Contents lists available at ScienceDirect

International Journal of Cognitive Computing in Engineering

journal homepage: www.elsevier.com/locate/ijcce

An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier

Saloni Kumari[a], Deepika Kumar[b,*], Mamta Mittal[c]

[a] Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[b] Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[c] Department of Computer Science & Engineering, G B Pant Govt. College of Engineering, New Delhi, 110020, India

ARTICLE INFO

Keywords:
Diabetes
Machine Learning
ensemble
soft voting
classification

ABSTRACT

Diabetes is a dreadful disease identified by escalated levels of glucose in the blood. Machine learning algorithms help in identification and prediction of diabetes at an early stage. The main objective of this study is to predict diabetes mellitus with better accuracy using an ensemble of machine learning algorithms. The Pima Indians Diabetes dataset has been considered for experimentation, which gathers details of patients with and without having diabetes. The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms viz. random forest, logistic regression, and Naive Bayes for the classification. Empirical evaluation of the proposed methodology has been conducted with state-of-the-art methodologies and base classifiers such as AdaBoost, Logistic Regression,Support Vector machine, Random forest, Naïve Bayes, Bagging, GradientBoost, XGBoost, CatBoost. by taking accuracy, precision, recall, F1-score as the evaluation criteria. The proposed ensemble approach gives the highest accuracy, precision, recall, and F1_score value with 79.04%, 73.48%, 71.45% and 80.6% respectively on the PIMA diabetes dataset. Further, the efficiency of the proposed methodology has also been compared and analysed with breast cancer dataset. The proposed ensemble soft voting classifier has given 97.02% accuracy on the breast cancer dataset.

1. Introduction

Diabetes is commonly referred to as diabetes mellitus by doctors and health professionals. It's a state where the body is unable to make blood glucose commonly known as blood sugar (Joshi & Alehegn, 2017). Diabetes attacks lots of people worldwide and is generally divided into Type1 and Type2 diabetes (Ndisang, Vannacci & Rastogi, 2017). Type1 diabetes is also called insulin-dependent diabetes most often begins in childhood. In type1, the pancreas with antibodies is attacked by the body, after that it destroys internal body parts and stops making insulin. Type2 is also known as adult-onset diabetes or non-insulin-dependent. it is mostly lenient than type1, but it is still very harmful and cause dangerous complication, specifically in the small blood vessels in your eyes, kidney, and nerves (Himsworth & Kerr, 1939). In 2014, the World Health Organization (WHO) reported globally that adult diabetes patients have nearly doubled since 1980, rising from 4.7% to 8.5%. In 2012, 1.5 million people died due to diabetes (Standl, Khunti, Hansen & Schnell, 2019). A WHO survey among pregnant women states that 2 to 17.8% have gestational diabetes. Diabetes mellitus is one of the primary concerns in medical science research because of the extreme social effect of the specific disease, which inevitably generates huge amounts of data. Undoubtedly, therefore, when it comes to diagnosis, management and other associated clinical administration aspects, machine learning and data mining approaches in diabetes mellitus are of great concern. . Various methods have been developed in the context of this study and therefore, an ensemble approach has been proposed using machine learning and data mining for classification of diabetes. Obesity and Gestational diabetes mellitus are obstructions which obtain during pregnancy and eventually affect the generation of the baby during postnatal and fatal life (World Health Organization, 2013). Non-pregnant female individuals have different criteria for diabetes recognition based on the interconnection between the risk of diabetes microvascular obstruction and plasma glucose quantity (Kandhasamy & Balamurali, 2015). Diabetes is common in people's daily life due to the development of living standards, the early identification and determination of diabetes is the only way to stay away from its complications (Vijayan & Anjali, 2015).

Lots of research has been done in disease prediction such as diagnosis, prediction, classification, therapy etc. Recent research shows that various ML (Machine Learning) algorithms have been used for disease identification and prediction. They have resulted in remarkable efficiency and improvement in profound conventional and ML methods (Goyal, Malik, Kumar, Rathore & Arora, 2020; Kumar et al., 2020;

[*] Corresponding author.
E-mail address: deepika.kumar@bharatividyapeeth.edu (D. Kumar).

https://doi.org/10.1016/j.ijcce.2021.01.001
Received 23 November 2020; Received in revised form 27 January 2021; Accepted 28 January 2021
Available online 30 January 2021

Mittal, Arora, Pandey & Goyal, 2020). ML has shown their abilities to efficiently and strongly deal with high numbers of variables while making strong predictive models. Supervised ML methods focus to examine the dependent term in the form of the independent terms/variables. Predictive modeling is extensively applied in many data mining and healthcare sectors such as brain tumor detection and salient object detection, In general, 85% of those were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules (Chen et al., 2016; Kaur et al., 2018; Kavakiotis et al., 2017; Kumar & Batra, 2018, 2020, 2021; Mittal et al., 2019; Mittal, Kaur, Pandey, Verma & Goyal, 2019; Sudharsan, Peeples & Shomali, 2014). It's an important mechanism to convert biomedical datasets into workable information, high-level research in clinical, and enhances healthcare. The need for classification of diabetes patients as mentioned above led to many advances in ML techniques. The National institute of diabetes and digestive and kidney diseases has initially provided this Pima Indian diabetes dataset. It has 769 data points in which 500 are diabetes negative and 268 are diabetes positive (Pima Indians Diabetes dataset May 2008). The research history on this dataset shows that various machine learning algorithms and ensemble approaches have been applied for the classification of disease but none of them able to achieve an accuracy of more than 76% (Fatima & Pasha, 2017; Grudzinski, 2008; Husain & Khan, 2018). Therefore, we proposed an ensemble approach to improvise results. This research mainly focuses on the performances, theoretical, and properties of learning models and algorithm methods. The classification approach has been utilized rather than regression for the prediction of disease. The ensemble approach has been proposed using a soft voting classifier to classify diabetes. The logistic regression Naïve Bayes and Random Forest algorithms have been ensembled and the performance of the ensemble approach has achieved better results as compared to base classifiers. The performance of the aforementioned algorithm has been evaluated by taking accuracy, precision, recall, F1-score as evaluation criteria.

The organization of the paper is as follows: Section2 discusses the literature done in this area with a detailed discussion of various machine learning and ensemble approaches. Section3 explains the proposed methodology where a soft voting classifier has been used with an ensemble of three ML algorithms viz. Naïve Bayes, Random forest, and Logistic Regression. Section 4 discusses the results and analysis of the proposed methodology and the results of the proposed methodology have been compared and analysed with conventional ML algorithms and state-of-art methodologies.

2. Related work

Remarkable amounts of research in the context of recognition of diabetes patients using machine learning models and data mining operations have been embossed excitedly in recent years. Researchers used the ensemble technique in which several single models are combined to give better prediction results. In 2014, Vijayan et al. used various data mining techniques for diabetes mellitus (Vijayan & Ravikumar, 2014). In 2017, the author shared the importance of AdaBoost and bagging techniques of machine learning using J48 as the foundation for diabetes prediction. It impressively classifies diabetic and non-diabetic patients based on risk factors of diabetes. It was observed that the AdaBoost learning algorithm outshines than bagging and J48 algorithm (M. Fatima & Pasha, 2017). Smith, Everhart, Dickson, Knowler & Johannes (1988) proposed a neural network ADAP algorithm to build an associative model in which they randomly selected data for training and the accuracy achieved was 76%. Quinlan (1993) used a C4.5 learning model and the model performed well with an accuracy of 71.1%. Naive Bayes, J48, Radial basis function, Artificial neural network had been used for diabetes type 2 diagnosis. Naive Bayes achieved an accuracy of 76.95% and outperformed results of J48, RBF, with accuracies of 76.52%, 74.34% respectively (Nai-Arun & Moungmai, 2015). In 2015, Nongyao et al. proposed a model that speculates the risk of diabetes. The methodology used four different machine learning algorithms for the classification purpose: ANN, DT, LR, and NB. Bagging and boosting have been used for improvisation in results. The experiments demonstrate that the random forest algorithm outperformed as compared to other algorithms (Soltani & Jafarian, 2016). Sahan et al. used a 10- fold cross-validation method with a weighted artificial immune system and got a prediction accuracy of 75.87% (Sahan, Kodaz & Gunes, 2005). Anand et al. used the CART model and the algorithm an achieved accuracy of 75% (Anand & Shakti, 2015). Rani and Jyothi (2016) proposed ensemble algorithms that use the ANN, NB, KNN, zeroR, J48, simple cart, filtered classifier, and cv parameter selection for the classification. The proposed methodology achieved an accuracy of 77.01%. Li L. proposed a methodology with SVM, ANN, Naïve Bayes, and a weighted-based study for the classification (Li, 2014).Bashir et al. proposed a ensembled model of CART, ID,3, and C4.5 which achieved an accuracy of 76.5% (Bashir, Qamar, Khan & Javed, 2017).

The research demonstrates the enhanced approach using ensemble of three machine learning algorithms using soft voting classifier. To accomplish the goals, results have been compared and analysed with breast cancer dataset. The research objective have been summarised as follows:

- An ensemble of machine learning algorithms viz. random forest, logistic regression, and Naïve Bayes with soft voting classifier have been proposed. The proposed methodology binary classifies the diabetes mellitus disease data into positive and negative classes.
- Experiments have been conducted on two different datasets viz. PIMA diabetes and breast cancer dataset.
- Accuracy, precision, recall, F1-score, AUC curve have been taken as the evaluation criteria for testing the robustness of proposed methodology.
- A comparison with existing techniques of the proposed methodology reveals superior results with the same number of defined parameters.
- Empirical evaluation of the proposed methodology has been conducted with conventional base classifiers such as AdaBoost, Logistic Regression,Support Vector Machine, Random forest, Naïve Bayes, Bagging, GradientBoost, XGBoost, CatBoost.

3. Proposed methodology

This research extensively works upon improving the results and accuracy of diabetes detection. Authors have proposed an ensemble of machine learning algorithms by using a soft voting classifier for the binary classification of disease into positive and negative. The data preprocessing has been applied before giving input to the model, which is followed by data augmentation. The flow diagram of proposed ensemble approach using soft voting classifier is comprehended in Fig. 1.

3.1. Data description

In this research work, the Pima Indian Dataset has been considered for the experimentation (Pima Indians Diabetes dataset May 2008). The dataset has 9 columns and one output column with a dichotomous value to specify if the person has diabetes positive or diabetes negative. It has 768 rows, in which 500 patients are non-diabetics while the other 268 are diabetic patients. There are nine feature columns in the dataset such as pregnancy month, glucose, plasma, BP, fold thickness of triceps skin, quantity of insulin, BMI, Pedigree function, Age of patients, and one target column (0 or 1).

3.2. Data pre-processing

Data Pre-processing is an important step that is used to transform the data in a useful and efficient format so that it can be fed to the machine learning algorithm. The first technique used for data pre-processing is data normalization. This technique is used to perform linear transformation of data. It is also called Min-max normalization, where all the

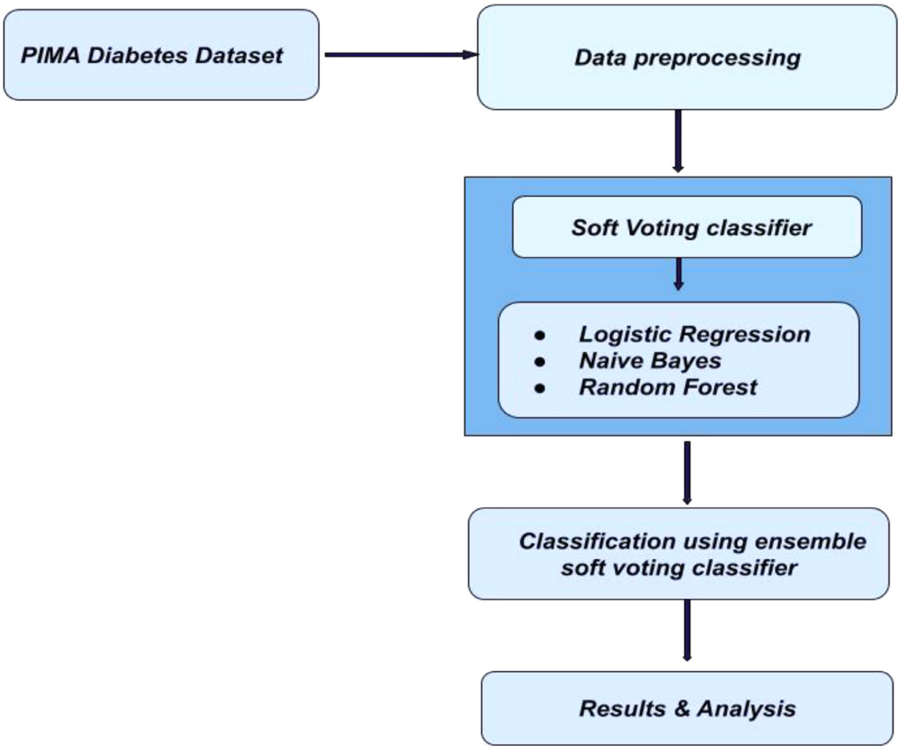


Fig. 1. The flow diagram of proposed ensemble approach using soft voting classifier.

values of the attributes ranges between [0,1]. The next pre-processing technique used is label encoding. This technique is applied to the dependent variable i.e. the person is having diabetes or not. So all the string values in the output variable are replaced by 0 and 1 determining the output class. There were many missing values for various attributes in the dataset i.e. pregnancy attribute has 111, insulin attribute has 374, and skin thickness attribute has 227 missing values. All the missing values were replaced by the median for a particular attribute. This technique of data-pre-processing is also called replacement by median.

3.3. Model architecture

In this proposed methodology, we have used the ensemble of machine learning algorithms such as Logistic Regression, Naive Bayes, and Random Forest classifiers. The above-mentioned algorithms have been ensembled with a soft voting classifier to enhance accuracy. These algorithms are briefly discussed in this section.

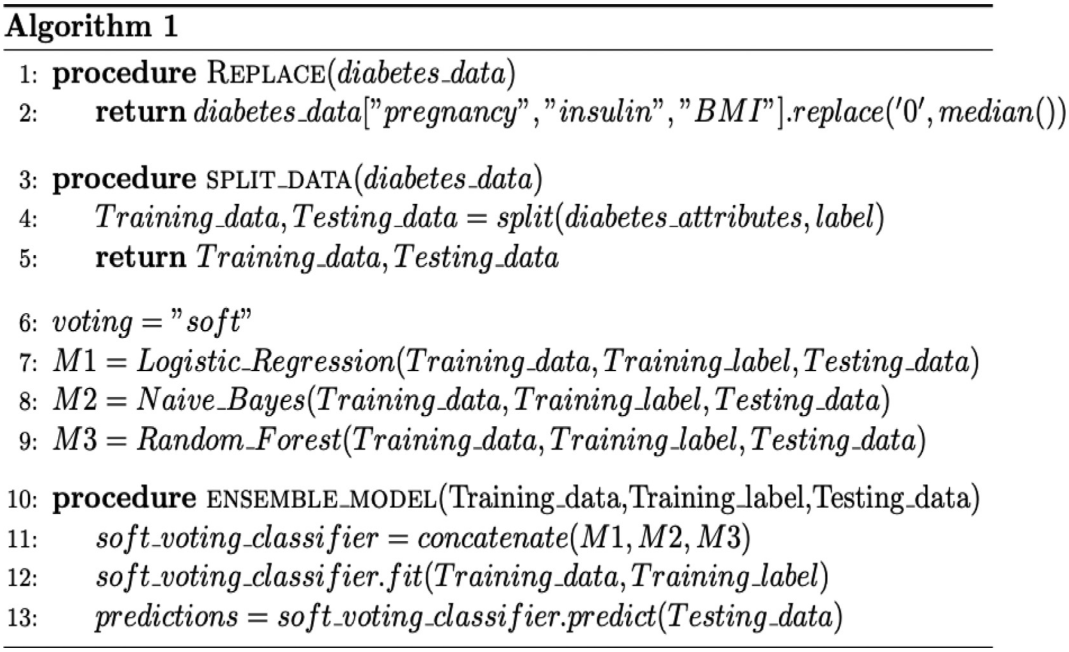
- **Logistic regression**: Statistical technique is used in logistic regression to speculate dichotomous results ($y = 0$ or 1). Logistic regression is a linear learning algorithm. The predictions of logistic regression are done in terms of probabilities of an event occurring. LR algorithm maps each data point using the sigmoid function. The standard logistic function gives an S- shaped curve. Eq. (1) shows the sigmoid function (Kleinbaum, Dietz, Gail, Klein & Klein, 2002; Pregibon, 1981).

$$\text{sigmoidequation} = \frac{1}{1 + e\hat{}(-x)} \tag{1}$$

- **Naive Bayes:** This classifier is based on a probabilistic method. It uses Bayes' theorem with the assumption that the existence of one feature in a given class is not related to the existence of another feature in the same class. Joint probabilities of categories and terms are used to speculate the probabilities of given categories. Due to this independence assumption, the parameters for every term can be studied individually, hence it accelerates computation operations. The Bayesian network consists of a structural model and a set of conditional probabilities (Jiang, Zhang & Cai, 2009).
- **Random Forest Classifier:** This classifier is a type of ensemble classifier. It uses decision tree models to get better prediction results. It develops many trees and a bootstrap technique is applied to every tree from the set of training data. In classification, procedure input is given to each tree present in the forest, and then, each tree votes individually for that class. In the end, the RF selects the class, which has got the greatest number of votes (Pal, 2005).
- **Proposed Ensemble soft voting classifier**: This classifier is a meta-classifier for merging same or conceptually dissimilar machine learning models for prediction through majority voting. A voting classifier uses two types of voting techniques, hard and soft. In hard voting, the final prediction is done by a majority vote in which the aggregator selects the class prediction that comes again and again among the base models. In soft voting, base models should have the Predict_proba method. The voting classifier presents better overall results than other base models, as it combines the predictions of different models. In the proposed model, Logistic Regression, Naive Bayes,and Random Forest classifier have been ensembled. A soft voting classifier has been utilized which uses the predict_proba attribute column that gives the probability of each target variable[42]. Then it shuffles training data & data points, and these data points are passed to logistic regression, Naïve Bayes, and Random Forest model. Each model calculates individual prediction with voting aggregator and soft voting technique, the majority voting is computed which yields the final prediction. The algorithm for the proposed methodology has been illustrated in Fig. 2.

4. Result and analysis

The proposed methodology uses ensemble of three machine learning models viz. Random Forest, Logistic Regression, Naïve Bayes with soft voting classifier. Experimentation has been conducted using PIMA diabetes dataset. The dataset has 769 data points and 10 feature columns where zero has been replaced with their median values. The dataset has been divided into testing and training datasets with 20% and 70% respectively. Accuracy, Precision, Recall, F1 score are the most common evaluation metrics adopted for checking the robustness and efficiency



Algorithm 1

```
procedure REPLACE($diabetes\_data$)
    return $diabetes\_data["pregnancy", "insulin", "BMI"].replace('0', median())$

procedure SPLIT_DATA($diabetes\_data$)
    $Training\_data, Testing\_data = split(diabetes\_attributes, label)$
    return $Training\_data, Testing\_data$

$voting = "soft"$
$M1 = Logistic\_Regression(Training\_data, Training\_label, Testing\_data)$
$M2 = Naive\_Bayes(Training\_data, Training\_label, Testing\_data)$
$M3 = Random\_Forest(Training\_data, Training\_label, Testing\_data)$

procedure ENSEMBLE_MODEL(Training_data,Training_label,Testing_data)
    $soft\_voting\_classifier = concatenate(M1, M2, M3)$
    $soft\_voting\_classifier.fit(Training\_data, Training\_label)$
    $predictions = soft\_voting\_classifier.predict(Testing\_data)$
```

Fig. 2. Algorithm for proposed ensemble soft voting Classifier.

Table 1
Comparison of various Machine Learning models (PIMA dataset).

Algorithms	Accuracy	Precision	F1_score	Recall	AUC value
Logistic Reg	74.89%	64.47%	62.82%	61.25%	80.10%
K-Nearest	71.92%	58.33%	59.75%	61.25%	66.31%
Support Vector	74.02%	67.24%	56.52%	48.75%	0.0%
Naive Bayes	74.12%	61.90%	63.41%	65%	79.01%
Decision Tree	71.42%	58.13%	60.24%	62.50%	68.94%
Random Forest	77.48%	71.21%	64.38%	58.75%	78.10%
Soft Voting Classifier	**79.08%**	**73.13%**	**71.56%**	**70%**	**80.98%**
AdaBoost	75.32%	68.25%	60.13%	53.75%	74.98%
Bagging	74.89%	62.5%	65.47%	68.75%	70.11%
GradientBoost	75.32%	70.90%	57.77%	48.75%	71.89%
XGBoost	75.75%	64.28%	65.85%	67.50%	69.01%
CatBoost	75.32%	64.19%	64.59%	65.00%	74.56%

of the algorithms. True positive(tp) means when the value of the predicted class is 1 and the value of the actual class also 1. True negative (tn) means the value of the predicted class is 0 and the actual class also 0. false negatives(fn) and false positives (fp) occur when your predicted class controvert the actual class. Accuracy is the most important measure and it is a ratio of total correctly predicted observation to the total number of observations. Precision, Accuracy, Recall, and F1 score can be calculated by the following formulas:

$$Accuracy = \frac{tp + tn}{tn + tp + fp + fn} \tag{2}$$

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

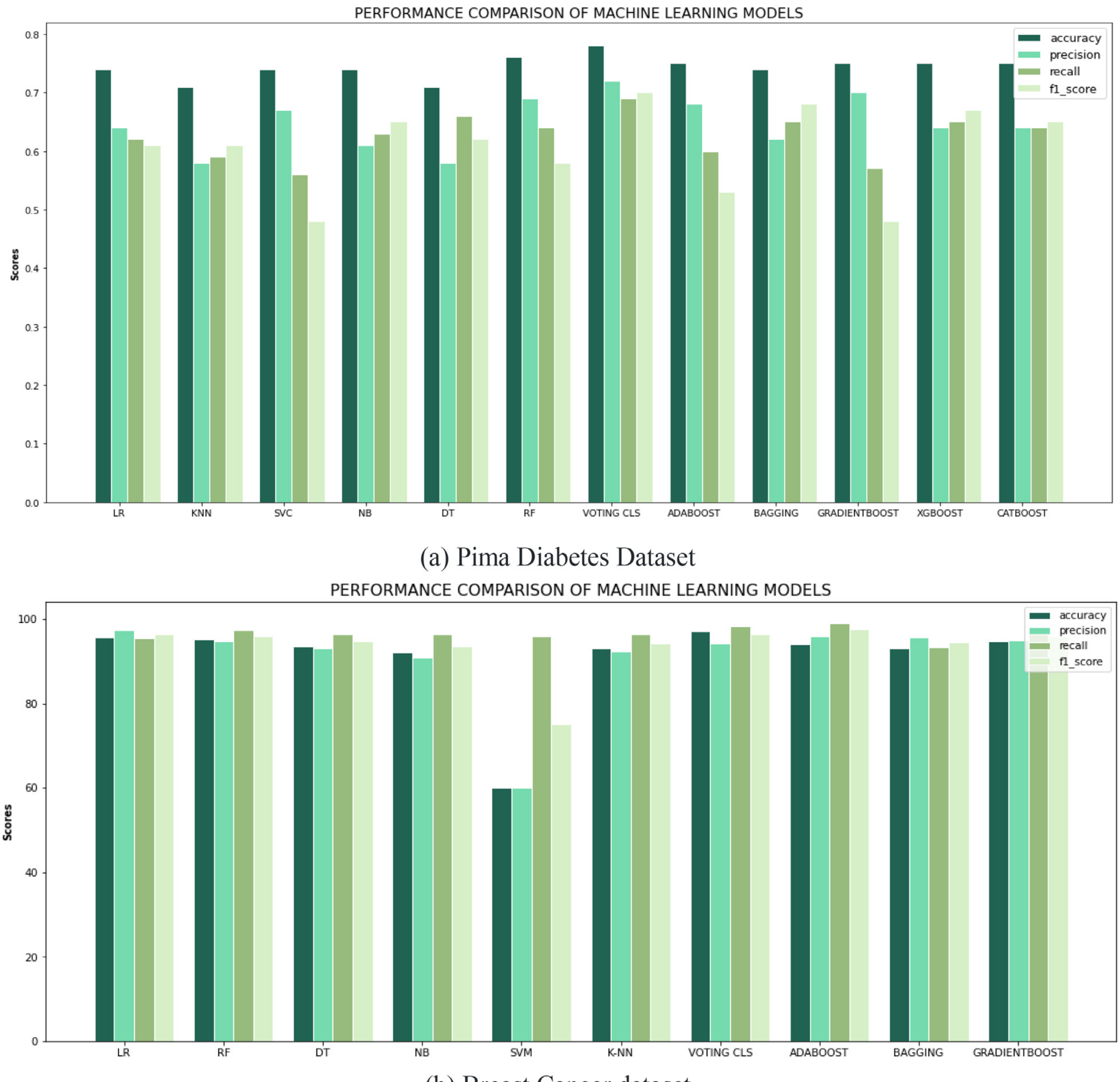
$$\text{F1score} = \frac{2 \times Precision \times recall}{Precision + recall} \tag{5}$$

A comparative analysis of all the conventional machine learning algorithms has been done in this section for diabetes mellitus classification into positive and negative classes. It has been done for comparing and analysing accuracies of all the conventional algorithms. Two datasets have been used for experimentation, one is PIMA diabetes mellitus which have two classes positive and negative dataset. Second dataset is breast cancer dataset which binary classifies the dataset into benign and malignant. Table 1 depicts the comparison of various machine learning models results using PIMA dataset. It can concluded from Table 1, that the ensemble soft voting classifier has achieved maximum Accuracy, Precision, F1 score, Recall, AUC value of 79.08%, 73.13%, 71.56%, 70%, 80.98% respectively as compared to other machine learning algorithms. It can be observed from the table that the KNN and decision tree does not performed good on the given dataset with an accuracy of 71.92%, 71.42%, respectively. If we see the individual performance of each Logistic regression, random forest and Naïve Bayes algorithm, all these algorithms give an accuracy of 74.89%, 77.48%, 74.12%, respectively. The detailed analysis of all the algorithms has been depicted in Table 1:

Breast Cancer is a major reason for dying and set out top among females worldwide (Dash, Acharya, Mittal & Abraham, 2020; Harris, Lippman, Veronesi & Willett, 1992; Montazeri, Montazeri, Montazeri & Beigzadeh, 2016). The proposed methodology has also been compared with breast cancer dataset to validate the robustness of the ensemble approach. The proposed methodology has outperformed on the breast cancer dataset too. It is depicted in Table 2 depicts the comparison of various machine learning algorithms using breast cancer dataset. The ensemble soft voting classifier using breast cancer dataset has achieved maximum Accuracy, Precision, F1 score, Recall, AUC value of 97.27%, 94.40%, 98.33%, 96.32%, 98.87% respectively as compared to other

Table 2
Comparison of various Machine Learning models (Breast cancer dataset).

Algorithms	Accuracy	Precision	Recall	F1_score	AUC value
Logistic Reg	95.74%	97.29%	95.57%	96.428%	95.90%
Random Forest	95.21%	94.82%	97.34%	96.06%	96.80%
Decision Tree	93.61%	93.16%	96.46%	94.78%	98.99%
Naive Bayes	92.02%	90.83%	96.46%	93.56%	95.58%
SVM	60.10%	60.10%	96..10%	75.08%	0.0%
K-NN	93.08%	92.37%	96.46%	94.37%	92.80%
Soft Voting Classifier	**97.27%**	**94.40%**	**98.33%**	**96.32%**	**98.87%**
AdaBoost	94.10%	95.96%	99.16%	97.54%	98.71%
Bagging Class	93.08%	95.72%	93.33%	94.51%	97.45%
GradientBoost	94.68%	95.08%	96.66%	95.86%	97.45%



(a) Pima Diabetes Dataset

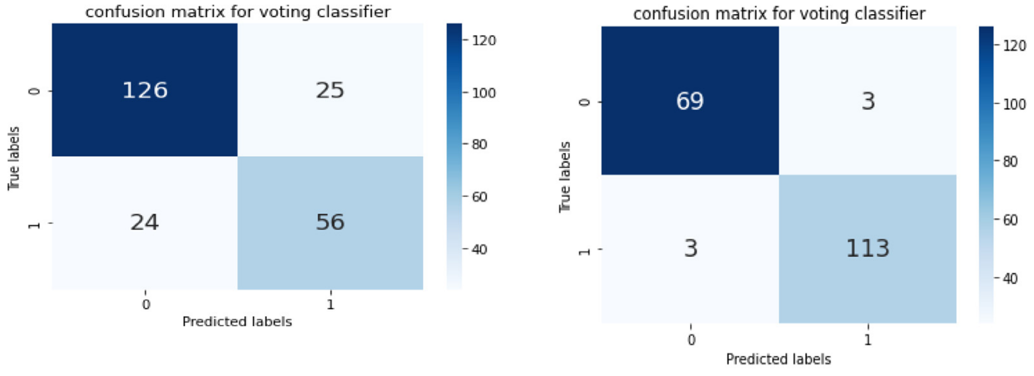
(b) Breast Cancer dataset

Fig. 3. (a-b): Comparative graph of Accuracy, Precision, Recall and F1_score.

machine learning algorithms. It can be analysed from the table that the SVM does not performed good on the given dataset with an accuracy of 60.10%. If we see the individual performance of each Logistic regression, random forest and Naïve Bayes algorithm, all these algorithms give an accuracy of 97.54%, 95.21%, 92.02% respectively. The detailed analysis of all the algorithms has been depicted in Table 2:

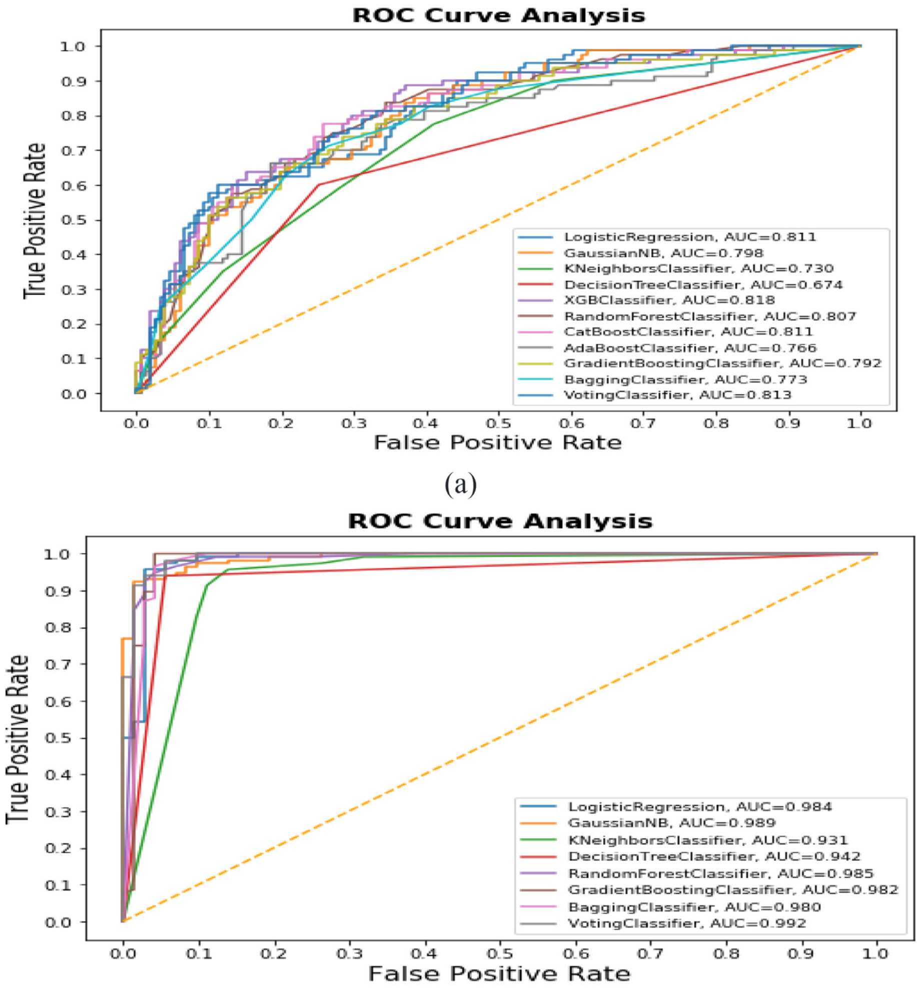
The comparative analysis graph of the various ML algorithm on the PIMA datasets and breast cancer dataset has been shown in Fig. 3(a-b).

Fig. 4(a-b) depicts the confusion matrix for both diabetes and breast cancer patients that are correctly or incorrectly predicted by the proposed ensemble soft voting classifier. In the diabetes dataset,126 were true positive,56 false-positive, and a total of 49 was false positive and



(a) (b)

Fig. 4. Confusion matrix. (a)Pima diabetes dataset (b) Breast cancer dataset.



(a)

(b)

Fig. 5. ROC Curve comparison of (a) Pima Diabetes dataset and (b) Breast Cancer dataset.

false negative. In breast cancer, dataset 69 were true positive,113 true negatives. Hence confusion matrix has given us a better picture of our predictions.

Fig. 5 gives the comparative ROC (Receiver Operating Characteristic) curves for machine learning models which are made by arranging (TPR) true positive rate as opposed to (FPR) false positive rate at different thresholds. The values for the covered surface under the *ROC* curves are computed using Tables 1 and 2. By exploring the ROC curve of breast cancer and diabetes datasets, it has been observed that the proposed model has a more covered area percentage of 81.3% (diabetes dataset) and 99.2% (breast cancer dataset). All other base classifiers and ensemble models such as AdaBoost, gradient boosting, random forest have area under percentage lesser than 81.3%, visible in Fig. 7.

The proposed methodology has been compared with the state-of-the-art methods too. It has been illustrated in Table 3. The results show that the proposed approach has achieved better results as compared with the state-of-art methods.

Table 3
Comparative analysis using state-of-art methods.

Source	Accuracy(%)
(Verma & Mishra, 2017)	76.8
(Bhargava, Sharma, Purohit & Rathore, 2017)	75.5
(Mirshahvalad & Zanjani, 2017)	74
Ensemble Soft voting classifier	**79.08**

5. Conclusion

Diabetes mellitus is an illness that is commonly found in adults nowadays. Hence the early recognition of this disease is the need of an hour. The main objective of this research work has to get the best accuracy and algorithm for predicting diabetes patients. Machine learning algorithms that have been applied in the previous five years were examined regarding their accuracy. Therefore, the authors have proposed a soft voting classifier model by ensemble of three machine learning algorithms such as random forest, logistic regression, Naive Bayes. The Pima Indians diabetes dataset has been taken for experimentation, after that the proposed model has been applied on the breast cancer dataset. The ensemble soft voting classifier has given 79.08% accurate results on the Pima Indians diabetes dataset and 97.02% correct results on the breast cancer dataset. In the future, this accuracy may be enhanced by using different deep learning models.

Declaration of Competing Interest

The authors do not have any conflict of interest with other entities or researchers.

References

Anand, A., & Shakti, D. (2015). Prediction of diabetes based on personal lifestyle indicators. In *Proceedings of the 1st international conference on next generation computing technologies (NGCT)* (pp. 673–676).

Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2017). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In *Proceedings of the 12th international conference on frontiers of information technology* (pp. 226–231).

Bhargava, N., Sharma, S., Purohit, R., & Rathore, P. S. (2017). Prediction of recurrence cancer using J48 algorithm. In *Proceedings of the 2nd international conference on communication and electronics systems (ICCES)* (pp. 386–390).

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Bio Medical Research International*.

Dash, S., Acharya, B.R., .Mittal, M., & Abraham, A. (2020). Deep learning techniques for biomedical and health informatics. A. Kelemen (Ed.).

Fatima, M., & Pasha, M. (2017a). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications, 9*(01), 1.

Fatima, M., & Pasha, M. (2017b). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications, 9*(01), 1.

Goyal, M., Malik, R., Kumar, D., Rathore, S., & Arora, R. (2020). Musculoskeletal abnormality detection in medical imaging using GnCNNr (Group Normalized Convolutional Neural Networks with Regularization). *SN Computer Science, 1*(6), 1–12.

Grudzinski, Karol (2008). Towards heterogeneous similarity function learning for the k-nearest neighbors' classification. *Artificial Intelligence and Soft Computing, 5097*, 578–587.

Harris, J. R., Lippman, M. E., Veronesi, U., & Willett, W. (1992). Breast cancer. *New England Journal of Medicine, 327*(5), 319–328.

Himsworth, H. P., & Kerr, R. B. (1939). Insulin-sensitive and insulin-insensitive types of diabetes mellitus. *Clinical Science, 4*, 119–152.

Husain, A., & Khan, M. H. (2018). Early diabetes prediction using voting based ensemble learning. In *Proceedings of the* international conference on advances in computing and data sciences (pp. 95–103). Springer.

Jiang, Liangxiao, Zhang, H., & Cai, Zhihua (2009). A novel Bayes model: hidden Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering, 21*(10), 1361–1371.

Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology, 4*(10), 426–435.

Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science, 47*, 45–51.

Kaur, B., Sharma, M., Mittal, M., Verma, A., Goyal, L. M., & Hemanth, D. J. (2018). An improved salient object detection algorithm combining background and foreground connectivity for brain image analysis. *Computers and Electrical Engineering, 71*, 692–703.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. New York: Springer-Verlag.

Kumar, D., & Batra, U. Classification of invasive ductal carcinoma from histopathology breast cancer images using stacked generalized ensemble. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1–16.

Kumar, D., & Batra, U. (2018). Epidemiology of breast cancer in Indian women: Population and hospital based study. *EAI Endorsed Transactions on Pervasive Health and Technology, 4*(16).

Kumar, D., & Batra, U. (2020). An ensemble algorithm for breast cancer histopathology image classification. *Journal of Statistics and Management Systems*, 1–12.

Kumar, D., Jain, N., Khurana, A., Mittal, S., Satapathy, S. C., Senkerik, R., et al. (2020). Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks. *IEEE Access : Practical Innovations, Open Solutions, 8*, 142521–142531.

Li, L. (2014). Diagnosis of diabetes using a weight-adjusted voting approach. In *Proceedings of the IEEE international conference on bioinformatics and bioengineering (BIBE)* (pp. 320–324).

Mirshahvalad, R., & Zanjani, N. A. (2017). Diabetes prediction using ensemble perceptron algorithm. In *Proceedings of the 9th international conference on computational intelligence and communication networks (CICN)* (pp. 190–194).

Mittal, M., Arora, M., Pandey, T., & Goyal, L. M. (2020). Image segmentation using deep learning techniques in medical images. *Advancement of Machine Intelligence in Interactive Medical Image Analysis*, 41–63.

Mittal, M., Goyal, L. M., Kaur, S., Kaur, I., Verma, A., & Hemanth, D. J. (2019a). Deep learning based enhanced tumor segmentation approach for MR brain images. *Applied Soft Computing, 78*, 346–354.

Mittal, M., Kaur, I., Pandey, S. C., Verma, A., & Goyal, L. M. (2019b). Opinion mining for the tweets in healthcare sector using fuzzy association rule. *EAI Endorsed Transactions on Pervasive Health and Technology, 4*(16).

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care, 24*(1), 31–42.

Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science, 69*, 132–142.

Ndisang, J. F., Vannacci, A., & Rastogi, S. (2017). *Insulin resistance, type 1 and type 2 diabetes, and related complications*.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes data. Accessed: 1st of (2008, May).

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9*(4), 705–724.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, Calif: Morgan Kaufmann Publishers.

Rani, A. S., & Jyothi, S. (2016). Performance analysis of classification algorithms under different datasets. In *Proceedings of the 3rd international conference on computing for sustainable global development (INDIACom)* (pp. 1584–1589).

S. Sahan, K. Polat, Kodaz, H., & Gunes, S. (2005). The medical applications of attribute weighted artificial immune system (awais):DIagnosis of heart and diabetes diseases. In *Proceedings of the ICARUS* (pp. 456–468).

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261).

Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications, 7*, 89–94.

Standl, E., Khunti, K., Hansen, T. B., & Schnell, O. (2019). The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European Journal of Preventive Cardiology, 26*(2_suppl), 7–14.

Sudharsan, B., Peeples, M., & Shomali, M. (2014). Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *Journal of diabetes science and technology, 9*(1), 86–90.

Verma, D., & Mishra, N. (2017). Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In *Proceedings of the international conference on intelligent sustainable systems (ICISS)* (pp. 533–538).

Vijayan, V. V., & Anjali, C. (2015). Prediction and diagnosis of diabetes mellitus—A machine learning approach. *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 122–127.

Vijayan, V., & Ravikumar, A. (2014). Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. *International Journal of Computer Applications*, (17), 95.

World Health Organization. (2013). Diagnostic criteria and classification of hyper glycaemia first detected in pregnancy No. WHO/NMH/MND/13.2.