
14	59	0	3.41	32.9	96.5	29.9	31	16.8	6.62	132	1
15	28	1	2.26	26.9	119	41.2	34.6	15.6	5.27	222	1

Fig. 2 Snapshot of the original dataset

As the ratio of the three classes were not balanced, we adopted a SMOTE balancing technique for ensuring proper balancing among all the classes of the dataset, so that the machine learning technique used for training would not get biased.

The snapshot of the dataset can be seen below in Fig. 2.

4.1 Algorithm Used (Decision Tree Classifier)

Decision Tree Classifier is a rule-based classifier which works on the basis of entropy. It uses different criterion functions like Gini Index and Information Gain for splitting the given data into one of the classes [32]. It can be clearly represented using a hierarchical tree-based diagram, where the classes are represented at the leaf level and the splitting features are represented at the interior nodes. This algorithm is mostly suitable for the decision-making problems where it mostly classifies the given problem into different classes more accurately. A snapshot of the Decision tree can be seen in Fig. 3.

The implementation of the Decision Tree Classifier was done in python using the sklearn library. Gini Index was taken as the criterion function.

5 Results and Discussion

All the experiments were conducted using hold-out method. The training data was taken a 70%, and testing data was taken as 30%. The splitting was done using stratified random sampling. Results of three models, two Decision Trees with Actual and SMOTED dataset and one proposed reduced rule-based method are evaluated and are presented in Table 2.

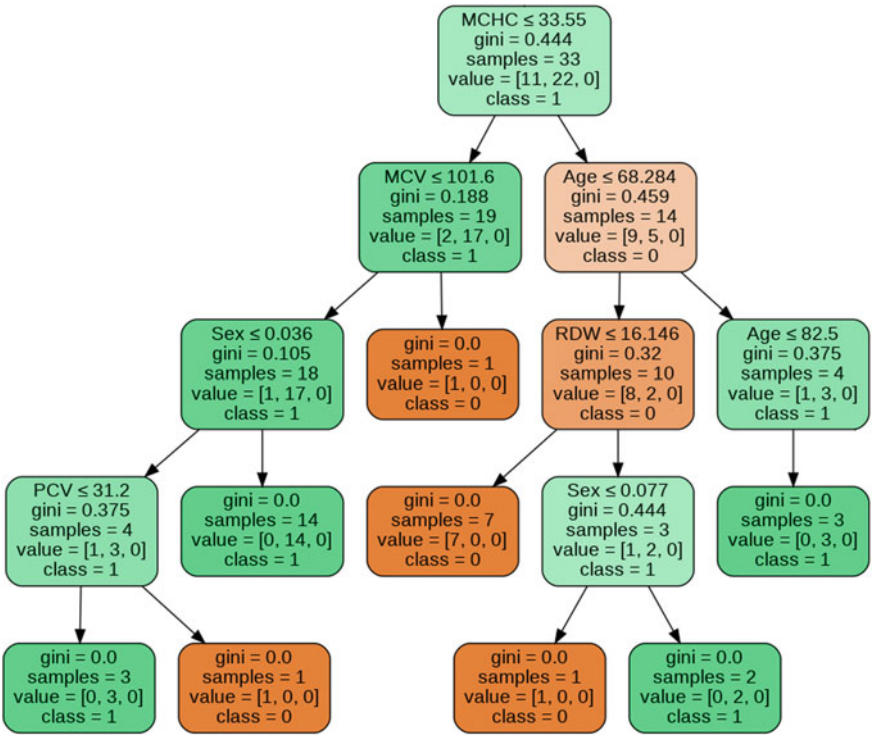

Fig. 3 Example of a decision tree

Table 2 Results of the decision tree using different datasets and reduced rules-based method

Metric	Actual dataset (decision tree)	SMOTED dataset (decision tree)	Coded reduced rules-based method
Accuracy	84.72	93.1	**92.03**
Recall			
Mild class	90.2	94.0	**96.48**
Moderate class	72.2	88.0	**78.26**
Severe class	66.7	95.0	**100**
Precision			
Mild class	97.9	95.0	**95.54**
Moderate class	72.2	88.0	**78.26**
Severe class	28.6	95.0	**100**

Table 3 Efficacy of the reduced rules

S. No	Rule number	Class	Efficacy
1	Rule 3	Severe	100
2	Rule 4	Severe	87.5
3	Rule 10	Moderate	68.47
4	Rule 15	Moderate	47.82
5	Rule 17	Mild	74.60
6	Rule 19	Mild	92.96

Accuracy, Recall and Precision have been used as performance metrics for measuring the efficacy of the results. These three can be defined as follows

- Accuracy can be defined as the ratio of total number of patients correctly classified irrespective of class to the total number of patients.

$$\text{Accuracy} = \frac{\text{Total number of correctly predicted samples(TP + FN)}}{\text{Total number of samples(TP + TN + FN + TF)}}$$

- Recall can be defined as the ratio of the number of patients correctly classified as Anemia class to the total number of Anemia class patients present in the dataset

$$\text{Recall} = \frac{\text{number of correctly predicted samples of a class(TP)}}{\text{Total number of samples in that class(TP + FN)}}$$

- Precision can be defined as the ratio of the number of patients correctly classified as Anemia class to the total number of patients classified as Anemia class.

$$\text{Precision} = \frac{\text{number of correctly predicted samples of a class(TP)}}{\text{Total number of samples classified as class(TP + FP)}}$$

The rules extracted using Actual dataset were 232 which is actually a huge number, and it would be difficult for the end user to interpret or use these many rules. Moreover, the results using these rules were not so significant. This is due to the imbalance nature of the dataset. Hence, SMOTE technique has been employed to balance the dataset. Using the balanced dataset with the Decision Tree Classifier, we obtained 26 rules which were more concise than earlier method and were also giving improving results. The rules using the balanced dataset are shown in Appendix A. Though the rules were concise, their efficacy was not so promising, and they were containing some not so important rules. Due to this, it would be difficult for the end user to use these set of rules handy.

So, in order to evaluate the efficacy of the rules, the 26 rules were coded and their efficacy in detecting the Anemia were computed by applying these rules on the Actual dataset. Then, the rules were sorted in descending order of their efficacy of their respective classes. Top two rules from each class are extracted as

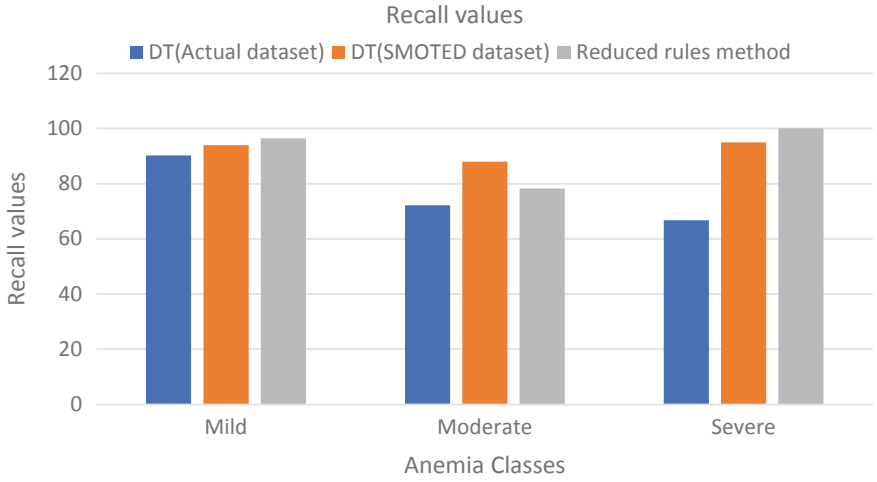

Fig. 4 Recall values for the three classes using decision tree and reduced rules-based method

reduced rules whose efficacies are shown below in Table 3. The reduced rules are presented in Appendix B.

Efficacy is computed using the following formula,

$$\text{Efficacy} = \frac{\text{number of correctly classified patients of a given class by a specific rule}}{\text{Total number of patients in that class}}$$

Next, these reduced rules were also coded and their performance metrics like Recall, Precision and Accuracy were also computed on the Actual dataset which are shown in Figs. 4, 5 and 6.

It can be observed from Table 2 that the Decision Tree Classifier using the Actual dataset has given the Accuracy, Recall and Precision values which are very low. This is due to the imbalance nature of the dataset. Whereas, in the case of Decision Tree Classifier using the SMOTED dataset, it can be observed that all the three metrics have improved compared to the results of the Actual dataset. While in the case of proposed Reduced rules-based method, it is evident that due to the elimination of not so important rules the recall value for the Anemia dataset has improved significantly in the case of Mild class and Severe class. Though there is slight reduction of 1% in the accuracy of Reduced rules-based method compared to Decision Tree Classifier using SMOTE dataset, it might be noted that the rules were reduced upto 77% which is a promising and significant contribution of this work. Having minimum rules which can be used for detection of Anemia would be an important tool for the end user to use.

Notations used in the following figures are,

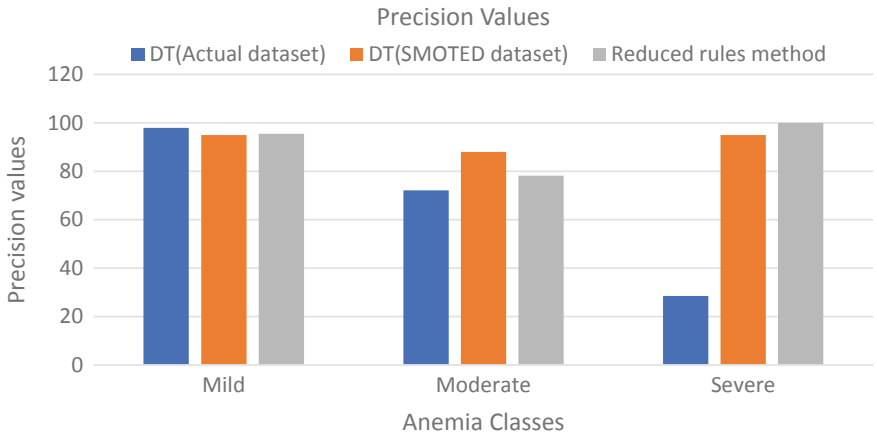

Fig. 5 Precision values for the three classes using decision tree and reduced rules-based method

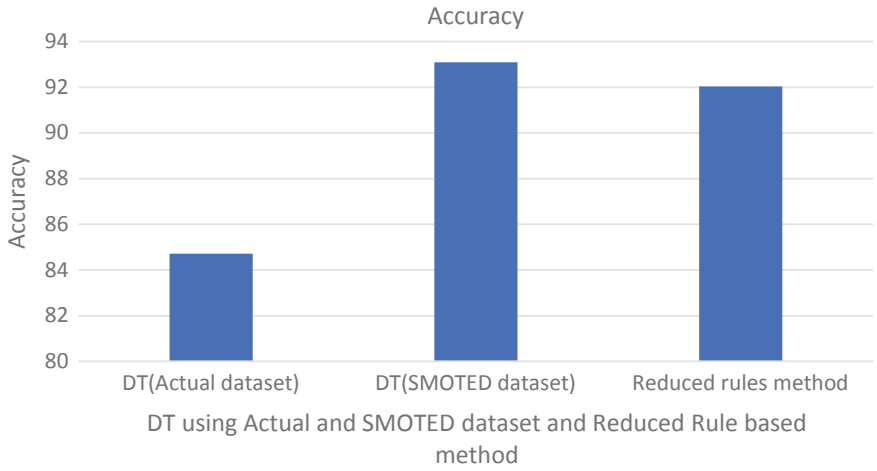

Fig. 6 Accuracy values for the three classes using decision tree and reduced rules-based method

- DT (Actual dataset): Decision Tree using the Actual dataset
- DT (Actual dataset): Decision Tree using the SMOTE dataset
- Reduced rules method: Reduced rule-based method on Actual dataset

Figure 4 shows the recall values of the three classes using the three different methods, two Decision Tree with Actual and SMOTED dataset and one with proposed Reduced rules-based method. It can be seen that the Reduced rules-based method has given good recall values of 96.48 and 100% compared to other methods in the case of Mild and Severe class. Whereas in the case of Moderate class, the Decision Tree Classifier with reduced (SMOTED) dataset has given good recall value of 88%.

Figure 5 shows the results of the precision values of the three different methods. It can be observed that in the case of Decision Tree Classifier using the Actual dataset, the precision values have been decreasing across different classes. While in the case of the Decision Tree Classifier using the SMOTED dataset and coded Reduced rules method the precision values have been slightly differing across all the classes. The precision value in case of Mild is topped by the Actual dataset, whereas in the case of Severe class the proposed Reduced rules-based method has given high result. While in case of Moderate class the Decision Tree Classifier using SMOTED dataset has topped.

As far as Accuracy is concerned, Fig. 6 shows that the accuracy has been improved in case of the Decision Tree Classifier using SMOTE dataset compared to Actual dataset and reduced rule dataset. The Reduced rules-based method has ranked second compared to the Decision Tree Classifier using SMOTE dataset. This may be due to the loss of information due to the reduction in the rules.

This significant improvement in Accuracy, Recall and Precision of the Decision Tree Classifier using SMOTE data and Reduced rules method can be attributed to the unbalanced nature of the dataset.

Since recall is an important parameter which helps in identifying the target class, it is significant to get a high recall which can classify any given test sample more confidently. Hence, as the Reduced rules-based method has given highest Recall for the Mild and Severe class, it helps in detecting the presence of Anemia at the early stages there by helping the medical fraternity. The use of reduced decision rules would come handy for the medical practitioners in detecting the Mild class Anemia and thereby giving suitable medication for delaying from further deterioration. This is the novel contribution of this work.

6 Conclusion and Direction for Future Work

Anemia detection is one of the challenging issues in current scenario. To address this issue, we have taken Anemia dataset from India. Due to the unbalanced nature of the dataset, we have used SMOTE technique to balance the classes. Decision Tree Classifier and Reduced rule-based method has been used to detect the presence of Anemia from a given dataset. The Reduced rule-based method was able to achieve significant results, especially in the case of Mild class compared to the Decision Tree Classifier using Actual dataset and SMOTE dataset. Due to the smaller number of rules given by reduced rule-based method, it can also be used as handy tool for detection of Anemia. As a future work deep learning algorithm can be used to anemia classifying and optimization-based methods may be used for rules reduction.

Appendix A

Rules Obtained Using Decision Tree on Balanced (SMOTE) Dataset

S. No.	Antecedent	Consequent
1.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548)	Moderate
2.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) and if(Age <= 24.53) and if(PLT/mm3 <= 214.05)	Moderate
3.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) and if(Age <= 24.53) and if(PLT/mm3 <= 214.05) else	Severe
4.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) and if(Age <= 24.53) else	Severe
5.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) else if(MCHC <= 31.89) and if(RDW <= 15.54)	Moderate
6.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) else if(MCHC <= 31.89) and if(RDW <= 15.54) else	Severe
7.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) else if(MCHC <= 31.89) else	Moderate
8.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) else if(RBC <= 2.14)	Severe
9.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) else if(RBC <= 2.14) else	Moderate
10.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) and if(MCHC <= 38.66)	Moderate
11.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) and if(MCHC <= 38.66) else if(Age <= 64.0)	Moderate
12.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) and if(MCHC <= 38.66) else if(Age <= 64.0) else	Mild
13.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) else if(Age <= 24.23)	Moderate
14.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) else if(Age <= 24.23) else	Severe
15.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) and if(PCV <= 35.27)	Moderate
16.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) and if(PCV <= 35.27) else if(MCHC <= 30.58)	Moderate
17.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) and if(PCV <= 35.27) else if(MCHC <= 30.58) else	Mild

(continued)

(continued)

S. No.	Antecedent	Consequent
18.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) else if(MCH <= 20.92)	Moderate
19.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) else if(MCH <= 20.92) else	Mild
20.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) else if(RBC <= 3.46)	Moderate
21.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) else if(RBC <= 3.46) else if(MCH <= 27.66) and if(MCV <= 81.63)	Mild
22.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) else if(RBC <= 3.46) else if(MCH <= 27.66) and if(MCV <= 81.63) else	Moderate
23.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) else if(RBC <= 3.46) else if(MCH <= 27.66) else	Mild
24.	if(PCV <= 29.22) else if(PCV <= 36.48) else if(MCHC <= 29.08) and if(PCV <= 37.64)	Moderate
25.	if(PCV <= 29.22) else if(PCV <= 36.48) else if(MCHC <= 29.08) and if(PCV <= 37.64) else	Mild
26.	if(PCV <= 29.22) else if(PCV <= 36.48) else if(MCHC <= 29.08) else	Mild

Appendix B

Reduced Rules

S. No.	Antecedent	Consequent
1.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) and if(Age <= 24.53) and if(PLT/mm3 <= 214.05) else	Severe
2.	if(PCV <= 29.22) and if(PCV <= 26.13) and if(TLC <= 4.548) else if(MCHC <= 38.43) and if(PCV <= 24.88) and if(Age <= 24.53) else	Severe
3.	if(PCV <= 29.22) and if(PCV <= 26.13) else if(RDW <= 17.42) and if(MCHC <= 38.66)	Moderate
4.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) and if(PCV <= 35.27)	Moderate
5.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) and if(PCV <= 35.27) else if(MCHC <= 30.58) else	Mild
6.	if(PCV <= 29.22) else if(PCV <= 36.48) and if(MCHC <= 32.64) and if(RBC <= 4.71) else if(MCH <= 20.92) else	Mild

References

1. J.R. Khan, N. Awan, F. Misu, Determinants of anemia among 6–59 months aged children in Bangladesh: evidence from nationally representative data. BMC Pediatr. **16**(1), 3 (2016)
2. J.E. Ewusie, C. Ahiadeke, J. Beyene, J.S. Hamid, Prevalence of anemia among under 5 children in the Ghanian population: estimates from the Ghana demographic and health survey. BMC Public Health **14**(1), 626 (2014)
3. M. Abdullah, S. Al-Asmari, Anemia types prediction based on data mining classification algorithms, in *Communication, Management and Information Technology*, ed. by Sampaio de Alencar (2017)
4. B. Sow, S. Hiroki, M. Hamid, A. Hafiz Farooq, Using biological variables and social determinants to predict malaria and anemia among children in Senegal. IEICE Swim **117**, 13–20 (2017)
5. J.G.I. Paulo, A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw. **15**(1), 11–39 (2002)
6. S. Smys, W. Haoxiang, Naïve Bayes and entropy based analysis and classification of humans and chat bots. J. ISMAC **3**(01), 40–49 (2021)
7. World Health Organization, *The World Health Report 2002: Reducing Risks, Promoting Healthy Life* (World Health Organization, 2002)
8. B.J. Brabin, M. Hakimi, D. Pelletier, Iron deficiency anemia: reexamining the nature and magnitude of the public health problem. J. Nutr. **131**, 6045–6155 (2001)
9. E. Mclean, M. Cogswell, I. Egli, B. Wojdyla, B. De Benoist, Worldwide prevalence of anemia, WHO vitamin and mineral nutrition information system, 1993–2005. Public Health Nutr. **12**(4), 444–454 (2009)
10. G. Battineni, G.G. Sagaro, N. Chinatalapudi, F. Amenta, Applications of machine learning predictive models in the chronic disease diagnosis. J. Pers. Med. **10**(2), 21 (2020)
11. X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang, Q. Liu, Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J. Med. Sci. **29**(2), 93–99 (2013)
12. S.B. Choi, W.J. Kim, T.K. Yoo, J.S. Park, J.W. Chung, Y.H. Lee, E.S. Kang, D.W. Kim, Screening for prediabetes using machine learning models. Comput. Math. Methods Med. **2014**, 618976 (2014)
13. W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, Applications of support vector machine modeling for prediction of common diseases: the case of diabetes and pre diabetes. BMC Med. Inform. Decis. Mak. **10**(1), 16 (2010)
14. C.H. Hsieh, R.H. Lu, N.H. Lee, W.T. Chiu, M.H. Hsu, Y.C. Li, Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines and artificial neural networks. Surgery **149**(1), 87–93 (2011)
15. Y. Zhao, B.C. Healy, D. Rotstein, C.R. Guttmann, R. Bakshi, H.L. Weiner, C.E. Brodley, T. Chitnis, Exploration of machine learning techniques in predicting multiple sclerosis disease course. PloS one **12**(4), e0174866
16. S.A. Sanap, M. Nagori, V. Kshirsagar, Classification of anemia using data mining techniques, in *International Conference on Swarm, Evolutionary and Memetic Computing 2011* (Springer, Berlin, Heidelberg, 2011), pp. 113–121
17. M.B. Mengesha, Dadi, *Prevalence of anemia among adults at Hawassa University referral hospital, Southern Ethiopia.* BMC Hematol. **19**, 1 (2019)
18. S. Benjamin, T. Shripad, R. David, Machine learning approaches to the social determinants of health in the health and retirement study. SSM Popul. Health **4**, 95–99 (2018)
19. A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis. Mech. Syst. Signal Process. **21**(6), 2560–2574 (2007)
20. B. Sow, H. Mukhtar, H.F. Ahmad, H. Suguri, Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. Inform. Health Soc. Care **45**(3), 229–241 (2020)

21. M. Jaiswal, A. Srivastava, T.J. Siddiqui, Machine learning algorithms for anemia disease prediction, in *Recent Trends in Communication, Computing, and Electronics* (Springer, Singapore, 2019), pp. 463–469
22. N. Almugren, N. Alrumayyan, R. Alnashwan, A. Alfutamani, I. Al-Turaiki, O. Almugren, The effect of Vitamin B12 deficiency on blood count using data mining, in *5th International Symposium on Data Mining Applications* (Springer, Cham, 2018), pp. 234–245
23. S. Jatoi, M.A. Panhwar, M.S. Memon, J.A. Baloch, S. Saddar, Mining complete blood count reports for disease discovery. Int. J. Comput. Sci. Netw. Secur. **18**(1), 121–127 (2018)
24. K. Meena, D.K. Tayal, V. Gupta, A. Fatima, Using classification techniques for statistical analysis of Anemia. Artif. Intell. Med. **94**, 138–152 (2019)
25. C.C. Chern, Y.J. Chen, B. Hsiao, Decision tree–based classifier in providing telehealth service. BMC Med. Inform. Decis. Mak. **19**(1), 1–15 (2019)
26. K.S. Lakshmi, G. Vadivu, Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Comput. Sci. **115**, 290–295 (2017)
27. Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatr. **27**(2), 130–135 (2015)
28. T.K. Yıldız, N. Yurtay, B. Öneç, Classifying anemia types using artificial learning methods. Eng. Sci. Technol. Int. J. **24**(1), 50–70 (2021)
29. H. Mardiansyah, R.W. Sembiring, S. Efendi, Handling problems of credit data for imbalanced classes using SMOTEXGBoost. J. Phys. Confer. Ser. **1830**(1), 012011 (2021)
30. S. Kilicarslan, M. Celik, Ş. Sahin, Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification. Biomed. Signal Proces. Control **63**, 102231 (2021)
31. R. Vohra, J. Pahareeya, A. Hussain, Complete blood count anemia diagnosis. Mendeley Data **V1** (2021). https://doi.org/10.17632/dy9mfjchm7.1
32. S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)