INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Voting-Boosting: A novel machine learning ensemble for the prediction of Infants' Data

Sourabh [1]*, Vibhakar Mansotra[1], Paramjit Kour[1], Sachin Kumar[1]

1 Department of Computer Science & IT, University of Jammu, Jammu & Kashmir, India

OPEN ACCESS

Received: 01-05-2020

Accepted: 15-06-2020

Published: 24-06-2020

Editor: Dr. Natarajan Gajendran

Citation: S, Mansotra V, Kour P, Kumar S (2020) **Voting-Boosting: A novel machine learning ensemble for the prediction of Infants' Data**. Indian Journal of Science and Technology 13(22): 2189-2202. https://doi.org/10.17485/IJST/v13i22.468

***Corresponding author**.
Sourabh

Department of Computer Science & IT, University of Jammu, Jammu & Kashmir, India

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

Abstract

Background/Objectives: Owing to the continuous increase of electronic records and recent advances in machine learning, various automated disease diagnosis tools have been developed and proposed in healthcare sector. In the present study, an ensemble methodology using voting and boosting techniques has been proposed for optimal selection of features and prediction of infants' data of India. **Methods/Analysis:** For feature selection, the best-first search algorithm of wrapper technique has been used in addition to voting-boosting. The proposed ensemble consists of combination of heterogeneous classifiers including Random Forest, J48, JRip, CART and Stochastic Gradient Descent (SGD). The effectiveness of the proposed ensemble and single classifiers have been investigated in terms of classification accuracy, precision, f-measure, recall, MCC and PRC area using varied k-fold cross validation. **Findings:** The results depicted that the proposed Voting-Boosting ensemble (k=15) outperforms the individual classifiers using selected features. **Applications / Improvements:** The proposed Voting-Boosting ensemble can be extended by using more state-of-the art classification approaches and further utilized for other healthcare datasets for enhancing the performance.

Keywords: Machine learning; ensemble; feature selection; wrapper; voting and boosting

1 Introduction

India is one of the fastest growing economies and second most populous country of the world but poor health outcomes among infants has drawn global attention in health profile of India. Still, there are diverse challenges and shortfalls in terms of healthcare expenditure, vaccination, malnutrition, poor health facilities for newborns and widespread inaccessible geographical locations that need to be addressed urgently. India suffers a large proportion of the disease burden from which Infant Mortality Rate (IMR) is quite a big hurdle for the government. Infant Mortality Rate (IMR) is a standard measure for measuring infants' death per 1,000 live births less than one year of age [(1)]. IMR of India declined

from 81/1000 live births in 1990 to 34/1000 in 2016 and there was a total of 1.08 million deaths of under-5 children in 2016 [2]. India contributed 500,000 i.e. one-third of the global deaths annually and most of these are vaccine preventable deaths [3] and nearly 60 million children are malnourished every year [4]. There is a gradual decline in mortality rates globally but it is critical to regard that this reduction does not occur the same way in all countries. The health of newborns is an important indicator in the assessment of the development of any society and a growing concern at the global level.

Infant mortality has two additional dimensions viz. neonatal mortality rate and post neonatal period. The neonatal mortality rate is the death of newborns during the period from 0 to 27 days of life and this period is the highest risk of dying. The post neonatal period is the death of newborns from 28 days up to completing 1 year of age and expressed per 1,000 live births [5]. The death of newborns mostly happen due to preventable causes which include pre-maturity/preterm, neonatal infections, congenital malformations etc. whereas post-newborn period or children under-5 deaths mostly occur due to infectious diseases like pneumonia, diarrhea, malaria, measles, diarrhea, malnutrition and intrapartum related complications [6]. All these deaths can be preventable through ingenious and affordable intervention. The World Health Organization (WHO) and the United Nations (UN) have coordinated to reduce these mortality rates. They indited eight Millennium Development Goals (MDGs) with measurable targets and clear deadlines in which MDG 4 aimed to reduce child mortality throughout the world by 2015 [7]. In compliance to MDGs, India has witnessed considerable progress but still a lot of work is remaining in some areas. The Government of India has implemented various initiatives both at national and state levels mainly targeting the deprived sections and poor families with the aim of qualitative improvements in standards of public health, healthcare in the rural areas and mainly for the reduction of infant mortality, thereby initiated intensified schemes including [8] Universal Immunization Programme (UIP), Janani Suraksha Yojana (JSY), Janani Shishu Suraksha Karyakaram (JSSK), Rashtriya Bal Swasthya Karyakram (RBSK), Mission Indradhanush (MI), Pradhan Mantri Surakshit Matritva Abhiyan (PMSMA), Intensified Mission Indradhanush (IMI) and many others. Despite of all these initiatives, there is still a need of more flexible approach i.e. implementation of recent technologies for reducing the burden of present Infant Mortality Rate (IMR) in India.

The potentiality of the available data can be exploited only if it can be analyzed and transformed into useful information and in turn is used to generate knowledge to support decision making or development of intelligent automated system for early detection of problems. This study aims at applying multiple data mining and machine learning algorithms in healthcare domain in general and infants' data in particular. The performance of data mining algorithms used in predicting mortality rate is highly efficient with a good combination of salient features and proper implementation of prediction algorithms [9]. The individual classifiers seem to be inapt of ensuring optimal results in terms of prediction and stability. Thus, the effectiveness of heterogeneous ensemble approach exploits the strength of different classifiers at the same time and overcomes the weaknesses of single classifiers [10].

In this study, a new heterogeneous ensemble technique using voting and boosting has been proposed for feature selection and prediction. The removal of weak features is highly desirable for reduction of data dimensionality in any dataset [11,12]. Therefore, the objective of feature selection is to identify the subset of relevant and non-redundant features in the dataset. The selected features contribute towards the final prediction and rejected features will not be used for subsequent modules and analysis. The wrapper method has been applied on full training dataset in the present work. Afterwards, majority voting has been applied as baseline method for wrapping the output of each boosted classifier with k-fold cross-validation.

Furthermore, the remainder of this paper is arranged as follows. The work related to proposed ensemble methodology in healthcare domain was studied and explored in section 2. In the third section, the proposed ensemble methodology, architecture and algorithm based on wrapper feature selection were thoroughly explored. In section 4, working environment and the results were discussed and clearly presented. In fifth section, conclusion, future scope and benefits of the proposed methodology have been elaborated.

2 Related Work

In the recent years, use of data mining and machine learning techniques has been increased to predict the possibility and track of diseases. A numerous number of ensemble methodologies and toolkits have been created, proposed and studied by researchers in the healthcare sector. In this section, a few worth works that are closely related to the proposed ensemble methodology in healthcare sector are presented and the marvelous potential of ensemble techniques is highlighted.

Moreira et al.[13] proposed an ensemble of the nearest neighbor classifiers using the random subspace algorithm which classifies unbalanced pregnancy database. The performance was evaluated by Area under Curve (AUC) and other indicators of the well-known confusion matrix using 10-fold cross-validation method wherein results indicated that Subspace KNN ensemble showed high predictive accuracy of 0.937. This approach predicted the Apgar score, intrauterine growth restriction problems and gestational age during childbirth which can be strongly associated with the neonatal death risk and also predicted fetus-related problems that develop hypertensive disorders in pregnancy. Kabir and Ludwig[14] focused to improve the performance of classification algorithm by using stacked-ensemble technique which finds the optimal weighted average of various learning models. In stacked-ensemble technique, Gradient Boosting Machine (GBM), Random Forest (RF) and Deep Neural Network (DNN) were used as base learners and Generalized Linear Model (GLM) as meta-learner. The results indicated that the stacked ensemble outperformed the individual base learners. Bashir et al.[15] used Bootstrap Aggregation consisting of heterogeneous classifiers namely Naive Bayes, Linear Regression, Quadratic Discriminant Analysis, Instance Based Learner and Support Vector Machine on five different heart disease datasets. The proposed bagging method (BagMOOV) achieved 84.16% accuracy, 93.29% sensitivity, 96.70% specificity and 82.15% f-measure with 10-fold cross-validation. Huang et al.[16] created SVM ensemble based on bagging and boosting over small and large scale breast cancer datasets. The performance of proposed ensemble was evaluated by classification accuracy, ROC, f-measure and computational training time. This approach predicted that SVM ensemble performed slightly better than single SVM classifiers.

Cong et al.[17] created a new selected ensemble method integrated with K-Nearest Neighbor, Support Vector Machine and Naive Bayes in order to diagnose breast cancer using both ultrasound and mammography images. The new indicator R proposed to choose the base classifier for ensemble learning. The proposed new selected ensemble method achieved an accuracy of 88.73% and sensitivity of 97.06% with the evidence that classifier-fusion method was better than the feature-fusion method. Das and Sengur[18] investigated the use of powerful ensemble learning techniques viz. bagging, boosting and random subspace with K-Nearest Neighbors (K-NN), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) as base classifiers. The results indicated that ensemble method was more effective for diagnosing valvular heart disease. Rijn[19] build online performance estimation framework for dynamic data stream that weight the votes of individual classifiers members across the data stream and rely only on Hoeffding trees as base-level classifier. The performance was estimated using two functions based on window and fading factors. The results showed that BLAST with fading factors outperformed than BLAST using the window approach. Santos[20] proposed an ensemble feature ranking by using Information Gain, Gain Ratio, Symmetrical Uncertainty, Chi-Square and other classifiers on breast cancer dataset with the best performance of Naïve Bayes having higher AUC and lower FPR. Khan et al.[21] presented a technique called as Optimal Tree Ensemble (OTE) by integrating trees that were accurate and diverse. The performance of OTE was assessed by applying it on 35 different data sets and compared it with k-nearest neighbor, classification and regression tree, random forest, node harvest and support vector machine. The results revealed that the size of the ensemble was reduced significantly and in most of the cases, better results were obtained. As a subject of corollary, there are analogous proposed methods to uncover the hidden patterns in healthcare domain by broadening the canvas of their literature.

3 Methodology

3.1 Proposed Voting-Boosting Ensemble Model

This section introduces the proposed system architecture, algorithm and methodology which concerns its fundamental design, feature selection design and ensemble methodology for prediction. The aim of consolidating multiple classifiers is to obtain preferable performance as compared to individual classifiers. In this work, two popular ensemble techniques viz. majority voting and adaboosting have been used. The majority voting is an ensemble strategy that selects one of many alternatives based on the predicted classes with the most votes. AdaBoosting is a boosting meta-algorithm which iteratively re-weights based on the training error of the base classifier[22]. The proposed framework consists of two main components viz. Feature Selection module and Model Building & Evaluation module for infants' data prediction**.** The pseudocode of proposed voting-boosting ensemble model algorithm applied in the present research is shown in Algorithm 1.

Algorithm 1: Proposed Voting-Boosting Ensemble with K-fold Cross Validation
INPUT: *Dataset, D*
OUTPUT: *Best model for Infants' Data*
FS = {F1, F2, F3, . . . , Fn } is set of features

1. Feature Selection

   (a) Apply Wrapper method with Best-First search
   (b) Apply Voting and Boosting ensemble technique
   (c) Apply different classification algorithms (Algo1, Algo2, Algo3,..., AlgoN)
   (d) Feature Subset Generation (F1, F2, F3, ,Fm)
   (e) ***If*** obtained desired features then
   (f) Proceed further (Evaluation)
   (g) ***Else***
   (h) Repeat from step **c**.

2. Model Building & Evaluation

   (a) With selected feature subset and K-fold cross-validation by setting the value of k at 5, 10, 15.
   (b) Apply AdaBoosting to each classifier and
   (c) Obtain output {M1, M2, M3, , Mn}
   (d) Apply Majority Voting to Boosted classifier
   (e) Evaluate (Mv using various evaluation measures
   (f) Choose Best Classifier

Output: *Best suited model for Infants' dataset.*

3.1.1 Feature Selection module

The feature selection process attempts to reduce a dataset by removing irrelevant or redundant features to enhance the classifier performance and reduce data noise[23,24] and are categorized mainly into three methods viz. filter, wrapper and embedded methods[25]. The wrapper method uses a classifier to evaluate multiple models by their predictive accuracy (on test data) after statistical re-sampling or cross-validation of the dataset to find the optimal combination that maximizes model performance[26].

In the present study, Wrapper-Voting-Boosting (WVB) feature selection approach has been proposed and the various algorithms and ensemble techniques have been combined with wrapper method in order to prove their

usability in detecting the important features of infants' dataset. Let D_s is the dataset, f_n is the set of feature vectors, t_n is the set of target variables and L = {$Algo_1$, $Algo_2$, $Algo_3$, …, $Algo_n$} is the set of algorithms that have to be applied on dataset to acquire a good performance in the domain of feature selection. In this phase of work, the classical wrapper search algorithm viz. best-first search method has been used. WVB selects 'm' number of relevant features from the 'n' original features. The schematic flow of feature selection is shown in Figure 1. In the process of feature selection, five algorithms viz. CART, J48, JRip, Random Forest and SGD have been used that select different subsets and then yield different results. If desired feature subset is generated, then the process has to be stopped otherwise the other classifiers have to be selected to repeat the process. This process stops at validation procedure. The ensemble feature selection process not only reduces the risk of selecting an unstable subset but also avoids the problem of local optima as the ensemble techniques are usually superior to the single models [27].

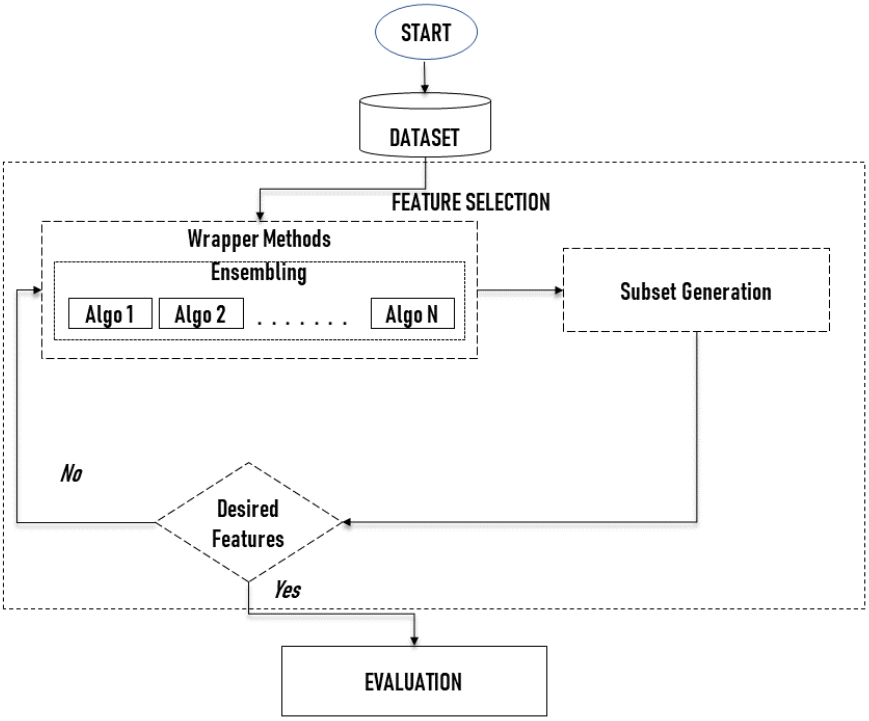

Fig 1. Proposed Wrapper-Voting-Boosting (WVB) ensemble architecture.

3.1.2 Model Building & Evaluation

Ensemble method is a machine learning technique having collection of several classifiers to improve the overall predictive performance [28,29]. The present section of model building & evaluation deals with heterogeneous ensemble called Voting-Boosting (VB) ensemble and its architecture is shown in Figure 2. The VB ensemble is flexible to choose different classification algorithms for selection and prediction of healthcare datasets. The selected subset of 'm' features obtained from WVB has been further used as input for processing of the model. This approach focuses on techniques to enhance the performance of ensemble learning method with different classification algorithms {C_1, C_2, …, C_n}. Each classification algorithm gets boosted {B_1, B_2, …, B_n} and then wrapped using majority voting technique for calculation of final output. The same set of five different classifiers including Random Forest, J48, JRip, CART and SGD has been used. The best classifier has to be evaluated on the basis of various performance measures {M1, M2, …, Mn} with varied k-fold cross-validation.

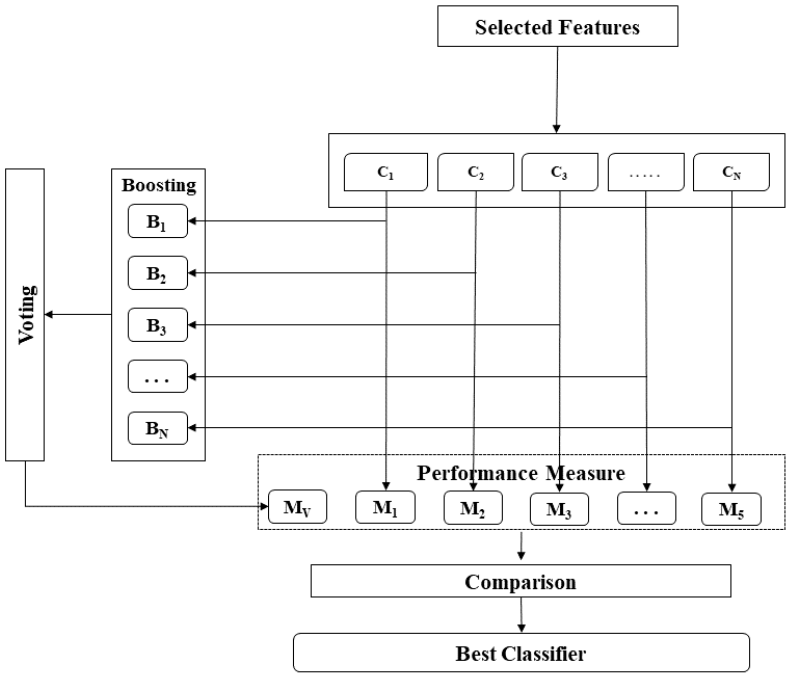

Fig 2. Proposed Voting-Boosting (VB) Ensemble Architecture.

4 Experiments

The experiments have been conducted on infants' dataset with selected subset of attributes. For the classification of data, a class label is required. The class label used for the present work is IMR with two values viz. high and low. The high signifies the districts of India that have the IMR value of 33 or greater whereas low signifies the districts of India that have the IMR value of less than 33. Numerous individual classifiers and the proposed ensemble have been applied on each test set with varied fold of cross validation and their results are analyzed and compared to find out the best model.

4.1 Dataset

The dataset used in this study has been taken from Health Management Information System (HMIS), a portal of Ministry of Health and Family welfare (MoHFW), Government of India from 2014-18. The dataset initially contained 40 features and after applying WVB ensemble with five different classification algorithms viz. CART, J48, JRip, Random Forest and SGD, a subset of 15 features has been extracted and shown in Table 1.

Table 1. Selected Feature Set

S. No.	Features	Code
1.	Child immunization – BCG	BCG
2.	Child immunization - Pentavalent 1	PENTA_1
3.	Child immunization - Pentavalent 2	PENTA_2
4.	Child immunization - OPV 0 (Birth Dose)	OPV_0
5.	Child immunization - Hepatitis-B0 (Birth Dose)	HEP-B0
6.	Child immunization - Hepatitis-B3	HEP_B3
7.	Child immunization - Measles, Mumps, Rubella (MMR) Vaccine	MMR_V
8.	Children more than 5 years received DPT5 (2nd Booster)	DPT5_2B

Continued on next page

Table 1 continued

S. No.	Features	Code
9.	Children more than 10 years received TT10	TT_10
10.	Number of cases of AEFI – Abscess	AEFI_A
11.	Number of cases of AEFI – Death	AEFI_D
12.	Number of cases of AEFI – Others	AEFI_O
13.	Immunization sessions planned	IS_P
14.	Immunization sessions held	IS_H
15.	Number of children more than 16 months of age who received Japanese Encephalitis (JE) vaccine	JE_16M

4.2 Working Environment

The experiments have been carried out on open source software, Waikato Environment for Knowledge Analysis (WEKA) toolkit. WEKA is an aggregation of multifarious machine learning algorithms for data visualization, classification, clustering, regression etc. and is widely used for study, research, implementation, construction or development of new machine learning schemes[30].

4.3 K-fold Cross-Validation

K-fold cross-validation is a common technique used in statistical learning to evaluate the performance of a model or the generalization of a trained model[31]. This protocol is used to partition the dataset into k mutually exclusive partitions as the first subset is used as a validation set for training model on the remaining k–1 subset[32]. The overall performance is obtained by averaging the performance of all k subsets and reduces the bias associated with random selection of samples from each data set[33]. In this study, K-fold cross validation with K= 5, 10 and 15 have been used.

4.4 Evaluation Measures

The performance was evaluated using several standard performance metrics such as:

4.4.1 Accuracy

It refers to the ability of a classifier to measure accurate values i.e. calculates the percentage of correct predictions and mainly used with the cases where the data classes are nearly balanced.

True Positive (TP): Observation which is Positive and is also predicted to be Positive.
False Positive (FP): Observation which is Negative but is predicted to be Positive.
True Negative (TN): Observation which is Negative and is also predicted to be Negative.
False Negative (FN): Observation which is Positive but is predicted to be Negative.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

4.4.2 Precision

It is the ratio of correct positive values and measures how two or more values are closed to each other.

$$Precision = \frac{TP}{TP+FP}$$

4.4.3 Recall

The number of correctly predicted positive values out of the total positive values that are true in that particular class.

$$Recall = \frac{TP}{TP+FN}$$

4.4.4 F-measure

F-measure or F-score is the weighted average of precision and recall i.e. both interpreted together rather than individually.

$$F-measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.4.5 MCC

Matthew's correlation coefficient (MCC) is a well-balanced measure for the quality of binary classifications ranging from -1 (anti-correlation) to +1 (a perfect classifier) with values around 0 corresponding to a random guess[34].

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

4.4.6 PRC

The Precision-recall curve (PRC) shows precision values for corresponding sensitivity (recall) values where recall values plotted on the x-axis and precision values on the y-axis and is determined by the ratio of positives and negatives[35].

5 Results and Discussion

In this study, the models were evaluated based on the accuracy, precision, recall, MCC and PRC discussed above but the prediction accuracy has been considered as the most significant factor. Additionally, the value of batch size is taken as 100. In this section, the performances of individual classifiers and proposed Voting-Boosting ensemble with varied fold of cross-validation are reported. The obtained results are discussed and displayed in the following subsections.

5.1 Results without Feature Selection and Ensembling

This section describes the results obtained by applying individual classifiers i.e. without feature selection and ensembling with varied fold of cross-validation. Table 2 reported the performance of 5 applied classifiers with all evaluation measures. It can be seen that Random Forest has better performance as compared to other classifiers. The highest accuracy for Random Forest has been achieved as 89.31% when k=15.

Table 2. Results obtained without feature selection & ensembling

Classifiers	C-V	Evaluation Measures					
		Accuracy	**Precision**	**Recall**	**F-Measure**	**MCC**	**PRC Area**
Random Forest	K=5	86.79%	87%	86.8%	86.8%	73.8%	94.8%
	K=10	89.02%	89.2%	89%	89%	78.2%	95.4%
	K=15	89.31%	89.60%	89.30%	89.30%	78.90%	95.60%
J48	K=5	82.04%	82.20%	82%	82%	64.20%	79.90%
	K=10	82.34%	82.30%	82.30%	82.30%	64.70%	79.60%
	K=15	83.82%	83.80%	83.80%	83.80%	67.70%	82.70%
CART	K=5	80.80%	80.70%	80.70%	61.50%	80.20%	80.80%
	K=10	83.80%	83.80%	83.80%	83.80%	67.70%	83.30%
	K=15	84.12%	84.10%	84.10%	84.10%	68.30%	81.30%
JRip	K=5	82.07%	82%	82%	82%	64.10%	80%
	K=10	81.15%	81.20%	81.20%	81.20%	62.40%	81.20%

Continued on next page

Table 2 continued

Classifiers	C-V	Evaluation Measures					
	K=15	81.15%	81.20%	81.20%	81.20%	62.30%	81.80%
	K=5	81.60%	82%%	81.60%	81.50%	63.60%	75.90%
SGD	K=10	81.89%	82.20%	81.90%	81.80%	64.10%	76.20%
	K=15	82.19%	82.60%	82.20%	82.10%	64.70%	76.50%

C-V: Cross-Validation

5.2 Results with Feature Selection and Ensembling

The proposed model incorporates well-known ensemble techniques viz. voting and boosting to improve the performance of the traditional classifiers. The results presented in Table 3 were obtained after applying Voting-Boosting ensemble for feature selection and prediction. The results depicted that Voting-Boosting (VB) ensemble outperforms with 90.5% of accuracy at k=15. The proposed model provides noteworthy effectiveness in terms of all applied evaluation measures.

Table 3. Results obtained with feature selection & ensembling

Classifiers	C-V	Evaluation Measures					
		Accuracy	**Precision**	**Recall**	**F-Measure**	**MCC**	**PRC Area**
Random Forest	K=5	87.09%	87.20%	87.10%	87.10%	74.30%	95.10%
	K=10	88.57%	88.70%	88.60%	88.60%	77.20%	95.40%
	K=15	89.02%	89.10%	89%	89%	78.10%	95.60%
J48	K=5	80.71	80.80%	80.70%	80.70%	61.50%	78.60%
	K=10	82.04%	82.10%	82%	82%	64.10%	79.90%
	K=15	84.71%	84.80%	84.70%	84.70%	69.50%	83%
CART	K=5	79.37%	79.40%	79.40%	79.40%	58.80%	79.30%
	K=10	83.23%	83.30%	83.20%	83.20%	66.50%	81.80%
	K=15	84.27%	84.30%	84.30%	84.30%	68.60%	82.50%
JRip	K=5	80.56%	80.70%	80.60%	80.60%	61.20%	79.70%
	K=10	81%	81.20%	81%	81%	62.20%	79.20%
	K=15	81.30%	81.50%	81.30%	81.30%	62.90%	82.20%
SGD	K=5	79.97%	80.40%	80%	79.90%	60.30%	74.10%
	K=10	80.71%	81%	80.70%	80.70%	61.70%	74.90%
	K=15	80.26%	80.60%	80.30%	80.20%	60.90%	74.40%
VB Ensemble	K=5	87.38%	87.50%	87.40%	87.40%	74.90%	94.90%
	K=10	89.02%	89%	89%	89%	78.10%	95%
	K=15	90.50%	90.60%	90.50%	90.50%	81.10%	95.70%

C-V: Cross-Validation

5.3 Overall Comparison

The overall comparison of accuracy, precision, recall, f-measure, MCC and PRC area was compared without feature selection & ensembling and with feature selection & ensembling at varied K-fold cross-validation which is reflected in bar graph viz. Figures 3, 4, 5, 6, 7 and 8. According to the results, the proposed Voting-Boosting ensemble outperformed the traditional classification techniques. The best results achieved by Voting-Boosting ensemble indicated

an accuracy rate of 90.50% at K=15 cross-validation which was found to be greater than individual classifiers. The remarkable results with precision of 90.60%, recall of 90.50%, F-measure of 90.50%, MCC of 81.10% and PRC area of 95.70% using Voting-Boosting (VB) ensemble with K=15 cross-validation has been achieved.

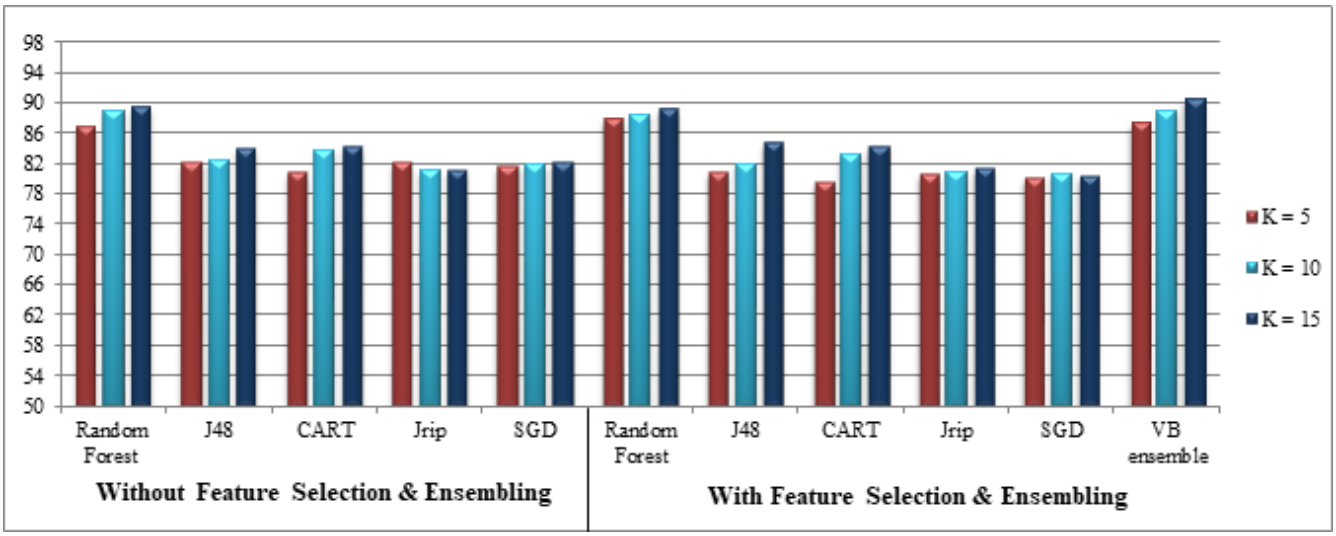

Fig 3. Overall Accuracy Comparison.

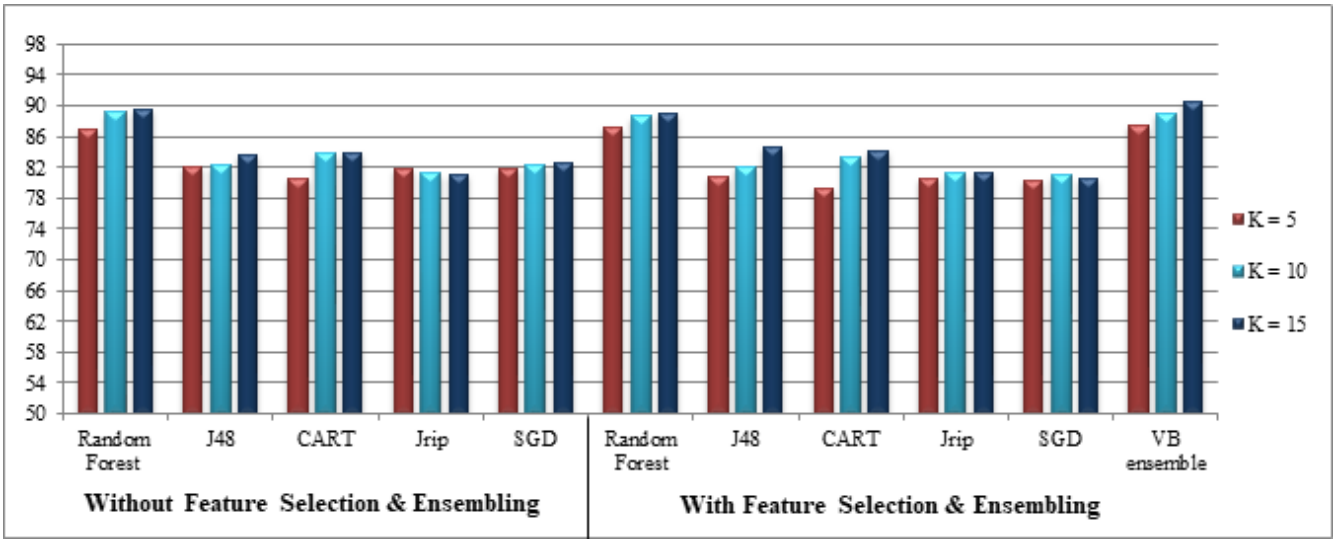

Fig 4. Overall Precision Comparison.

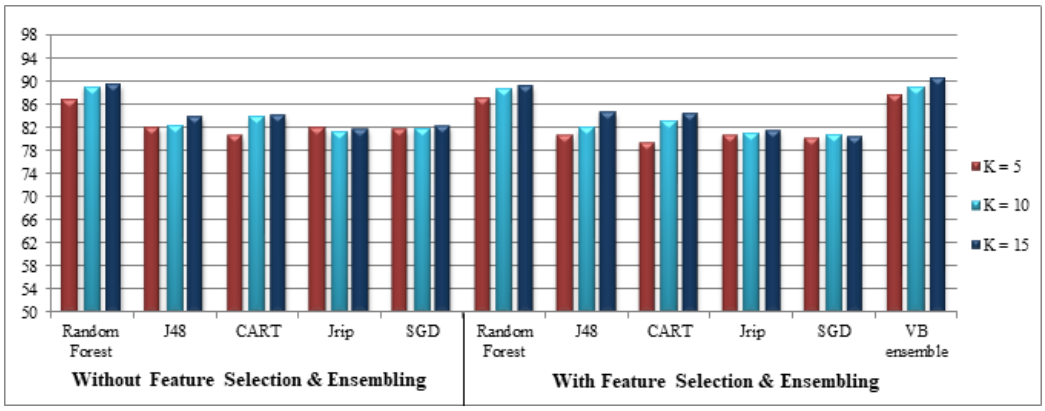

Fig 5. Overall Recall Comparison.

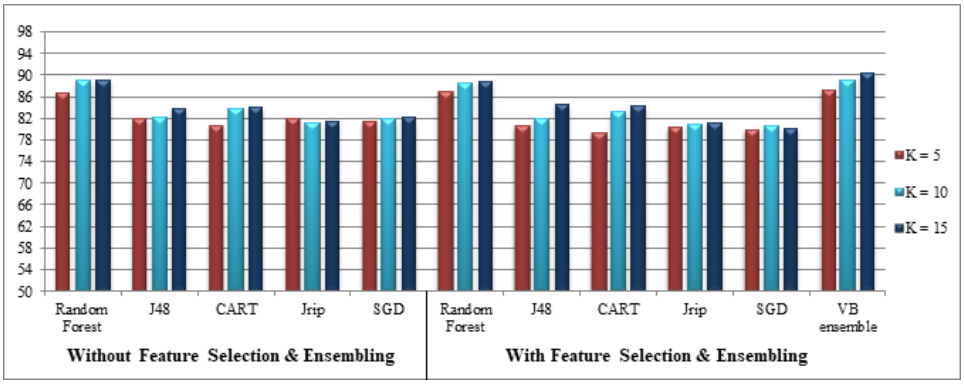

Fig 6. Overall F-Measure Comparison.

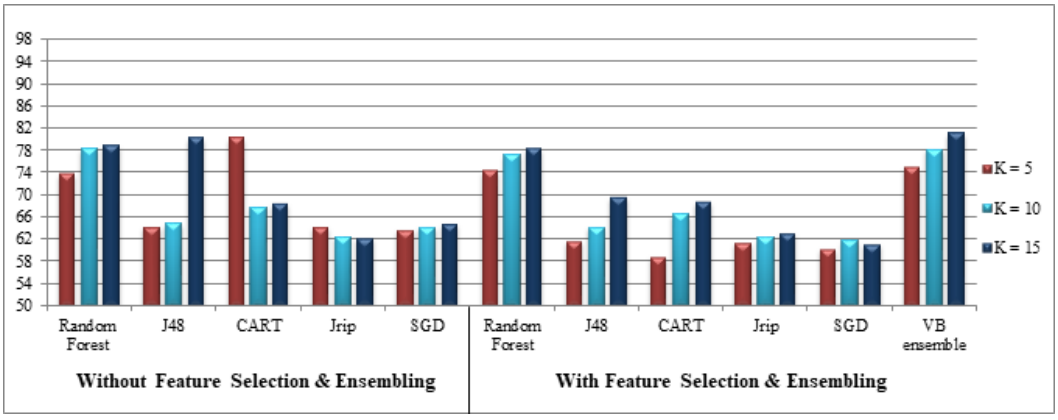

Fig 7. Overall MCC Comparison.

6 Conclusion

With the rapid development of technologies, experts from various fields are working for the well-being of the society by investigating electronic health records. Gigantic amount of data gets analyzed by data mining and machine

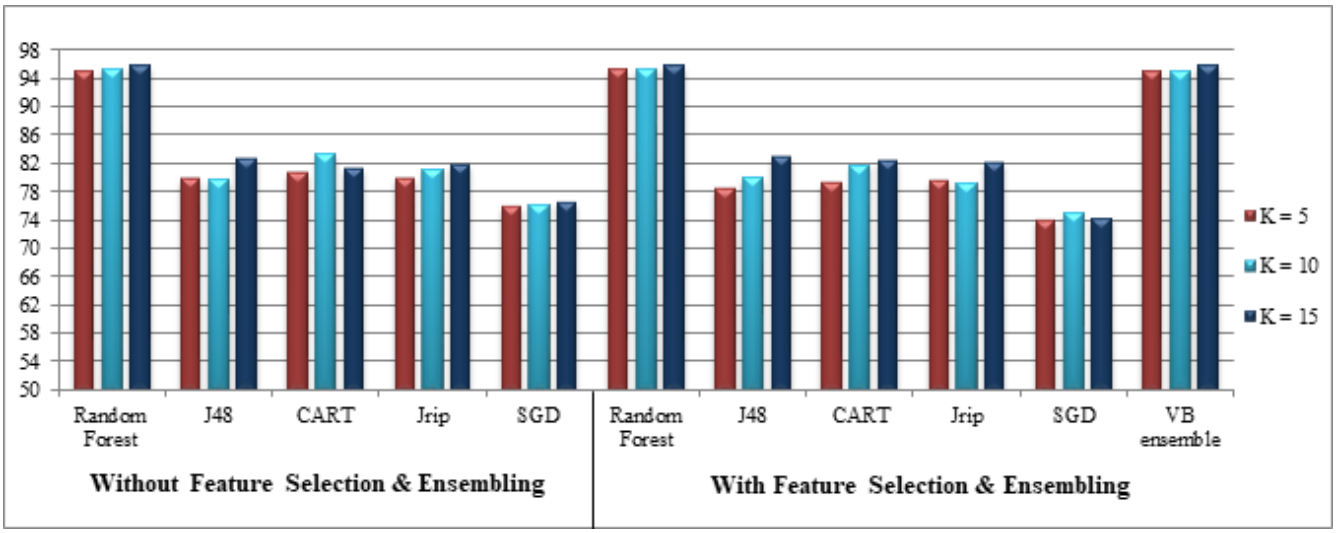

Fig 8. Overall PRC Area Comparison.

learning techniques and various new methodologies and automated systems have been developed. The exploitation of ensemble technique in the field of healthcare sector plays a vital role for disease prediction and classification. A novel ensemble technique equipped with a good feature selection function contributes effectively to classification and prediction performance. The wrapper method with best-first search algorithm finds the optimal subset of features for enhancing the predictability of model. The proposed ensemble model smites the limitations of conventional data mining techniques by employing the ensemble of five heterogeneous classifiers viz. Random Forest, J48, JRip, CART and SGD. The reliability of the system was evaluated by computing different parameters including accuracy, precision, recall, f-measure, MCC and PRC area. The proposed Voting-Boosting (VB) ensemble achieved noteworthy effectiveness based on accuracy, precision, recall, f-measure, MCC and PRC area by varying K-fold cross-validation by working on best feature subset obtained after feature selection process. Thus, from this study, it has been educed that Voting-Boosting ensemble was best suited model for classifying the infants' data into different categories. In future, the focus shall be on utilizing other techniques such as bagging, stacked generalization, blending, random-subspace and other ensemble techniques to improve the performance of the present work. In addition, the same methodology shall also be implemented on other datasets related to healthcare with diverse attributes to confirm the robustness of Voting-Boosting ensemble.

References

1) Deb AK, Dutta S, Hnichho C, Vanlalpeki M, Phosa HT, Rakhu K, et al. A case control study investigating factors associated with high infant death in Saiha district of Mizoram, India bordering Myanmar. *BMC Pediatrics*. 2017;17(1):1–9. Available from: https://dx.doi.org/10.1186/s12887-017-0778-z.
2) Bhatia M, Dwivedi LK, Ranjan M. Priyanka Dixit, Venkata Putcha. Trends, Patterns and Predictive Factors of Infant and Child Mortality in Well-Performing and Underperforming States of India: A Secondary Analysis Using National Family Health Surveys. *BMJ Open*. 2019;9(3).
3) Francis MR, Nohynek H, Larson H, Balraj V, Mohan VR, Kang G, et al. Factors associated with routine childhood vaccine uptake and reasons for non-vaccination in India: 1998–2008. *Vaccine*. 2018;36(44):6559–6566. Available from: https://dx.doi.org/10.1016/j.vaccine.2017.08.026.
4) Kumar C, Singh PK, Rai RK. Under-Five Mortality in High Focus States in India: A District Level Geospatial Analysis. *PLoS ONE*. 2012;7(5):e37515–e37515. Available from: https://dx.doi.org/10.1371/journal.pone.0037515.
5) Guerra AB, Guerra LM, Probst LF, Gondinho BVC, Ambrosano GMB, Melo EA, et al. Can the primary health care model affect the determinants of neonatal, post-neonatal and maternal mortality? A study from Brazil. *BMC Health Services Research*. 2019;19(1). Available from: https://dx.doi.org/10.1186/s12913-019-3953-0.
6) Lassi ZS, Mallick D, Das JK, Mal L, Salam RA, Bhutta ZA. Essential interventions for child health. *Reproductive Health*. 2014;11(Suppl

1):S4–S4. Available from: https://dx.doi.org/10.1186/1742-4755-11-s1-s4.
7) O'Hare B, Makuta I. An analysis of the potential for achieving the fourth millennium development goal in SSA with domestic resources. *Globalization and Health*. 2015;11(1):8–8. Available from: https://dx.doi.org/10.1186/s12992-015-0092-1.
8) Shrivastwa N, Gillespie BW, Kolenic GE, Lepkowski JM, Boulton ML. Predictors of Vaccination in India for Children Aged 12–36 Months. *American Journal of Preventive Medicine*. 2015;49(6):S435–S444. Available from: https://dx.doi.org/10.1016/j.amepre.2015.05.008. doi:10.1016/j.amepre.2015.05.008.
9) Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*. 2019;36:82–93. Available from: https://dx.doi.org/10.1016/j.tele.2018.11.007.
10) Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Computing and Applications*. 2019. Available from: https://doi.org/10.1007/s00521-019-04082-3.
11) Panthong R, Srivihok A. Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm. *Procedia Computer Science*. 2015;72:162–169. Available from: https://dx.doi.org/10.1016/j.procs.2015.12.117.
12) Masoudi-Sobhanzadeh Y, Motieghader H, Masoudi-Nejad A. FeatureSelect: a software for feature selection based on machine learning approaches. *BMC Bioinformatics*. 2019;20(1). Available from: https://dx.doi.org/10.1186/s12859-019-2754-0.
13) Mario WL, Moreira J, Rodrigues G, Marcondes AJV, Neto V, Furtado. 2018. Available from: https://doi.org/10.5753/sbcas.2018.3671.
14) Kabir MF, Ludwig AS. Enhancing the Performance of Classification Using Super Learning. *Data-Enabled Discovery and Applications*. 2019;3(1). Available from: https://dx.doi.org/10.1007/s41688-019-0030-0.
15) Bashir S, Qamar U, Khan FH. BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. *Australasian Physical & Engineering Sciences in Medicine*. 2015;38:305–323. Available from: https://dx.doi.org/10.1007/s13246-015-0337-6.
16) Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. *PLOS ONE*. 2017;12(1):e0161501–e0161501. Available from: https://dx.doi.org/10.1371/journal.pone.0161501.
17) Cong J, Wei B, He Y, Yin Y, Zheng Y. A Selective Ensemble Classification Method Combining Mammography Images with Ultrasound Images for Breast Cancer Diagnosis. *Computational and Mathematical Methods in Medicine*. 2017. Available from: https://doi.org/10.1155/2017/4896386.
18) Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Systems with Applications*. 2010;37(7):5110–5115. Available from: https://dx.doi.org/10.1016/j.eswa.2009.12.085.
19) van Rijn JN, Holmes G, Pfahringer B, Vanschoren J. The online performance estimation framework: heterogeneous ensemble learning for data streams. *Machine Learning*. 2018;107(1):149–176. Available from: https://dx.doi.org/10.1007/s10994-017-5686-9.
20) Santos V, Datia N, Pato MPM. Ensemble Feature Ranking Applied to Medical Data. *Procedia Technology*. 2014;17:223–230. Available from: https://dx.doi.org/10.1016/j.protcy.2014.10.232.
21) Khan Z, Gul A, Perperoglou A, Miftahuddin M, Mahmoud O, Adler W, et al. Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*. 2020;14:97–116. Available from: https://dx.doi.org/10.1007/s11634-019-00364-9.
22) Large J, Lines J, Bagnall A. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Mining and Knowledge Discovery*. 2019;33(6):1674–1709. Available from: https://dx.doi.org/10.1007/s10618-019-00638-y.
23) Boughaci D, shuayree Alkhawaldeh AA. Three local search-based methods for feature selection in credit scoring. *Vietnam Journal of Computer Science*. 2018;5(2):107–121. Available from: https://dx.doi.org/10.1007/s40595-018-0107-y.
24) Rajab DK. New Hybrid Features Selection Method: A Case Study on Websites Phishing. *Hindawi Security and Communication Networks*. 2017. Available from: https://doi.org/10.1155/2017/9838169.
25) Miao J, Niu L. A Survey on Feature Selection. *Procedia Computer Science*. 2016;91:919–926. Available from: https://dx.doi.org/10.1016/j.procs.2016.07.111.
26) Abusamra H. A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma. *Procedia Computer Science*. 2013;23:5–14. Available from: https://dx.doi.org/10.1016/j.procs.2013.10.003.
27) Zhao S, Zhang Y, Xu H, Han T. 2019. Available from: https://doi.org/10.1155/2019/4318463.
28) Adnan OM, Abuassba D, Zhang X, Luo A, Shaheryar H, Ali. 2017. Available from: https://doi.org/10.1155/2017/3405463.
29) Nagi S, Bhattacharyya DK. Classification of microarray cancer data using ensemble approach. *Network Modeling Analysis in Health Informatics and Bioinformatics*. 2013;2(3):159–173. Available from: https://dx.doi.org/10.1007/s13721-013-0034-x. doi:10.1007/s13721-013-0034-x.
30) Ma H, Xu CF, Shen Z, Yu CH, Li YM. Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China. *Hindawi BioMed Research International*. 2018. Available from: https://doi.org/10.1155/2018/4304376.
31) Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*. 2018;10:100–107. Available from: https://dx.doi.org/10.1016/j.imu.2017.12.006.
32) Cawley GC, Talbot NLC. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*. 2008;71(2-3):243–264. Available from: https://dx.doi.org/10.1007/s10994-008-5055-9.
33) Bashir S, Qamar U, Khan FH. A Multicriteria Weighted Vote-Based Classifier Ensemble for Heart Disease Prediction. *Computational Intelligence*. 2015. Available from: https://doi.org/10.1111/coin.12070.

34) Pu L, Naderi M, Liu T, Wu HC, Mukhopadhyay S, Brylinski M. eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology and Toxicology*. 2019;20(1). Available from: https://dx.doi.org/10.1186/s40360-018-0282-6.
35) Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. 2015;10(3):e0118432–e0118432. Available from: https://dx.doi.org/10.1371/journal.pone.0118432.