AN ENSEMBLE FRAMEWORK FOR CLASSIFICATION OF MALARIA DISEASE

T. Sajana and M. R. Narasingarao
Department of Computer Science Engineering, K L E F, Vaddeswaram, Guntur, Andhra Pradesh, India
E-Mail: sajana.cse@kluniversity.in

ABSTRACT

Malaria disease is one whose presence is rampant in semi urban and non-urban areas especially resource poor developing countries. It is quite evident from the datasets like malaria etc., where there is always a possibility of having more negative patients (non-occurrence of the disease) compared to patients suffering from disease (positive cases). Developing a model based decision support system with such unbalanced datasets is a cause of concern and it is indeed necessary to have a model predicting the disease quite accurately because most of the conventional machine learning algorithms are showing very poor performance to classify the skewed distribution data i.e., whether a patient is affected by malaria disease or not because in imbalanced data, majority (unaffected) class samples are dominates the minority (affected) class samples which leading to class imbalance problem. To overcome this nature of class imbalance problem ensemble methods are used which produces the better accuracy in classification of minority samples. The aim of this research is to propose a comparative study on classifying the imbalanced and balanced malaria disease datasets using various ensemble methods like boosting, bagging and voting algorithms for accurate classification of affected patient. Experimental outcomes shows that Random Forest algorithm shows outstanding performance for the classification of imbalanced malaria disease.

Keywords: malaria, imbalanced data, balanced data, AdaBoost, random forest, voting.

1. INTRODUCTION

Malaria disease is one of the elevating issues among the vector borne diseases in medical domain [1]. It becomes one of the global health problem caused by a mosquito bite [1] [2]. Malaria, which is a vector borne disease, affects the rural population for many years. Even though people maintain with healthy life style with good food habits and with neat surroundings, still due to the climate changes or for any other reason, many people are affected by malaria disease [1], [3]. According to 2016 report of World Health Organization (WHO), one million people are dying annually due to the vector borne diseases like malaria [4]. Irrespective of the age factor, Millions of deaths have occurred which is estimated to be 839000 in 2000 (range: 653000-1.1 million) to 438000 in 2015 (range: 236000-635000) i.e., 48% have been recorded. Overall, it is estimated that, the incessant presence of Malaria disease has decreased the world population by 60% [4-5].

One way to solve this problem is to properly identify the patient with a disease and providing accurate diagnosis. Malaria a vector borne disease, the data of which always creates a class imbalance problem because of the presence of samples of negative class (unaffected patients) dominating the samples of positive class (affected patients) [6-7]. Handling such a kind of imbalanced data, is a critical task because prediction of a patient with or without a disease becomes an important problem in the medical scenario [6-8]. Unbalanced data sets are include not only medical domain but also domains like credit card fraud detection[9], detection of problems in software's[9-10],detection of oil spills in satellite radar images[11-12], frauds in telecommunications[9], detection of frauds in financial sector[9] etc.

Analysing an imbalanced data set is very pertinent in different diseases like malaria [6]. Handling such a complex nature of data, becomes an issue in Data Mining and Machine Learning domains [6] and it is observed that, most of the traditional machine learning algorithms is very sensitive with imbalanced data [13-14]. In general, the main goal of machine learning algorithms is to achieve good accuracy for classification/prediction. While dealing with imbalanced data, the classifier predicts towards the majority class rather than minority class because the classifier will be trained with many majority class examples rather than minority class examples and hence the prediction of the classifier will be more oriented towards majority class examples. So, accurate classification of minority class samples is important than majority class samples especially in medical diagnosis [6]. For example, misclassification of an affected patient is more dangerous than unaffected patient [13-14]. The organization of rest of the paper as follows: Section 2 describes the literature review of various methods of classification of balanced and imbalanced malaria disease data. Section 3 presents the experimental frame work for classification of both imbalanced and balanced malaria disease data using Ensemble frame work and Section 4 shows the experimental outcomes.

2. LITERATURE REVIEW

Machine learning and Data mining algorithms are well performed on classification of datasets especially in medical domain. Because all these traditional algorithms are trained equally with samples which are belongs to target classes' i.e. positive class and negative class. Various researchers suggested different kinds of machine learning and data mining algorithms for the classification of malaria disease data as stated below:

- Miranda I. Teoboh - Ewungkem *et al.* [15] determined the gametocyte stage of malaria parasite using Self-Knowledge Model (SKM) and Dual-Knowledge Model (DKM).

- Noah H. Paul *et al.* [16] investigated a method for detection of malaria parasite P. Falciparum using Rapid diagnostic test, Nested PCR methods.

- Kshipra C.Charpe *et al.* [17] investigated malaria parasite stages using Image processing and classification techniques.

- J. Somasekar *et al.* [18] proposed a method for detection of effected erythrocytes based on Adaptive median filter, edge enhancement and Fuzzy C-Means clustering techniques.

- J.E. Arco *et al.* [19] investigated and estimated parasite density leading to the detection of Malaria using Adaptive threshold and Connected Component Analysis.

- Hanung Adi Nugroho *et al.* [20] examined the classification of malaria parasite and proposed a method for detection of malaria parasite stages using K-means clustering and Multilayer Perceptron Neural Networks.

- H. Chiroma *et al.* [21] developed a method for density estimation of malaria parasites using Jordon – Elman Neural Network.

Based on the outcomes of above stated methodologies classification of any sample will be done accurately with balanced distribution of samples. But, in the era of imbalanced datasets all these conventional algorithms are failed to classify the minority class samples especially because the learning rate of all traditional algorithms are biasing towards with the majority class samples only and hence accurate classification of minority class samples not to be done by the classifiers [9-10]. Concentrating into the medical domain in the view of imbalanced data early stage identification is very important for an affected patient otherwise; sometimes it may even leads to death also. But, still there is lack of research in identification of an affected patient (positive class sample) in imbalanced malaria disease dataset [13-14], [22], [23] because of the biasing problem of skewed distribution data. Hence, classifying the minority class samples from the past few years became a challenging issue to many researchers which also leads to class imbalance problem in many fields [22-23]. Consider the literature review on imbalanced malaria disease as follows:

- Manoj Gambhir *et al.* [24] investigated that malaria disease still becomes a global burn disease even though there is a control on vectors and also suggested that there is a necessity of additional control intervention called mass drug administration to reduce the malaria risk.

- Ewan Cameron *et al.* [25] suggested a relationship between the infection of different aged prevalence and clinical studies of parasite count of Plasmodium falciparum by using an ensemble approach of statistical Bayesian which combines regression based model with Markov Chain Monte Carlo sampling.

- Daniel Ruiz *et al.* [26] investigated on the records of temperature, rainfall etc. and suggested an ensemble malaria model for assessment of long term impact of climate conditions.

- Salma Jamal *et al.* [6] developed a Predictive model of anti-malarial molecules inhibiting apicoplast formation for an imbalanced malaria disease dataset using Cost sensitive Naive Bayesian, Cost sensitive Random Forest and Meta Cost J48 algorithms and found that Cost sensitive Random Forest algorithm is the best algorithm. 2013

- Raquel M. Goncalves *et al.* [27] invented that Plasmodium Vivax has more impact on human functionality than the effect of Plasmodium falciparum by conducting statistical analysis.

- Bruno B Andrade *et al.* suggested that severe stage of malaria disease causes to reducing of inflammatory cytokines which is a high level imbalance class problem in medical domain [8].

3. RESEARCH DESIGN & METHODOLOGY

Skewed distribution data classification is playing a vital role especially in medical domain. Many researchers are proposed well defined sophisticated methods for handling and learning imbalanced data sets like data sampling methods, cost sensitive methods, kernel function methods [28-29]. But the drawback of existing methods are loss of data or over fitting, increasing in misclassification cost error rate. Hence, we are proposing ensemble methods which are best methods for classification of imbalanced datasets particularly because set of classifiers are grouped together to improve the accuracy performance of classifiers and they can reduces the data over fitting or loss and may nullify the misclassification error rate [30-31]. Consider the design of proposed methodology as shown in Figure-1 which describes the most popular methods of ensemble methods like Boosting, Bagging and Voting as stated below:

Boosting: An ensemble algorithm that creates set of models that attempt to correct the mistakes of the models before them in the sequence. One of the most popular methods for Boosting is AdaBoost algorithm as described in Algorithm 1 Algorithm 1 - AdaBoost

- Fit a sequence of weak learners (i.e., models that are only such as small decision trees)
- Repeatedly call the learners by using modified versions of the data.
- Make the predictions from classifiers.
- The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.

Bagging: Bootstrap Aggregation or bagging involves taking multiple samples from training dataset (with replacement) and training a model for each sample. The final output prediction is averaged across the predictions of all of the sub-models. Let us consider the popular method of Bagging is Random Forest. Which is a forest of decision trees in which samples of the training dataset are taken with replacement, but the trees are constructed in a way that reduces the correlation between individual classifiers. Specifically, only a random subset of features is considered for each split. Let us describe the Random Forest algorithm as stated in Algorithm 2.
Algorithm 2 - Random Forest

- Samples are drawn with replacement if bootstrap =True by default.
- Fits a number of decision tree classifiers on various sub-samples of the dataset
- Use averaging to improve the predictive accuracy and control over-fitting.

Voting: The idea behind the `Voting Classifier` is to combine conceptually different machine learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses as described in the Algorithm 3.
Algorithm 3 - Voting

- Samples are trained on different machine learning classifiers like MLP and LDA Classifiers.
- Count the majority vote on classification of sample.
- Consider majority vote to improve the predictive accuracy and control over-fitting.

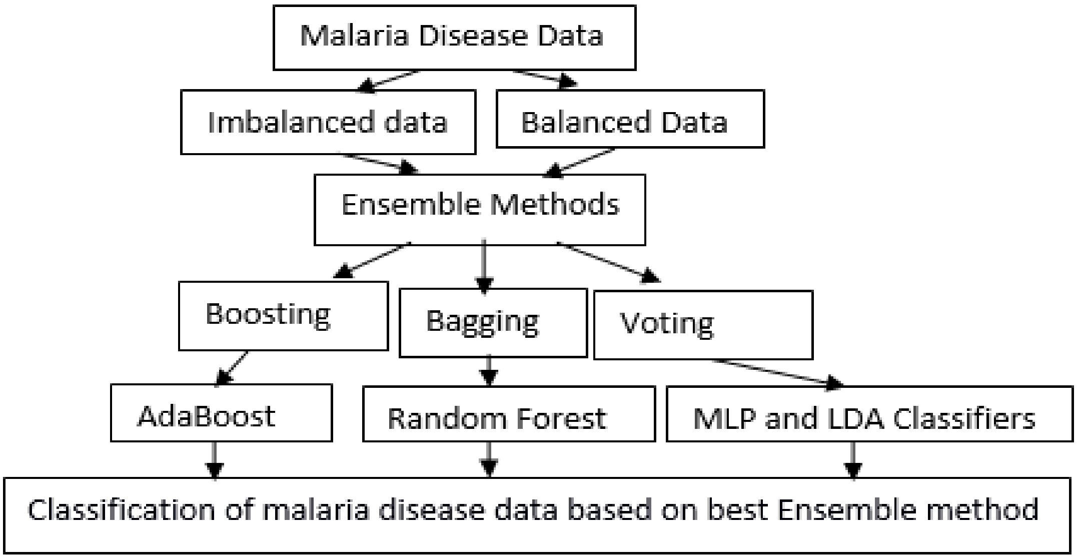

Figure-1. Proposed ensemble method for classification of both imbalanced & balanced malaria disease datasets.

Dataset description: we have collected 165 patients' data from the medical wards of Narasaraopet in which it consists of various attributes like Age, Haemoglobin, RBC, Hct, Mcv, Mch, Mchc, Platelets, WBC, Granuls, Lymphocytes, Monocytes, Malaria.

Consider the Class Imbalance Distribution of Malaria Disease Dataset as follows:

Total no of instances - 165,
Total no of Attributes - 13,
Class (Minority, Majority) - (Positive, Negative),
% of Class - Minority: 0.03, Majority: 96.9,
Class Imbalance Ratio (IR) - 32.30.

To balance the class distribution we applied an over sampling technique called SMOTE - Synthetic Minority Oversampling Technique on imbalanced malaria disease data that inference the balanced malaria disease dataset.

4. RESULTS AND DISCUSSIONS

Considered a set of well-defined algorithms of ensemble methods namely AdaBoost, Random forest and voting algorithms. Presenting a well-defined methodology by giving common distribution of samples to the frame work as shown in Figure-1. To measure the performances of these proposed ensemble methods the metrics used are Accuracy, precision, recall and F1-score are used which are derived from confusion matrix as stated in Table-1:

Table-1. Confusion matrix.

		Predicted	
		Positive class “Effected”	**Negative class “Unaffected”**
Actual	Positive Class “Effected”	TP (True Positive)	FN (False Negative)
	Negative Class “Unaffected”	FP (False Positive)	TN (True Negative)

Where Accuracy = [TP + TN] / [TP + FN + FP + TN]
Precision = [TP] / [TP + FP]
Recall = [TP] / [TP + FN]
F1-Score = [2 x Precision x Recall]/[Precision + Recall]

Initially we measures the performance metrics of all these proposed ensemble methods namely AdaBoost, Random forest and Voting algorithms with imbalanced malaria disease data and balanced malaria disease data as shown in Figures 2-10 and then conducted a comparison study between the proposed ensemble methods which are shown in Figures 11-14 to suggest which method is best for classification of both imbalanced and balanced malaria disease datasets

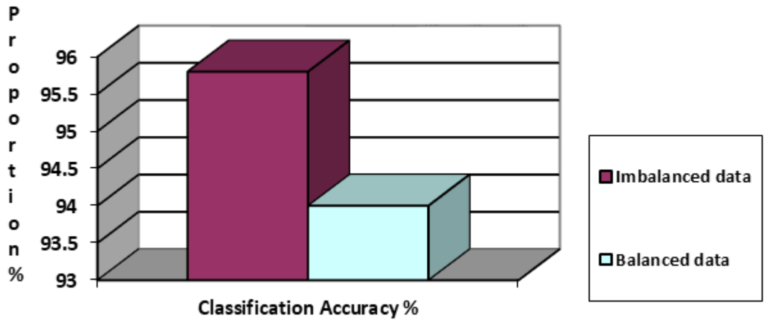

Figure-2. Classification Accuracy % of both imbalanced & balanced malaria disease data using AdaBoost algorithm.

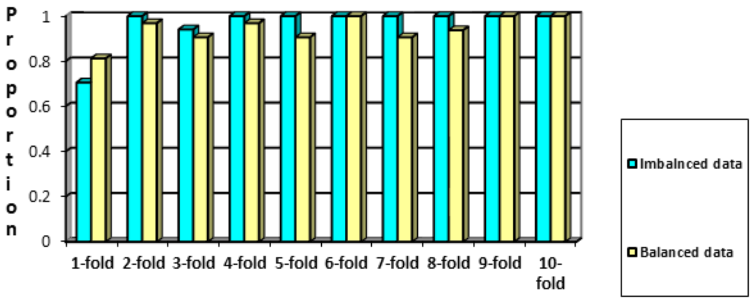

Figure-3. AdaBoost 10 - fold cross validation result on both imbalanced & balanced malaria disease datasets.

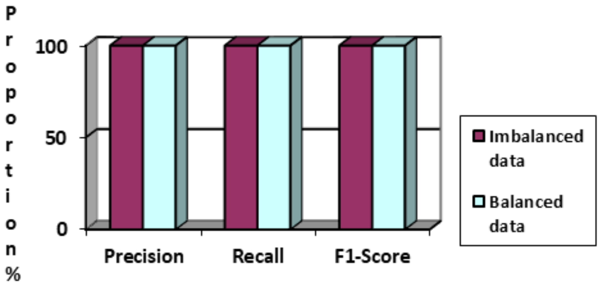

Figure-4. Classification report of AdaBoost algorithm on Imbalanced & Balanced malaria disease datasets.

For the classification of malaria disease data AdaBoost performs good accuracy of 95.8% on imbalanced malaria disease data when compared with balanced malaria disease data as 94.0% which is shown in Figure-3 and it also reduces the bias variance at one level with imbalanced data whereas the bias variance not that much reduced with balanced data but AdaBoost performed equal proportion of Precision, Recall and F1-score measures with both imbalanced and balanced malaria disease datasets which are shown in Figures 2, 3, 4.

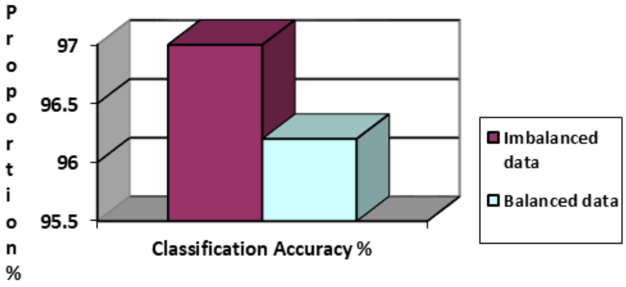

Figure-5. Classification Accuracy % of both imbalanced & balanced malaria disease data using Random Forest algorithm.

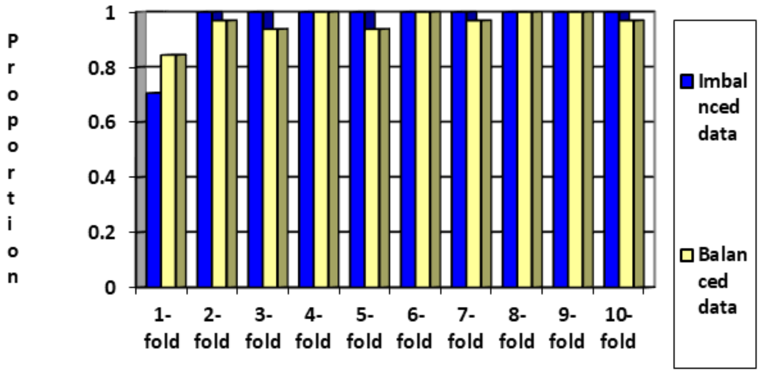

Figure-6. Cross validation results of random forest algorithm on both imbalanced & disease datasets.

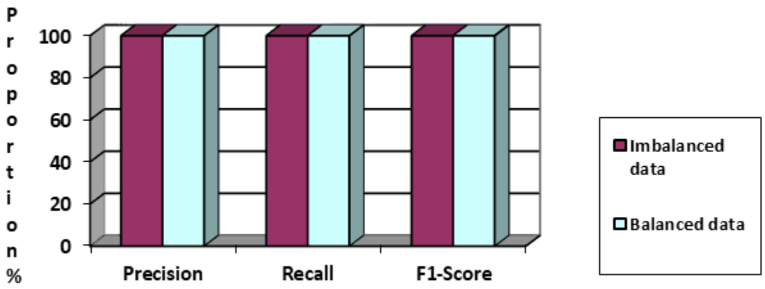

Figure-7. Classification report of random forest algorithm on both imbalanced & balanced malaria disease datasets.

Random Forest Tree, which is one of the best Bagging algorithm that shows outstanding classification performance of 97.0% with imbalanced malaria disease data where as it shows 96.2% of accuracy on balanced malaria disease data as shown in Figure-5. Bias variance also reduced by Random Forest at high level when conducted comparison between both Imbalanced and balanced malaria disease datasets as shown in Figure-6 but it shows an equal performance of classification metrics like Precision,Recall,F1-Score on both imbalanced and balanced malaria disease datasets as shown in Figure-7.

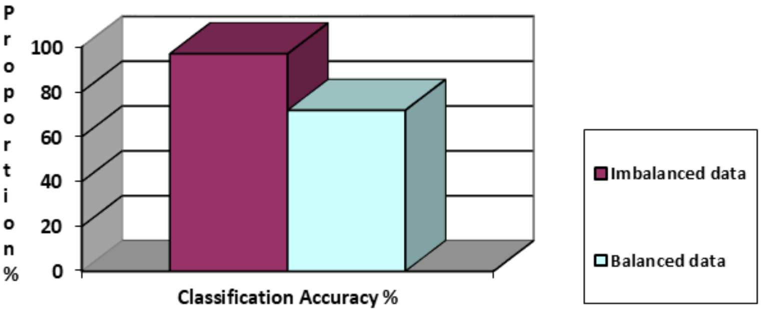

Figure-8. Classification accuracy % of both imbalanced & balanced malaria disease datasets using voting algorithm.

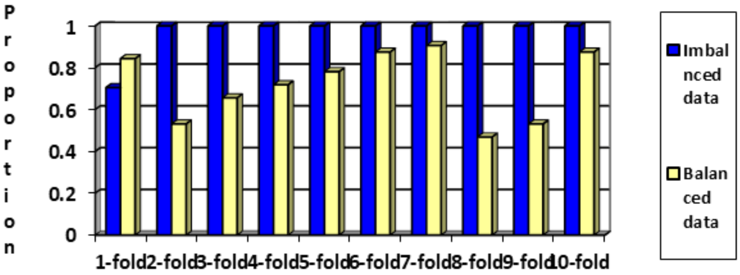

Figure-9. Cross validation results of voting algorithm on both imbalanced & balanced malaria disease datasets.

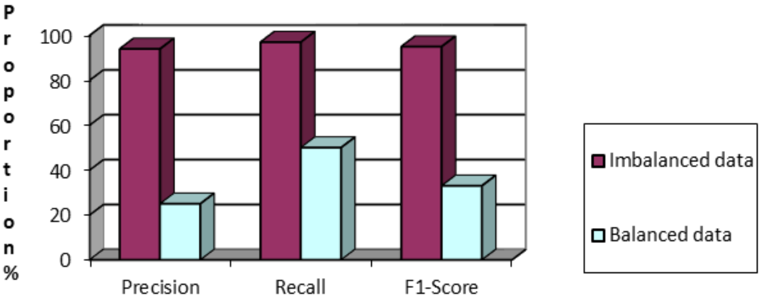

Figure-10. Classification report of voting algorithm on imbalanced & balanced malaria disease datasets.

Presenting a combination of machine learning algorithms like MLP and LDA Classifiers as an ensemble method of Voting which classifies the samples based on majority voting. Voting ensemble algorithm increases the classification accuracy performance of imbalanced malaria disease data i.e., 97% when compared with balanced malaria disease data as 71% which is as shown in Figure-8. But unlike other ensemble methods voting of MLP and LDA Classifiers shows good bias variance at each k-fold cross validation of both imbalanced and balanced malaria disease datasets which is shown in Figure-9 and also shows 94% of Precision, 97% of Recall and 95% of F1-Score on imbalanced malaria disease data and 25% of Precision, 50% of Recall and 33% of F1-Score with balanced malaria disease data as shown in Figure-10.

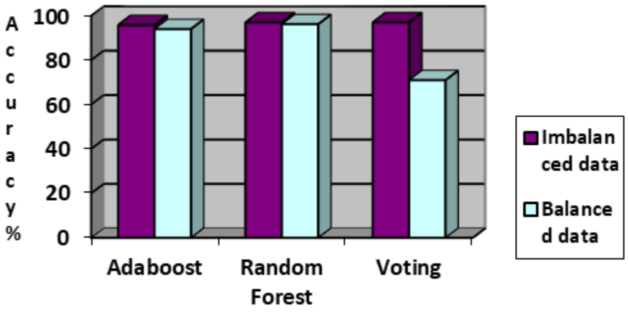

Figure-11. Accuracy comparison of proposed ensemble algorithms - AdaBoost, random forest and voting algorithms.

Looking into the overall performance of proposed ensemble methods Random Forest algorithm shows best accuracy performance on classification of both imbalanced and balanced malaria disease data sets when compared with AdaBoost and voting algorithms which is shown in Figure-11.

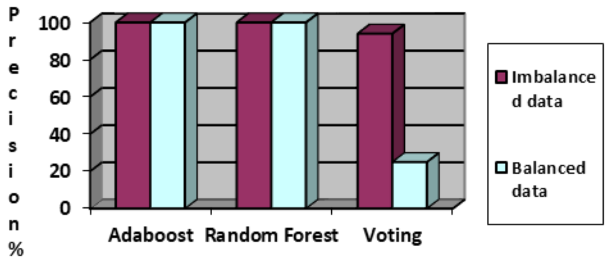

Figure-12. Precision comparison of proposed ensemble algorithms - AdaBoost, random forest and voting algorithms.

To know the exact performance of classifiers a metric called Precision is used on both imbalanced and balanced malaria disease datasets which is shown in Figure-12. Except voting algorithm all the remaining ensemble methods shows equal classification report on both imbalanced and balanced malaria disease datasets whereas voting shows good precision report on imbalanced data when compared with balanced malaria disease data.

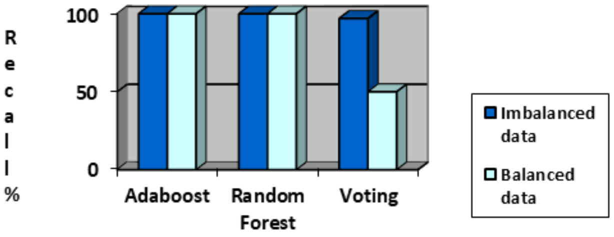

Figure-13. Recall comparison of proposed ensemble algorithms - AdaBoost, random forest and voting algorithms.

Another classification performance metric is Recall which defines the completeness of classification of samples. Like precision all the ensemble methods shows same performance for recall metric except voting which shows 97% on imbalanced and 50% on balanced malaria disease datasets as shown in Figure-13.

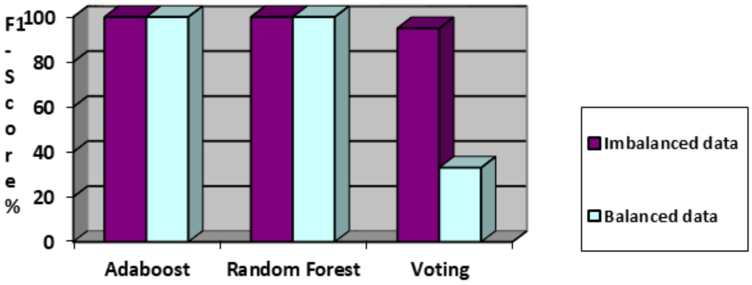

Figure-14. F1-Score comparison of proposed ensemble algorithms - AdaBoost, random forest and voting algorithms.

F1- Score or F-Score is one of the classification performances metric that depends on weighted average on both Precision and Recall metrics. Like precision and recall measures the ensemble methods shows equal performance on both imbalanced and balanced data sets except voting algorithm which is shows 95% and 33% on both imbalanced and balanced disease datasets as shown in Figure-14.

5. CONCLUSIONS

Malaria disease becomes one of the class imbalance problems in medical diagnosis. If we are examined in a nutshell, classification of such imbalanced data especially minority class samples is a key thing. So accurate prediction of effected patient and diagnosing within time is very important otherwise it may lead to death also. Many conventional classifiers are voted for classification of majority class samples only. So handling minority class samples and its classification is becoming a burning issue in medical field. Hence, proposing ensemble methods which shows best classification performance on imbalanced data when compared with balanced malaria disease dataset. Conducted a comparative study between ensemble methods like AdaBoost, Random forest and voting ensemble methods to classify the minority class samples especially. Out of all the ensemble methods Random Forest - Bagging ensemble shows best performance and presented the balanced class distribution of imbalanced malaria disease data by using Synthetic minority oversampling technique - SMOTE algorithm. Finally conducted a comparative study between both imbalanced and balanced malaria disease datasets by using ensemble methods for a better prediction system.

REFERENCES

[1] Thanh Quang Bui and Hai Minh Pham. 2016. Web based GIS for spatial pattern detection: application to malaria incidence in Vietnam. Bui and Pham Springer plus. 5: 1014, pp. 1-14.

[2] Dave A MacLeod, Anne Jones *et al.* 2015. Demonstration of successful malaria forecasts for Botswana using an operational seasonal climate model. Environmental research letters, IOP Publishing. 10 044005, pp. 1-11.

[3] Md Z Rahman, Leonid Roytman *et al.* 2015. Environmental Data Analysis and Remote Sensing for Early Detection of Dengue and Malaria. Proc. of SPIE. 9112: 1-9.

[4] WHO Malaria Report - 2016, http://www.who.int/mediacentre/factsheets/fs387/en/, 2016. World Malaria Report - 2015.

[5] Pages-x, xi. http://apps.who.int/iris/bitstream/10665/200018/1/978 92415651 58_eng.pdf.

[6] Salma Jamal, Vinita Periwal *et al.* 2013. Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics. 2013, 14: 55, 1-8.

[7] Tsige Ketema and Ketema Bacha. 2013. Plasmodium vivax associated severe malaria complications among children in some malaria endemic areas of Ethiopia. BMC Public Health. 2013, 13: 637, pp. 1-7.

[8] Bruno B Andrade, Antonio Reis-Filho *et al.* 2010. Severe Plasmodium vivax malaria exhibits marked inflammatory imbalance. Malaria Journal. 2010, 9: 13, pp. 1-8.

[9] Guo Haixiang, Li Yijing, *et al.* 2016. Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications. pp. 1-49.

[10] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. pp. 1-12.

[11] Xiaoheng Deng, Weijian Zhong *et al.* 2016. An Imbalanced Data Classification Method Based on Automatic Clustering Under-Sampling. IEEE transaction. pp. 1-8.

[12] Aida Ali, Siti Mariyam Shamsuddin *et al.* 2015. Classification with class imbalance problem: a Review. International journal of Advances in Soft Computing and its Applications. 7(3): 176-204.

[13] N. Poolsawad, C. Kambhampati *et al.* 2014. Balancing Class for Performance of Classification with a Clinical Dataset. Proceedings of the World Congress on Engineering. 1: 1-6.

[14] M. Mostafizur Rahman and D. N. Davis. 2013. Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing. 3(2): 224-228.

[15] Miranda I. Teboh - Ewungkem and Thomas Yuster. 2016. Evolutionary implications for the determination of gametocyte sex ratios under fecundity variation for the malaria parasite. Elsevier- Journal of Theoretical Biology. 408, pp. 260-73.

[16] Noah H. Paul, Arthur Vengesai *et al.* 2016. Prevalence of Plasmodium falciparum transmission reducing immunity among primary school children in a malaria moderate transmission region in Zimbabwe. Elsevier - Acta Tropica. 163, pp. 103-108.

[17] Kshipra C. Charpe, Dr. V. K. Bairagi *et al.* 2015. Automated Malaria Parasite and there Stage Detection in Microscopic Blood Images. IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO).

[18] J. Somasekar, B. Eswara Reddy. 2015. Segmentation of erythrocytes infected with malaria parasites for the diagnosis using microscopy imaging. Elsevier - Computers and Electrical Engineering. pp. 336-351.

[19] J.E. Arco, J.M. Górriz et al. 2015. Digital image analysis for automatic enumeration of malaria parasites using morphological operations. Elsevier - Expert Systems with Applications. 42, pp. 3041-3047.

[20] Hanung Adi Nugroho, Son Ali Akbar *et al.* 2015. Feature Extraction and Classification for Detection Malaria Parasites in Thin Blood Smear. IEEE- 2nd Int. Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Indonesia.

[21] H. Chiroma, S. Abdul-Kareem *et al.* 2014. Malaria Severity Classification through Jordan-Elman Neural network Based on Features extracted From Thick

Blood Smear. Neural Network World, 5/15, pp. 565-584.

[22] Yazan F, Roumani *et al.* 2013. Classifying highly imbalanced ICU data. Health care Manag Sci. 16, pp. 119-128.

[23] Jia Pengfei, Zhang Chunkai *et al.* 2014. A New Sampling Approach for classification of Imbalanced Data sets with High Density. IEEE transaction. pp. 217-222.

[24] Manoj Gambhir and Chathurika Hettiarachchige. 2017. Making sense of consensus: comparative modelling of malaria interventions. Population Health, IBM Research - Australia. Vol. 5.

[25] Ewan Cameron, Katherine E.Battle *et al.* 2015. Defining the relationship between infection prevalence and clinical incidence of Plasmodium falciparum malaria. Nature Communications. 6: 8170, pp. 1-10.

[26] Daniel Ruiz, Cyrille Brun *et al.* 2014. Testing a multi-malaria-model ensemble against 30 years of data in the Kenyan highlands. Malaria Journal. 13: 206, pp. 1-14.

[27] Raquel M. Goncalves, Kezia K. G. Scopel et al. 2012. Cytokine Balance in Human Malaria: Does Plasmodium vivax elicit More Inflammatory Responses than Plasmodium falciparum? PLUS ONE. 7(9): 1-10.

[28] Haibo He, Edwardo A.Garcia *et al.* 2009. Learning from Imbalanced Data. IEEE Transaction on Knowledge and Data Engineering. 21(9): 1263-1282.

[29] Yanmin Sun, Mohamed S. Kamel *et al.* 2007. Cost-sensitive boosting for classification of imbalanced data. Elsevier-Pattern Recognition. 40, pp. 3358-3378.

[30] Mickel Galar, Alberto Fernandez *et al.* 2011. A Review on Ensemble for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics-PART C: Applications and Reviews. pp. 1-22.

[31] Taghi M. Khoshgoftaar, Jason Van Hulse *et al.* 2011. Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data. IEEE Transactions on Systems, Man, and Cybernetics - PART A: Systems and Humans. 41(3): 552-568.