PREDICTION OF ANEMIA USING MACHINE LEARNING ALGORITHMS

Prakriti Dhakal, Santosh Khanal, and Rabindra Bista

Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

ABSTRACT

Anemia is a state of poor health where there is presence of low amount of red blood cell in blood stream. This research aims to design a model for prediction of Anemia in children under 5 years of age using Complete Blood Count reports. Data are collected from Kanti Children Hospital which consist of 700 data records. Then they are preprocessed, normalized, balanced and selected machine learning algorithms were applied. It is followed by verification, validation along with result analysis. Random Forest is the best performer which showed accuracy of 98.4%. Finally, Feature Selection as well as Ensemble Learning methods, Voting, Stacking, Bagging and Boosting were applied to improve the performance of algorithms. Selecting the best performer algorithm, stacking with other algorithms, bagging it, boosting it are very much crucial to improve accuracy despite of any time issue for prediction of anemia in children below 5 years of age.

KEYWORDS

Machine learning, Anemia, Children, Prediction, Algorithm, Accuracy

1. INTRODUCTION

Machine Learning is based on the idea that a system can learn from data, identifying key patterns for better decision making that applies minimal human intervention [1]. Machine Learning algorithms has proved to be an efficient tool for early prediction of fatal disease such as Anemia, Hepatitis, Lung Cancer, Liver Disorder, Breast Cancer, Thyroid Disease, Diabetes etc. with higher accuracy in order to save human life. In medical science, healthcare related data are being used for predicting epidemics, for detecting various disease, for improving quality of life and avoiding early deaths [2]. Thus, Machine Learning plays an important role in Health Informatics.

Anemia is a nutritional deficiency disorder, global public health problem affecting people of both under developed and developed countries [3]. Anemia is a condition where the total concentration of Red Blood Cells (RBC) or Hemoglobin (Hb) in the blood is low. According to the World Health Organization (WHO), anemia is termed as 'a condition in which the number of red blood cells or their oxygen-carrying capacity is insufficient to meet physiologic needs' [4]. Anemia disease can be classified on the basis of morphology and etiology. The most reliable indicator of anemia is blood hemoglobin concentration, however, there are a number of factors that can cause anemia including iron deficiency, chronic infections such as HIV, Tuberculosis, vitamin deficiencies e.g. vitamins B12 and A, and acquired disorders that affect Red Blood Cell production and Hemoglobin synthesis. Therefore, prediction of anemia plays most important role in order to detect other associated diseases.

Children are the future of any country, the detection of anemia in early age helps to prevent other

associated diseases in future which may seriously hamper their growth and development. This issue emerges a social purpose to conduct this research. Furthermore, Anemia is typically diagnosed on a complete blood count as it is the main test for effective diagnosis of anemia. Henceforth, the main aim of this research is to design a model using different machine learning algorithms and compare the performances of those algorithms on the basis of evaluation criteria for prediction of Anemia using Complete Blood Count (CBC) for children under 5 years. The section II presents the related survey, section III presents the methodology with experimental setup, section IV shows results of experiments and section V concludes the paper.

2. LITERATURE SURVEY

Machine Learning has been an emerging tool for Prediction of Diseases. The work [5] has figured out that each algorithm has its own strength as well as weakness and its own area of implementation. The authors [6] identify those studies that applied more than one supervised machine learning algorithm on one disease prediction. Algorithms include Random Forest, Decision Tree, ANN, SVM, Logistic Regression, Naïve Bayes and K-nearest Neighbor. It shows that Support Vector Machine (SVM) algorithm is applied most frequently and Random Forest (RF) algorithm showed superior accuracy. The research [7] illustrates that many machine learning algorithms have shown good results. It is so as they identify the related attributes accurately.

The authors [8] investigated about supervised machine learning algorithms Naive Bayes, Random Forest and Decision Tree algorithm for prediction of anemia using Complete Blood Count (CBC) where Naive-Bayes technique performed well in terms of accuracy as compared to Decision Tree and Random Forest. The work [9] determined which individual classifier or subset of classifier in combination with each other achieves maximum accuracy in Red blood cell classification for anemia detection showing unique idea of use of subset of classifier and use of ensemble learning techniques. [10] specified anemia type for the anemic patients with dataset from the Complete Blood Count (CBC) which showed J48 Decision Tree as best performer.

The research [11] predicted the anemia status of children under five years taking common risk factors as features. The research concluded that ML methods in addition to the classical regression techniques can be considered to predict anemia. The authors [12] constructed some predictive models by using the identified risk factors through machine learning approach predict the anemia status of children under 36 months. The work [13] constructed a prediction model to predict the potential risk of anemia among infants from Multilayer Perceptron model (MLP) which identified three risk factors for anemia including exclusive breastfeeding, maternal anemia during pregnancy and non-timely supplementation of complementary food. The authors [14] examined the prevalence of anemia in under-five years children taking Ghanian population which showed higher prevalence below 2 years of age. The authors [1] investigates the prevalence of anemia as children grew from infancy to preschool-age for check the dynamic anemia status of children over time where children were at greater risk for developing anemia have persistent anemia between toddlerhood and preschool-age.

From the survey, we could only find few technical predictions for children with technical results that includes running the classifier algorithm and figuring out technical results that includes accuracy, precision and other technical factors. Technical prediction of anemia for other age groups, not for the case of children. Here, the prediction method only considered the risk factor, social, economic factor, not blood reports. Therefore, there became a need of conducting technical research that includes running classifier algorithm and producing technical output like accuracy, precision for prediction of anemia in children considering a detail analysis of Blood Report.

3. METHODOLOGY

We have proposed a research framework to answer our research questions. The research framework is guided by a computational framework. The research started with a literature survey, then advanced to data collection, algorithm processing, verification, validation and at last ended with a result.

3.1. Data Collection

For the data collection, our main research site was Kanti Children Hospital, from where we collected 700 data records of children below 5 years of age. Among the 18 attributes of Complete Blood Count report, RBC counts are mainly used for classifying anemia is a person. So, we selected 7 attributes in the RBC count section. The 7 attributes include Red Blood Cells (RBC), Hemoglobin (Hb), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), RDW-cv. We selected these attributes for prediction of Anemia. For children we consulted a doctor and referred to the pediatric reference range to estimate the cut-off range for children below 5 years of age.

After data collection, the raw data were preprocessed. The data has been recorded in Hematological Analyzer from which data was collected manually, then we prepared dataset for our research with different pre-processing techniques followed by data normalization.

3.2. Model Preparation

After the pre-processing of dataset, the dataset becomes ready to run in a classifier algorithm. For anemia prediction we selected six classifier algorithms i.e. Random Forest, Decision Tree, Naïve Bayes, Artificial Neural Network, Support Vector Machine and Logistic Regression. In the model, we have used 10-fold cross validation for verification and validation. We used 10-fold cross validation for separating data int o training set and testing set where data has been separated into folds i.e. in 10 k-folds.

3.3. Performance Evaluation

The evaluation is based on confusion matrix. There are formulas related with confusion matrix to calculate performance of any classifier algorithm. Along with performance measure from confusion matrix, we have also evaluated other additional performance metrics related to time as well. The performance evaluation was based on accuracy, precision, f-score, recall and area under the curve. We also calculated the CPU time and Wall time required for running the algorithms.

4. EXPERIMENTS AND RESULTS

All data related portion including data pre-processing, data analysis and building different machine learning models were done using Python Programming Language. Some of the tools that have been used for the experimentation part of this research include Google Colab, Python Programming Language, SciPy, Scikit-learn, NumPy, Pandas, Matplotlib.

4.1. Comparative Performance Evaluation

We have selected six classifier algorithms for experimentation. These algorithms have their own specific conditions for processing. These conditions i.e. the hyperparameters were normalized for experimentation. We performed three experiments.

Table 1. Comparative Performance Analysis of Classifier Algorithms

Algorithm	Accuracy	Precision	Recall	F1-score	AUC	CPU time (ms)	Wall time (ms)
Logistic Regression	0.807	0.834	0.862	0.848	0.788	85.2	86.1
Support Vector Machine	0.951	0.980	0.940	0.960	0.955	270	273
Naïve Baye	0.907	0.969	0.878	0.921	0.916	20.4	20.5
Decision Tree	0.972	0.975	0.981	0.978	0.969	32.3	36.9
Artificial Neural Network	0.961	0.970	0.967	0.969	0.959	9600	5110
Random Forest	0.984	0.981	0.988	0.985	0.979	1740	1750

From above Table 1, we can observe that Random Forest is the best performer in the case of all the parameters i.e. Accuracy, Precision, Recall, F1-score, Area Under the Curve. Talking about the time, the best performer took maximum CPU and Wall time to run the entire process. As the prediction system is of medical data, so priority is given to the performance rather than time. The fast performer here is Naïve Baye as it takes minimum CPU time and Wall time. Logistic Regression, on the other hand showed minimum accuracy along with all the other parameters. We can also analyze that as the accuracy improves concurrently there is improvement in performance of other parameters as well.

4.2. Feature Analysis

Our aim in this research is to improve accuracy of the classifier algorithms for making the accuracy more than that of the best performer. For that we conducted feature analysis i.e. selecting best features starting from 3 best to 6 best.

Table 2. Feature Analysis (3 best, 4 best, 5 best, 6 best)

Algorithm	3 best	4 best	5 best	6 best
Logistic Regression	0.764	0.797	0.801	0.821
Support Vector Machine	0.906	0.946	0.948	0.947
Naïve Baye	0.892	0.925	0.912	0.915
Decision Tree	0.921	0.974	0.975	0.968
Artificial Neural Network	0.861	0.927	0.947	0.927
Random Forest	0.938	0.975	0.982	0.981

Above Table 2 shows the corresponding accuracy from feature analysis of all the classifier algorithms on the basis of 3 best, 4 best, 5 best and 6 best features. Overall, we can analyze that the feature analysis and selection method was effective for improving accuracy only for Logistic Regression, Naïve Bayes and Decision Tree whereas for SVM, ANN and Random Forest it was not effective.

4.3. Ensemble Learning Methods

We performed experimentation through feature analysis but we were unable to improve the accuracy more than of the best performer. Then, we conducted experiment by applying the ensemble learning methods such as voting classifier, stacking, bagging, boosting. These ensemble methods were applied to improve the overall accuracy of the model.

4.3.1. Voting Classifier

Below Table 3 shows the results after applying voting classifier for combination of different algorithm along with its CPU time and Wall time. The accuracy of the best performer is 98.4%. Voting classifier was not able to improve accuracy above 98.4% with any of the combination of algorithms.

Table 3. Voting Classifier and corresponding accuracy

Voting Classifier			
Algorithm	**Accuracy**	**CPU time (ms)**	**Wall time (ms)**
RF+LR	0.982	1850	1800
RF+DT	0.972	1850	1810
RF+SVM	0.981	2000	2000
RF+NB	0.964	1790	1800
RF+ANN	0.975	1330	7490
DT+LR	0.972	148	149
DT+SVM	0.972	308	312
DT+NB	0.972	91.6	98.4
DT+ANN	0.972	10700	5.62
SVM+ANN	0.964	11000	5790
SVM+NB	0.937	293	255
SVM+LR	0.941	354	360
NB+ANN	0.942	10300	5470
NB+LR	0.922	129	130
ANN+LR	0.964	16800	5670
RF+LR+DT	0.978	1850	1870
RF+LR+DT+ANN	0.981	13200	7440
RF+LR+DT+NB	0.980	1860	1870
RF+LR+DT+NB+ANN	0.981	1360	7650
RF+DT+NB	0.984	2310	2770
RF+DT+ANN	0.984	7970	7960
RF+DT+ANN+NB+SVM+LR	0.978	14230	8210

4.3.2. Stacking

The Table 4 below shows results after stacking ensemble learning methods. Random Forest with Logistic Regression or ANN when stacked produced accuracy of 98.7% as well as Random Forest when stacked with SVM or Naïve Bayes produced accuracy of 98.6%. Furthermore, when all the six algorithms considered for the study when stacked also produced accuracy of 98.6% i.e. the accuracy improved.

Table 4. Feature Analysis (3 best, 4 best, 5 best, 6 best)

Stacking			
Algorithm	**Accuracy**	**CPU time (ms)**	**Wall time (ms)**
RF+LR	**0.987**	10200	10200
RF+DT	0.981	9790	9800
RF+SVM	**0.986**	12700	13100
RF+NB	**0.986**	10000	10100
RF+ANN	**0.987**	100000	307000
DT+LR	0.972	719000	746000
DT+SVM	0.974	890	892
DT+NB	0.971	321	335
DT+ANN	0.972	31800	17400
SVM+ANN	0.967	56200	30300
SVM+NB	0.958	1.35	1.36
SVM+LR	0.957	1860	1950
NB+ANN	0.960	52.6 s	28.3 s
NB+LR	0.927	607	615
ANN+LR	0.962	32300	17400
RF+LR+DT	0.982	9890	9920
RF+LR+DT+ANN	0.981	66000	40600
RF+LR+DT+NB	0.984	10500	10600
RF+LR+DT+NB+ANN	0.982	65000	39600
RF+DT+NB	0.984	10000	10000
RF+DT+MLP	0.984	64000	38700
RF+DT+ANN+MLP+SVM+LR	**0.986**	65000	39700

4.3.3. Bagging

The Table 5 below shows value of n and the corresponding accuracy. We have to specify how many random samples the data have to be separated i.e. the value of n. For Decision Tree accuracy reached 98.6% when n is 20, which can be said as increase in accuracy than the best performer. Finally, for Random Forest accuracy is 98.8% when value of n is 10, it was higher than the best performer.

Table 5. Bagging with Different Values

Bagging (value of n)	LR	SVM	NB	ANN	DT	RF
10	0.805	**0.959**	**0.918**	0.958	0.98	**0.988**
20	**0.812**	0.958	0.917	**0.962**	**0.986**	0.986
30	0.807	0.958	0.914	0.957	0.980	0.986
40	0.807	0.958	0.915	0.958	0.980	0.985
50	0.804	0.958	0.915	0.958	0.978	0.985
60	0.804	0.958	0.912	0.958	0.982	0.985
70	0.805	0.958	0.911	0.960	0.982	0.985
80	0.808	0.958	0.912	0.960	0.980	0.986
90	0.809	0.958	0.912	0.958	0.978	0.986
100	0.809	0.958	0.911	0.958	0.980	0.986

4.3.4. Boosting

We talk about two boosting approaches namely Adaptive Boosting and XGB Booster in this section.

Adaptive Boosting

The table 6 below shows the accuracy and corresponding CPU and Wall time with Ada Booster. The accuracy of Naïve Bayes was 90.7% and it reached up to 93.7% when applied Adaptive Boosting. For other base estimator, accuracy was not increased, neither the improved accuracy was better than best performer.

Table 6. Adaptive Boosting with Corresponding Accuracy

Boosting (Adabooster)			
Algorithm	**Accuracy**	**CPU time (ms)**	**Wall time (ms)**
Random Forest	0.984	1850	1850
Decision Tree	0.968	52	53
Artificial Neural Network	Nan		
Naïve Bayes	0.937	1650	1650
Support Vector Machine	0.942	4480	4490
Logistic Regression	0.624	3520	3530

XGB Booster

The Table 7 below shows the learning rate adjustment in Xgb Booster along with corresponding accuracy it generates. We could see the fluctuation in the accuracy of the model. Finally, at learning rate 0.07, 0.06, 0.05, 0.04 accuracy reached up to 99%.

Table 7. Learning Rate with Corresponding Accuracy

XGB Booster	
Learning Rate	**Accuracy**
2	0.972
1.9	0.974
1.8	0.982
1.7	0.978
1.6	0.980
1.5	0.980
1.5	0.981
1.4	0.978
1.3	0.986
1.2	0.983
1.1	0.980
1	0.980
0.9	0.986
0.8	0.982
0.7	0.986
0.6	0.978
0.5	0.978
0.4	0.982
0.3	0.984
0.2	0.984
0.1	0.983
0.09	0.984
0.08	0.983
0.07	**0.99**
0.06	**0.99**
0.05	**0.99**
0.04	**0.99**
0.03	**0.99**
0.02	0.984
0.01	0.972

4.4. Balanced Data

Data were balanced, trained, fit into different classifier algorithms model again and finally all the experiments were performed.

4.4.1. Comparative Performance Evaluation

All the parameter, conditions as well as hyperparameters were same for this experiment section.

Table 8. Comparative Table Performance Analysis of Classifier Algorithms (Balanced Data)

Algorithm	Accuracy	Precision	Recall	F1-score	Area Under Curve	CPU time (ms)	Wall time (ms)
Logistic Regression	0.837	0.908	0.750	0.822	0.837	75.4	78
Support Vector Machine	0.954	0.978	0.929	0.953	0.954	357	362
Naïve Baye	0.917	0.964	0.867	0.913	0.917	20.3	22.1
Decision Tree	0.973	0.979	0.972	0.973	0.973	42.5	45.6
Artificial Neural Network	0.964	0.967	0.956	0.962	0.964	11000	5890
Random Forest	0.986	0.984	0.988	0.986	0.986	1850	1830

Above Table 8 shows comparative performance analysis of algorithms from balanced dataset. We can see the change in accuracy and also in other metrics. We can observe similar results, Random Forest is the best performer whereas Logistic Regression is the weakest performer while considering balanced data. However, Accuracy of Random Forest and Logistic Regression was more than that of unbalanced data.

4.4.2. Feature Analysis

For balanced dataset we conducted feature analysis i.e. selecting best features starting from 3 best to 6 best in the same manner as for unbalanced data.

Table 9. Feature Analysis (3 best, 4 best, 5 best, 6 best: Balanced Data)

Algorithm	3 best	4 best	5 best	6 best
Logistic Regression	0.831	0.841	0.836	0.839
Support Vector Machine	0.929	0.951	0.948	0.956
Naïve Baye	0.878	0.926	0.915	0.923
Decision Tree	0.936	0.972	0.971	0.973
Artificial Neural Network	0.938	0.959	0.956	0.957
Random Forest	0.955	0.979	0.980	0.979

The Table 9 shows results for feature selection for balance data. The result is varying in the case of balanced data. For the case of unbalanced data feature selection was effective for Logistic Regression, Naïve Bayes and Decision Tree whereas for balanced data, feature selection was effective for Logistic Regression, SVM and Naïve Bayes. However, the overall accuracy could not be improved.

4.4.3. Ensemble Learning Methods

We proceeded towards experiment by applying the ensemble learning methods i.e. voting classifier, stacking, bagging, boosting for balance dataset.

Voting

The table 10 below shows accuracy for voting classifier with combination of different algorithm in balanced dataset. Voting classifier was not able to improve accuracy above 98.6% with any of the combination of algorithms for balanced data.

Table 10. Voting Classifier and corresponding accuracy (Balanced Data)

Voting Classifier Algorithm	Accuracy	CPU time (ms)	Wall time (ms)
RF+LR	0.978	1930	1940
RF+DT	0.972	1900	1900
RF+SVM	0.982	2190	2200
RF+NB	0.962	1860	1860
RF+ANN	0.979	13500	7690
DT+LR	0.972	141	147
DT+SVM	0.972	406	410
DT+NB	0.972	68.1	71.6
DT+ANN	0.972	10300	5.42
SVM+ANN	0.964	11.2 s	5.97 s
SVM+NB	0.938	374	378
SVM+LR	0.951	435	442
NB+ANN	0.942	10300	4160
NB+LR	0.922	102	109
ANN+LR	0.962	10500	5570
RF+LR+DT	0.978	1940	1950
RF+LR+DT+ANN	0.982	13700	7820
RF+LR+DT+NB	0.982	1970	1980
RF+LR+DT+NB+ANN	0.979	13600	7790
RF+DT+NB	0.982	1900	1910
RF+DT+ANN	0.981	13600	7800
RF+DT+SVM	0.982	2240	2240
RF+DT+ANN+NB+SVM+LR	0.975	13900	8140

Stacking

The Table 11 shows the combination of algorithms when applied stacking ensemble methods with corresponding accuracy. For unbalanced dataset stacking had proved effective to improve accuracy. Stacking ensemble learning method did not proved to be effective for improving overall accuracy than the best performer in the case balanced dataset.

Table 11. Stacking and corresponding accuracy (Balanced Data)

Stacking			
Algorithm	**Accuracy**	**CPU time (ms)**	**Wall time (ms)**
RF+LR	0.986	10800	10800
RF+DT	0.982	10600	10600
RF+SVM	0.986	11900	11900
RF+NB	0.986	10500	10500
RF+ANN	0.986	72000	42400
DT+LR	0.972	701000	725
DT+SVM	0.978	1820	1830
DT+NB	0.972	349	349
DT+ANN	0.978	58800	31200
SVM+ANN	0.965	61000	32800
SVM+NB	0.957	1720	1730
SVM+LR	0.957	2020	2040
NB+ANN	0.965	58700	31100
NB+LR	0.928	518	521
ANN+LR	0.964	58800	31300
RF+LR+DT	0.981	10900	10900
RF+LR+DT+ANN	0.982	71000	42600
RF+LR+DT+NB	0.980	11000	11000
RF+LR+DT+NB+ANN	0.982	71000	42600
RF+DT+NB	0.982	10700	10700
RF+DT+ANN	0.982	71000	42100
RF+DT+SVM	0.981	12100	12100
RF+DT+ANN+NB+SVM+LR	0.982	73000	44500

Bagging

The Table 12 shows the value of n and the accuracy it gives when it executed for balanced data. From the table we can elaborate that when the value of n was adjusted, the accuracy was improved for all the algorithms. Random Forest when bagged gave accuracy higher than best performer. Bagging proved to be the effective method to improve accuracy for best performer in the case of balanced data. In the case of unbalanced data also bagging proved to be effective.

Table 12. Bagging with Different Values (Balanced Data)

Bagging (value of n)	LR	SVM	NB	ANN	DT	RF
10	0.844	0.956	0.916	0.962	0.978	0.986
20	**0.845**	0.956	**0.918**	0.966	0.981	**0.987**
30	0.841	**0.957**	0.916	0.964	0.982	**0.987**
40	0.844	0.956	0.915	0.965	0.982	**0.987**
50	0.842	0.956	0.915	0.966	**0.983**	0.986
60	0.844	0.956	0.915	0.965	0.982	0.986
70	0.844	0.956	0.915	0.966	0.982	0.986
80	0.842	0.956	0.915	**0.969**	0.982	0.986
90	0.844	0.956	0.914	**0.969**	0.982	0.986
100	0.842	0.956	0.915	0.970	0.982	0.986

Boosting

Adaptive Boosting

The Table 13 shows the accuracy and corresponding CPU and Wall time with Ada Booster with balanced data. For Random Forest as a base estimator the accuracy remained the same. For other base estimators' accuracy was not increased. In this case, Adaptive Boosting did not prove to be effective for increasing accuracy than the best performer for balanced dataset.

Table 13. Adaptive Boosting with Corresponding Accuracy (Balanced Data)

Boosting (Adabooster)			
Algorithm	**Accuracy**	**CPU time (ms)**	**Wall time (ms)**
Random Forest	0.986	2040	2040
Decision Tree	0.973	63.8	75.9
Artificial Neural Network	Nan		
Naïve Bayes	0.940	206	210
Support Vector Machine	0.900	1530	1530
Logistic Regression	0.624	403	403

XGB Booster

The table 14 shows the learning rate adjustment in Xgb Booster along with corresponding accuracy it generates. We could observe the fluctuation in the accuracy of the model. At learning rate 0.03 accuracy reached up to 98.6% which was just equal to the best performer.

Table 14: Learning Rate with Corresponding Accuracy (Balanced Data)

XGB Booster	
Learning Rate	**Accuracy**
2	0.979
1.9	0.980
1.8	0.978
1.7	0.975
1.6	0.978
1.5	0.977
1.4	0.979
1.3	0.981
1.2	0.981
1.1	0.981
1	0.974
0.9	0.980
0.8	0.982
0.7	0.980
0.6	0.983
0.5	0.982
0.4	0.982
0.3	0.981
0.2	0.982
0.1	0.983
0.09	0.982
0.08	0.983
0.07	0.983
0.06	0.982
0.05	0.982
0.04	0.982
0.03	0.986
0.02	0.983
0.01	0.981

We can see some variations in results of unbalanced and balanced data. For both the cases, the best performer was Random Forest. For unbalanced data accuracy was 98.4% whereas accuracy increased up to 98.6% for balanced data. The accuracy for all the algorithms increased when data was balanced. In the case of unbalanced data, the F1-score was higher than the accuracy. However, F1-score should be lower than the accuracy as it is one of the major parameters in the case of medical data. For the balanced data this issue was solved as F1-score was not more than accuracy, rather it was less or equal to accuracy for all the classifier algorithms. The same scenario of trade off factor for time was observed for both the data nature. The fast performer for both the cases was Naïve Baye taking minimum CPU time and Wall time. Overall, ensemble methods proved to increase overall accuracy of the prediction system for anemia in children.

4.5. Proposed Prediction Method

The Fig. 1 and Fig. 2 show our proposed prediction method for unbalanced and balanced data respectively. For unbalanced data, Random Forest showed the highest accuracy. Stacking all the six-classifier algorithm, bagging Random Forest 10 times, Decision Tree 20 times and adjusting XGB Booster's Learning Rate from 0.06 to 0.03 the accuracy was improved. However, for balanced data the scenario was different. Accuracy was successful to increase by 0.3% when Random Forest was bagged where number of random samples equals to 20, 30 and 40. For other considered ensemble learning methods accuracy couldn't increase more than the best performer. For some cases accuracy was similar to the best performer for Prediction of Anemia in children below 5 years of age.

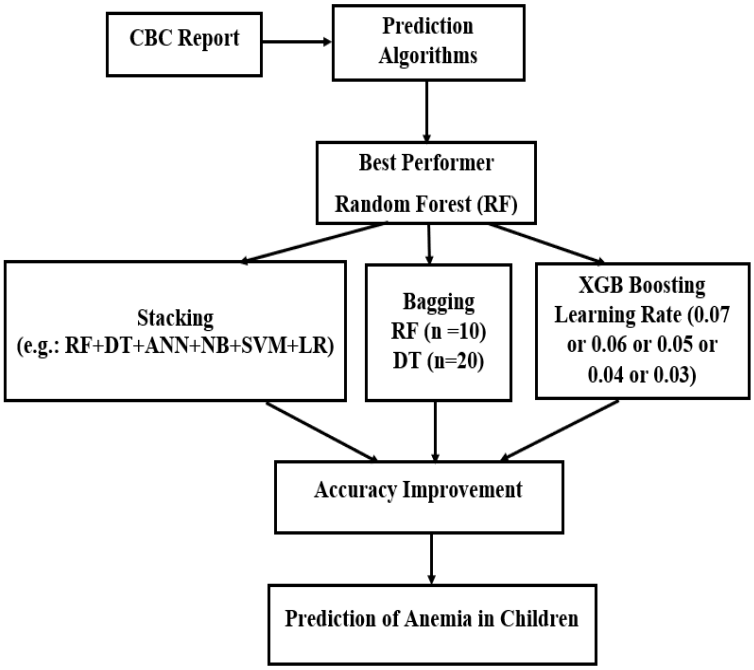

Fig. 1. Proposed Prediction Model (Unbalanced Data)

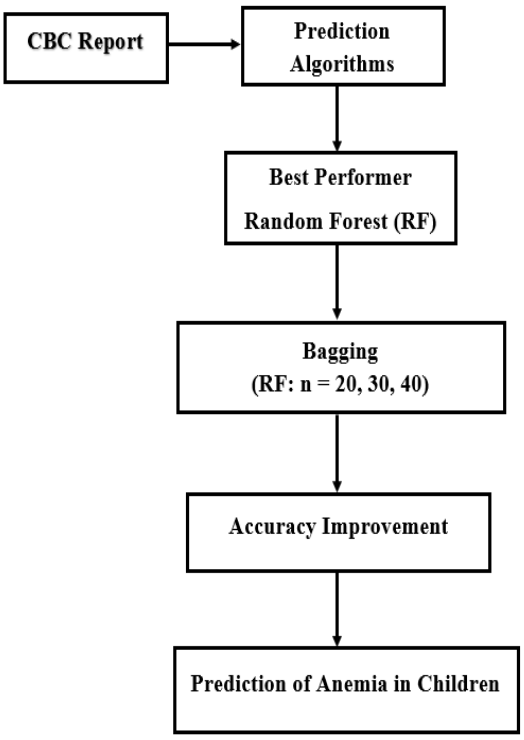

Fig. 2. Proposed Prediction Model (Balanced Data

5. CONCLUSION

The research study aims to predict anemia in children below 5 years of age. From the literature survey we found out that we could not find technical prediction for children. Moreover, the prediction method has only considered the risk factor, social, economic and demographic factor rather than giving emphasis for blood reports. So, we found a need for conducting our research study to technically predict anemia for children considering the blood report.

We ran the pre-processed data into classifier algorithms and conducted experiments for unbalanced as well as balanced data. For unbalanced data, results showed Random Forest with accuracy of 98.4%. We performed various experiments to improve accuracy of the model than the best performer. We conducted feature analysis from which we could not increase accuracy more than that of the best performer. After that, we applied ensemble learning methods in our experiment. The accuracy was increased up to 98.8% when applied with Stacking and Bagging. The accuracy increased by 0.2% i.e. it reached 98.6% when stacking Random Forest with SVM or Naïve Bayes, stacking all the six algorithms and bagging Decision Tree 20 times. Stacking Random Forest with Logistic Regression or ANN increased the accuracy by 0.3% i.e. it was increased by 98.7%. Accuracy was increased by 0.4% i.e. it reached 98.8% when Random Forest was bagged and the number of random samples equals to 10. Then, applying Extreme Gradient Boosting accuracy reached up to 99%. This was the case for unbalanced data. Then for balanced data, accuracy was increased by 0.1% which reached 98.7% when Random Forest was bagged where number of random samples equals to 20, 30 and 40. For other ensemble learning methods accuracy couldn’t increase or accuracy remained the same.

Finally, we developed a new proposed prediction framework for both unbalanced as well as balanced data which improves accuracy of existing algorithm. Therefore, we claim that selecting the best performer algorithm, stacking with other algorithms, bagging it, boosting it are very much crucial to improve accuracy despite of any time issue for prediction of anemia in children below 5 years of age.

REFERENCES

[1] L. Wang, M. Li, S. E. Dill, Y. Hu, and S. Rozelle, "Dynamic Anemia Status from Infancy to Preschool-Age: Evidence from Rural China," International journal of environmental research and public health, vol. 16, no. 15, pp. 2761, 2019.

[2] V. Arun, V. Shyam, and S. K. Padma, "Privacy of health information in telemedicine on private cloud," Int J Family Med Med Sci Res, vol. 4, no. 189, pp. 2, 2015.

[3] N. Soundarya and P. Suganthi, "A review on anaemia–types, causes, symptoms and their treatments," Journal of science and technology investigation, vol. 1, no. 1, 2017.

[4] N. Alli, J. Vaughan, and M. Patel, "Anaemia: Approach to diagnosis," SAMJ: South African Medical Journal, vol. 107, no. 1, pp. 23-27, 2017.

[5] H. Bhavsar and A. Ganatra, "A comparative study of training algorithms for supervised machine learning," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 4, pp. 2231-2307, 2012.

[6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1-16, 2019.

[7] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, pp. 1, 2017.

[8] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine Learning Algorithms for Anemia Disease Prediction," in Recent Trends in Communication, Computing, and Electronics, Springer, Singapore, 2019, pp. 463-469.

[9] P. T. Dalvi and N. Vernekar, "Anemia detection using ensemble learning techniques and statistical models," in 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, pp. 1747-1751.

[10] M. Abdullah and S. Al-Asmari, "Anemia types prediction based on data mining classification algorithms," Communication, Management and Information Technology–Sampaio de Alencar (Ed.).

[11] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine learning algorithms to predict the childhood anemia in Bangladesh," Journal of Data Science, vol. 17, no. 1, pp. 195-218, 2019.

[12] P. Anand, R. Gupta, and A. Sharma, "Prediction of Anemia among children using Machine Learning Algorithms," International Journal of Electronics Engineering, vol. 11, no. 2, pp. 469-480, 2019.

[13] J. Zhang and W. Tang, "Building a prediction model for iron deficiency anemia among infants in Shanghai, China," Food Science & Nutrition, 2019.

[14] J. E. Ewusie, C. Ahiadeke, J. Beyene, and J. S. Hamid, "Prevalence of anemia among under-5 children in the Ghanaian population: estimates from the Ghana demographic and health survey," BMC public health, vol. 14, no. 1, p. 626, 2014.

AUTHORS

Prakriti Dhakal, Recent Graduate of Masters in Computer Engineering Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

Santosh Khanal, Assistant Professor, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

Rabindra Bista*, Associate Professor, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal ,*- Corresponding Author