A comparative study of different machine learning tools in detecting diabetes

Ghosh, Pronab; Azam, Sami; Karim, Asif; Hassan, Mehedi; Roy, Kuber; Jonkman, Mirjam

Published in:
Procedia Computer Science

DOI:
10.1016/j.procs.2021.08.048

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 467–477

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

A Comparative Study of Different Machine Learning Tools in Detecting Diabetes

Pronab Ghosh [a], *[1]Sami Azam[b], Asif Karim[c], Mehedi Hassan[d], Kuber Roy[e], Mirjam Jonkman[f]

[a,d,e] *Daffodil International University, 102/1, Sukrabad, Dhaka 1207, Bangladesh*
[b,c,f] *College of Engineering, IT and Environment, Charles Darwin University, Casuarina 0810, NT, Australia*

Abstract

A significant proportion of people around the world are currently suffering from the harmful effects of diabetes and a considerable number of them not being identified at an early stage. Over time this may result in serious health problem such as blindness and kidney failure. To accurately classify the disease, different machine learning (ML) approaches can be utilized. In this context, four separate ML algorithms, namely Gradient Boosting (GB), Support Vector Machine (SVM) AdaBoost (AB), and Random Forest (RF) are evaluated using the Pima Indians diabetes dataset, first with based on all features, then to the features selected with the Minimal Redundancy Maximal Relevance (MRMR) Feature Selection (FS) approach. Seven different types of performance evaluation metrics were computed with a 10-fold cross-validation (CV) approach. Computational complexity is also evaluated. The best results were obtained with the Random Forest approach, achieving an accuracy of 99.35%.

Keywords: MRMR, Gradient Boosting, Support Vector Machine *(*RBF kernel*)*, AdaBoost, and Random Forest.

1. Introduction

Diabetes is one of the most common disease leading to disability and death worldwide and its incidence is increasing, particularly in developing countries. For example, a high number of people who live in Bangladesh experience the detrimental effects of diabetes [1] and the number of cases is expected to double by 2025 [2]. An additional problem is that cases are not always diagnosed in a timely manner. A recent study by the International Diabetes Federation estimated that the incidences of undiagnosed and diagnosed cases are among Bangladeshi citizens [2] are about equal. The increasing prevalence of diabetes is anticipated to have a major social and economic impact on a many countries. [3] Diabetes for maintaining proper physical and mental stability, we need to consume food having carbohydrate. Diabetics have a high level of glucose in the blood [4] and urine which can over time lead to major additional health problems, such as blindness and kidney failure. Timely and accurate detection of diabetes is crucial in managing this disease. Machine learning techniques are a promising approach to disease detection. The ultimate goal of the proposed study is to improve early diagnosis with the help of various machine learning algorithms. Performance metrics such

* *Corresponding author email address: asif.karim@cdu.edu.au*

10.1016/j.procs.2021.08.048

as Sensitivity (SEN), Specificity (SPE), Accuracy (ACC), False Negative Rate (FNR), False Discovery Rate (FDR), False Positive Rate (FPR) and Negative Predictive Value (NPV) are used to evaluate these algorithms.

After studying previous work, we propose to use multiple classification algorithms to develop a diabetes prediction model based on patterns generated from clinical datasets of different sizes. Our work is different from previous research in two ways: First, seven different machine learning techniques have been employed. Secondly, medical datasets of different sizes have been analysed using Python programming language. The fundamental characteristics of our proposed prediction system are as follows:

- Missing values are dealt with by the K-Nearest Neighbour's technique.
- A standardised scalar approach is used to obtain values in the range of [0, 1].
- Features are selected with the Minimal-redundancy-maximal-relevance feature selection algorithm (MRMR)
- The performances of different classifiers based on the MRMR selected features is investigated with the help of a 10-fold cross-validation approach.
- Various performance evaluation metrics are used to demonstrate the efficiency of different classifiers.

* Corresponding author – Sami Azam, Tel.: (08) 8946 6666.
E-mail address: sami.azam@cdu.edu.au

- Absolute Error (AE), Root Mean squared Error (RMSE) and Log Loss (LL) are evaluated to investigate the performance of the model.

2. Literature review

Researchers have proposed various Machine Learning (ML) based techniques to construct a diabetes prediction model.

Shafiqul et al. [5] utilized the information from a longitudinal clinical study, the San Antonio Heart Study. Eight ML strategies were utilized to predict whether an individual would develop type 2 diabetes within the next 7-8 years. An accuracy of 95.94% was accomplished by the ensemble Naïve Bayes (NB). Random Forest (RF) and Support Vector Machine (SVM) models were evaluated as well for this dataset, resulting in an acceptable accuracy but a low sensitivity (58.5% for RF and 51.9% for SVM). All results used to predict the rate of diabetes were computed based on a 10- fold Cross-Validation. Wang et al. [6] proposed a prediction algorithm for diabetes to deal with missing values. They applied this to the Pima Indians diabetes dataset from the UCI machine Learning Repository. Their motive was to overcome the limitations of this dataset due to missing values so that diverse ML methods could be used. They also use RF and confirmed the validity of this classifier utilizing the k-fold cross validation method. The classifier resulted in a precision of 87.10% and Area Under Curve (AUC) score of 0.928 when k = 5. SVM, however, only provided an accuracy of 65.1%. Chen et al. [7] investigated a diabetes classification model dependent on a boosting algorithm. The purpose of their work was to show how boosting algorithms performed for diabetes classification models based on a clinical dataset. They utilized two basic boosting algorithms named AdaBoost and LogitBoost for their machine learning model for diabetes. RF was used to check the impact of the boosting algorithms on classification. LogitBoost performed somewhat better than AdaBoost and achieved an accuracy of 95.30%, utilizing 10-fold cross validation with a 0.979 ROC Area score. By utilizing three distinctive ML procedures, Roshan and Ashish et al. in [8] examined whether Logistic Regression (LR), Gradient Boosting (GB), and Naive Bayes (NB) can be utilized for the analysis of diabetes disease. They also used the Pima Indians diabetes dataset to forecast the diagnosis of diabetes. Gradient Boosting achieved an accuracy of 86% vs 79% for LR and 77% for NB They utilized the Boruta algorithm to handle fundamentally co-related attributes like BMI and Plasma glucose. Yukai Li et al. in [9] used different machine learning techniques to improve feature selection and predict and classify diabetes. They worked with data from the Urumqi people in Xinjiang used these as input for SVM and AdaBoost. This dataset contained fasting blood glucose, age, and BMI. They found that the g-mean for AdaBoost was 94.65% and under the ROC curve was 0.9817, which was a more satisfactory results than on the result for SVM. Deepti et al. [10] proposed a model to detect diabetes at an early stage with the help of Support Vector Machine, Decision Tree and Naïve Bayes algorithms. They applied the Pima Indians Diabetes Database (PIDD), sourced from UCI machine learning repository.

Their result was that Naïve Bayes outperformed the other algorithms with the highest accuracy of 76.30%. Sujit et al. in [11] examined automatic diabetes prediction based on Random Forest and Gradient Boosting classifiers. They also used the Pima Indian Diabetes Dataset (PIDD) and found that tree-based ensemble learners could achieve an accuracy above 90%, with high sensitivity, specificity as well and AUC values as high as 94% with proper data pre-processing, hyper-parameter tuning and oversampling of the dataset. Wei et al. [12] explored a similar techniques with classifiers like deep neural network (DNN) and support vector machine (SVM). They too used the Pima Indian dataset to apply these techniques. The accuracies for each classifier were compared with several ways of data pre-processing. They improved the accuracy with parameters modification and obtained an accuracy of 77.86% with 10- fold cross-validation. T.P. Latchoumi et al. in [29] proposed a next-generation diabetes solution called the Smart Diabetes Diagnosis System with the help of various ML classifiers such as SVM, RF, Logistic Regression (LG) and so on. After applying all of the mentioned algorithms to Pima Indians Diabetes dataset (PIDD), LR is the first to be observed with 77.6% accuracy, RF generated a noteworthy accuracy rate of 74.09%, and SVM provided the lowest accuracy of 65.62%. W. Gou et al. in [30] examined different machine learning frameworks to identify the type 2 diabetes and their outcomes uncover a centre arrangement of gut microbiome highlights related with type 2 diabetes risk and future glucose increase. Around 249 cases without type 2 diabetes were examined through Microbiome Risk Score (MRS) with identified features. From their research they identified a robust combination of gut microbes associated with type 2 diabetes by integrating ML techniques. Novel Risk score successfully reproduce MRS type 2 diabetes association in another two independent cohorts. The MRS (counting 14 microbial features) reliably connected with type 2 diabetes, with a risk ratio for each 1-unit change in MRS 1.28 (95% CI 1.23–1.33), 1.23 (1.13–1.34), and 1.12 (1.06–1.18) across three cohorts. The MRS was emphatically related with future glucose increase ($P < 0.05$) and was corresponded with a variety of gut microbiota–derived blood metabolites.

3. Machine learning techniques

This section briefly discusses the ML algorithms that have been explored in this study.

3.1. Random forest (RF)

Random Forest (RF) is one of the most commonly used supervised machine learning techniques for classification and regression [13]. It works by creating a forest from a lot of random and unconnected Decision Trees (DTs) in the training process. Here the number of used n_estimators was 10. For a powerful and accurate analysis, ensemble methods may perform better than other models. Since the RF approach uses more features than the independent DTs in calculations, it can obtain a greater complexity, usually achieves greater accuracy when dealing with unknown datasets. Random forest can be defined as $\{h(x, \theta_k), k = 1, \ldots\ldots\ldots, n\}$, where x = the input data and θ_k = a is a mutually independent random vector parameter and [14]. Each DT, using a random vector as parameter, randomly selects the sample features and the subset of the sample data set to be used as training set.

3.2. Support vector machines (SVM)

SVM is used as a tool to find classification and regression rules, usually from a large amount of data. This technique adopts sigmoid as a kernel function. In order to classify test- data by SVM, an optimal hyperplane [15] is created before a collection of labelled training data is processed. The purpose of the SVM model is to find the space in the data matrix where different data groups can be drawn on the hyperplane with a wide separation. SVM is has been used in histopathological images to classifying the stage of cancer [16]. An n-levels hyperplane is defined by equation (1) as:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n = 0 \quad (1)$$

Where β_0, β_1, β_2... β_n depict hypothetical values and X_n displays data points in the n-dimension sample space. The initial aim of applying SVM was to solve the problem of 2-class classification but it was later expanded for multi-

class issues. The algorithm applies a 1-vs-rest method, attempting to isolate a single class is attempted from all other classes [17] as explained in Equation (2).

$$Z = \begin{cases} n, & \text{if } d_n(y) + t_l > 0 \\ 0, & \text{if } d_n(y) + t_l < 0 \end{cases} \quad (2)$$

Where $d_n(y) = max\{d_n\}_{i=1}^{Nl}$, $d_i(y)$ is the distance from y to the SVM hyperplane corresponding to class i, and t_l is the classification threshold.

3.3. AdaBoost

Adaptive Boosting (henceforth AdaBoost) is a boosting algorithm used for binary classification. It combines several weak classifiers to create a more robust collection [18]. This ML algorithm generates the outputs based on the 100 estimators. The training dataset instances are weighted with a starting weight [19] as shown in equation (3).

$$\text{Weight}(xi) = 1/N \quad (3)$$

Where ***N*** is the frequency of training instances, and x_i is i^{th} training instance. The decision stump gives an output for each input variable. The misclassification rate is then calculated using equation (4):

$$\text{Error} = (\text{correct} - N) / N \quad (4)$$

Where ***N*** is the frequency of training instances. Boosting simply means combining several simple trainers to achieve a more accurate prediction. AdaBoost fixes the weights which vary for both samples and classifiers. This enables the classifiers to focus on result that are relatively difficult to identify accurately. The classification e is calculated with equation (5).

$$H_k(p) = +/-(\sum_{k=1}^{k} a_k h_k(p)) \quad (5)$$

Equation (5) is a linear combination of all the weak classifiers (simple learners), where ***K*** is the total number of weak classifiers $h_k(p)$ is the output of weak classifier ***t*** (this can be either -1 or 1). a_k is the weight of classifier ***k***.

3.4. Gradient boosting

Gradient Boosting is often used for classification and regression problems [20]. Three factors are essential for this technique: An enhanced loss function, a weak learner to make predictions and an additive model to combine weak learners to minimize the loss function [21]. Gradient Boosting can enhance the efficiency of the algorithm by eliminating any occurrence of overfitting. If there is an imbalance between the numbers in each class, the model helps to improve the accuracy. Boosting methods such as regression tree learners, helps to get a higher predictive accuracy for a large variety of datasets, e.g. [22]. It makes use however of familiarity in a specific area. The Boosting process makes a difference to traditional machine learning by eliminating function space optimization. The optimal function ***F(X)*** is obtained after ***m*** iterations [23] as described by equation (6):

$$F(X) = \sum_{i=0}^{m} f_i(x) \quad (6)$$

where $f_i(x) = -\rho_i x\, gm(X)\ (i = 1, 2\ldots, M)$ indicates feature increments$f_i(x)$. The latest base-learner is the largest loss function correlated with negative gradients. The negative gradient for the m^{th} iteration is (7).

$$\mathrm{gm} = -\left[\frac{\partial \mathrm{L}(y,\mathrm{F}(\mathrm{X}))}{\partial \mathrm{F}(\mathrm{X})}\right]_{\mathrm{F}(\mathrm{X})=\mathrm{Fm}-1(\mathrm{X})} \quad (7)$$

where ***gm*** is the path where the loss function decreases most rapidly when $\boldsymbol{F(X) = F_m - 1(X)}$ [23]. A new decision tree aims to correct the error made by the previous base learner. The model is then modified to (8).

$$F_m(\mathrm{X}) = F_m - 1(\mathrm{X}) + \rho_m \, \mathrm{x} \, h_m(\mathrm{X}, \alpha_m) \quad (8)$$

4. Research methodology

This section provides a detailed explanation of the methodologies used in this study.

4 .1. Dataset

The Pima Indians Diabetes dataset (PIDD) [24], taken from the UCI machine learning repository, is used for this research. A total of 768 instances with 8 input attributes are used: glucose level, blood pressure, skin thickness, insulin, BMI, number of times pregnant, diabetes pedigree function, and age. There are some missing values for glucose level, blood pressure, skin thickness, insulin and BMI in the dataset, which have been filled by the KNN imputation technique.

4.2. Data pre-processing

For the successful use of the classifiers, pre-processing techniques such as eliminating missing values and application of standard scalars are applied to the dataset. Two common techniques for managing the dataset can be used: deletion and imputation. Deletion is often considered inefficient especially when it comes to medical datasets. As a consequence, the imputation method is used as an alternative method, where missing values are replaced by estimated values. In this, K- Nearest Neighbor (KNN) has been used where cases are imputed using values similar to K. Furthermore, the KNN classifier calculates the distance from point to point in the training data set using the Euclidean distance function [25] to classify an unknown instance and then replaces the missing values.

4.3. Feature selection

The machine learning process requires the selection of features as irrelevant features may affect the performance of the classifiers. Proper selection of features enhances the accuracy of classification and reduces the execution time for the model [26]. In this research, we have used the Minimal-redundancy-maximal-relevance feature selection algorithm feature selection (FS) algorithm.

The MRMR starts withcharacteristics that correspond to the target labels but some of these features are redundant. The MRMR algorithm uses an Heuristic search method to selects the features with maximal relevance and minimal redundancy. It checks one feature per cycle and then calculates redundancy in pairs [27]. The mutual knowledge (MK) between two functions is calculated with this algorithm. This method is determined for each pair of features. MRMR is not appropriate for selection problems with a very wide range of features. After applying the MRMR algorithm on the given dataset, the following outcomes have been generated and five features have been selected to create an the machine learning model. Pregnancies (PREG) and BMI (BMI), appear to be the most important role (see Table 1). However, the lowest value was obtained for the Diabetes Pedigree Function (DPF) feature (around 0.116%). A rank score of 0.513 was obtained for Glucose. The age feature also has a relatively high rank score (about 0.60).

Table 1. The tabular format of significant features with its rank by MRMR approach

Features Name	Features Code	Rank

Pregnancies	PREG	0.8453064
Glucose	GLU	0.5134933
BMI	BMI	0.8711340
DiabetesPedigreeFunction	DPF	0.1162514
Age	AGE	0.6040413

4.4. Stratified k-fold cross validation

The dataset set is separated into equal parts for the K- fold cross-validation [17]. K-1 groups are used to train the classifiers, and the rest to test the performance at each step. The validation process is repeated several times and the classifier performance is then calculated based on the results. Different values of K are selected for the Cross Validation (CV) process. We have used K = 10 in our study, where 80% of the data are utilized for training purposes and 20% of the data are used for testing . This process is repeated for 10 times. Before selecting and testing new sets for the new cycle, all instances in the training and test group are randomised over the entire dataset. At the end of the 10-fold process, the average of all performance metrics is calculated.

4.5. Statistical analysis of the proposed model to diagnose diabetes

First, we need to check missing value. If there is no missing value, the standard scaler can directly be applied [28]; otherwise, the KNN method should be used on the dataset. After that, a standard scaler is used bring all data in the range of [0 to 1]. Subsequently, MRMR technique is used to select the best features. A 10 -fold cross validation technique is then used on the selected features. Next, after separating the dataset into two different parts, 80% of dataset is used for training and the remainder for the testing process. Four different machines learning algorithms are applied: RF, SVM, AB and GB. In order to test the efficiency of different classifiers, different performance assessment metrics are utilized in this study. A diagram of the complete model is shown in Fig. 1.

We used the confusion matrix to show the predicted outcomes. The performance measurement indicators can be calculated based on the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rates [25].

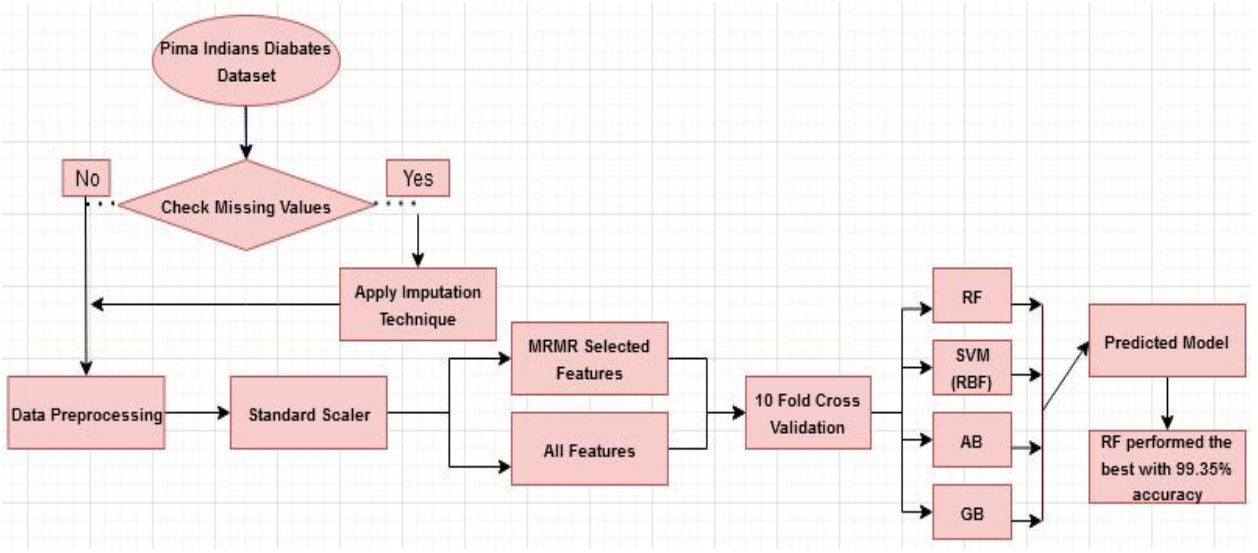

Fig. 1. The proposed architecture of diabetes disease to successfully diagnose the affected patients.

5. Experimental results and discussion

5.1 Correlation of Features

Both Age and Glucose are strongly correlated with Pregnancies where the value is close to 0.3. Likewise, the selected characteristics such as Age and BMI have a strong correlation with Glucose and their correlated value is almost 0.3. As can be expected there is no strong correlation between Diabetes Pedigree Function and Pregnancies or Age. There is some correlation between Diabetes Pedigree Function and BMI and Glucose though. In the following phase, with the help of selective algorithms, the optimum outcome has been achieved through the RF approach, an accuracy of 99.35%. Correlations between the different features are shown in Fig. 2.

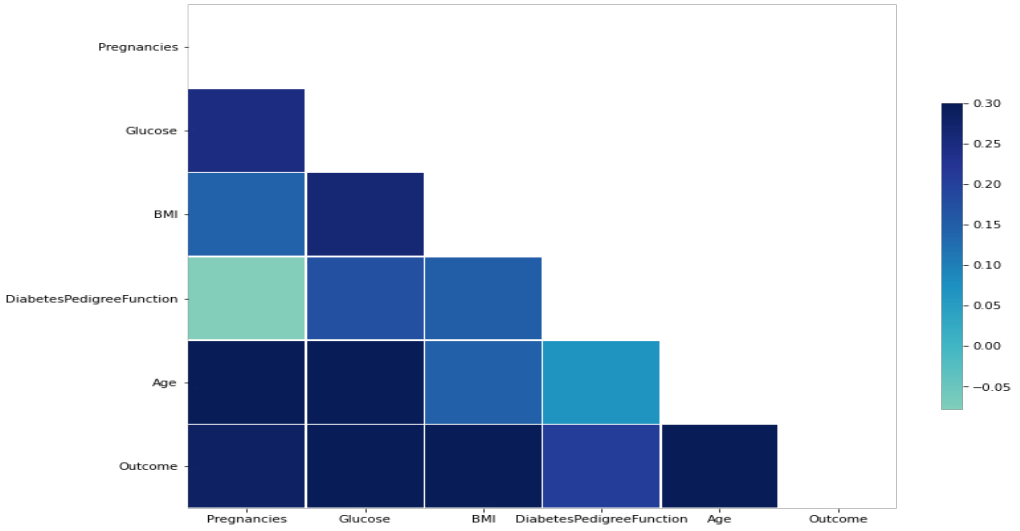

Fig. 2. The proposed architecture of MRMR approach with highly correlated features.

5.2. Outcomes for all features

The overall results that were computed based on the discussed algorithms such as RF, SVM, AB and GB have been shown in Table 2. The most desired outcomes of Accuracy, SEN, SPE and NPV have come from the RF algorithm and the scores of these dimensions on full features were about 99.35%, 99.01%, 100% and 98.15%. Additionally, the lowest results were observed considering these four dimensions by the SVM technique (85.06%, 85.98%, 82.14% and 72.22%). Both AB and GB have also shown considerably better results compared to the SVM model. The calculation results of SVM approach were significantly higher for FPR (17.86%) and FNR (14.02%), whereas the least scores were witnessed on the FPR and FNR values for the RF algorithm (approximately 0% and 0.99%). Neither of AB and GB was performed as high as SVM on the following dimensions (FNR, FPR scores), however the highest FDR value was produced by AB.

Table 2. Results of the various algorithms for all features

Dimensions	RF	SVM	AdaBoost	Gradient Boosting
Accuracy	99.35%	85.06%	88.31%	93.51%
SEN	99.01 %	85.98%	91.84%	96.88%
SPE	100%	82.14 %	82. 98%	87.93%
FPR	0%	17.86%	17.02%	12.07%
FNR	0.99%	14.02%	8.16%	3.125%
NPV	98.15%	72.22%	85.19%	94.44%
FDR	0%	8%	10%	7%

5.3. Analysis of full features outcomes

Three different types of tests have been done to further evaluate the effectiveness of our model: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Log Loss (LL), see Table 3. For all features, the RF algorithm on all features had the lowest error scores, with an MAE of 0.649% and an RMSE of 8.06%. The MAE and RMSE for the SVM model were the highest: 14.93% and 38.65%, respectively. There error rates for the AB (MAE 11.69%, RMSE 34.19%) and GB (MAE 6.49%, RMSE 25.48%) techniques fell between those of RF and SVM. The LL scores of SVM and AB (515.84% and 403.71%) were higher than those of GB (224.28%) and RF algorithm (just under 22.43%).

Table 3. The outcomes for all features

Dimensions	RF	SVM	AdaBoost	Gradient Boosting
MAE	0.649%	14.94%	11.69%	6.49%
RMSE	8.06%	38.65%	34.19%	25.48%
LL	22.43%	515.84%	403.71%	224.28%

5.4. Outcomes for the MRMR selected features

Based on the used algorithms such as RF, SVM, AB and GB, the overall results have been illustrated in Table 4. GB and SVM have also performed well enough, however, AB did not provide the best outcome for it, where RF touches the peak of the accuracy of 99.35%. It is almost 2% higher than its nearest rival GB. Third best is SVM among those 4 ML techniques with the score of 90.26%. The most desired results for SEN, SPE and NPV have come from the RF algorithm and the scores of these dimensions on the selected features of MRMR technique are about 99.01%, 100% and 98.15%. Additionally, the lowest results were observed by the SVM technique considering SEN and NPV dimensions (89.72% and 79.63%). For FPR (12.22%) and FDR (14%) scores, the AB method calculation results are

significantly higher, while the lowest scores are seen on the FPR and FDR values for the RF algorithm (approximately 0% and 0%). The result has been satisfactory to us, however, if we try to increase the accuracy level for the other algorithms, we must prepare our dataset properly. All of the categorical symptoms should be numbered equally. Furthermore, for increasing accuracy level there is no better option than data cleaning. The fact is that the more data are pre-processed, the more accurate prediction will be shown by this classifier.

Table 4. The outcomes for the MRMR features

Dimensions	RF	SVM	AdaBoost	Gradient Boosting
Accuracy	99.35%	90.26%	87.66%	97.40%
Sensitivity	99.01 %	89.72%	94.51%	97.06%
Specificity	100%	91.49%	87.77%	98.08%
false positive rate	0%	8.51%	12.22%	1.92%
false negative rate	0.99%	10.28%	5.49%	2.94%
Negative Predictive Value	98.15%	79.63%	90.74%	94.44%
False Discovery rate	0%	4%	14%	1%

5.5. Analysis of MRMR features outcomes

To further evaluate the different algorithms, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Log Loss (LL) were calculated for the MRMR features, see Table 5. For all features, the RF algorithm had the lowest MSE score, 0.649%. Scores for SVM (9.74%), AB (12.34%) and GB (2.59%) were all considerably higher. RMSE for AB and SVM were both quite high (35.13% and 31.21%). The Log Loss score of RF was very low (22.45%) with LL scores for SVM, AB and GB of 336.42%, 426.14% and 89.71% respectively.

Table 5. The outcomes for the MRMR features

Dimensions	RF	SVM	AdaBoost	Gradient Boosting
Mean Absolute Error	0.649%	9.74%	12.34%	2.59%
Root Mean Squared Error	8.06%	31.21%	35.13%	16.12%
Log Loss	22.43%	336.42%	426.14%	89.71%

5.6. Computational complexity of the introduced algorithms on the 5th chosen features

Table 6 illustrates the computational complexity of the four methods. We have computed both the Training complexity and the Prediction complexity. Additionally, the different types of parameters including n, the number of training samples, p, the number of attributes, n_{trees}, the number of trees and n_{sv}, the number of support vectors were found.

Table 6: Displayed outcomes of computational complexity

Models	Calculation Method - Training	Obtained Complexity - Training	Models	Calculation Method - Prediction	Obtained Complexity – Prediction
AB	$O(npn_{trees})$	O(384000)	RF	$O(pn_{trees})$	O(50)
GB	$O(npn_{trees})$	O(384000)	SVM (RBF Kernel)	$O(n_{sv}p)$	O(50)
RF	$O(n^2pn_{trees})$	O(29491200)	GB	$O(pn_{trees})$	O(100)
SVM (RBF Kernel)	$O(n^2p+n^3)$	O(455933952)	AB	$O(pn_{trees})$	O(500)

6. Conclusions and Future Work

After evaluating all performance metrics, it can be concluded that the Random Forest approach is the most suitable method, with an accuracy of 99.35%. SVM and AB both had the lowest performance. Therefore, a prediction system for early diagnosis of diabetes based on the Random Forest model holds significant promise. To further improve these results, careful data should be incorporated as a key aspect of the. Additionally, to improve the prediction accuracy, the use of pipeline structures for data pre-processing could be explored in the future. Development of an Android application is also a possibility as this may help practitioners to receive accurate diagnostic results instantly.

References

[1] S. Islam, A Lechner et al. “Healthcare use and expenditure for diabetes in Bangladesh,” BMJ Global Health, Vol. 2, Issue. 1, 2017.

[2] P. M. S. Sai, G. Anuradha, P. kumar, “Survey on Type 2 Diabetes Prediction Using Machine Learning,” Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020.

[3] P. K. Dhillon, P. Jeemon, N. K. Arora, et al., “Status of epidemiology in the WHO South-East Asia region: burden of disease determinants of health and epidemiological research, workforce and training capacity,” Int Journal of Epidemiology, Vol. 41, pp. 847–860, 2012.

[4] A. Anand, D. Shakti, “Prediction of Diabetes Based on Personal Lifestyle Indicators,” 1st International Conference on Next Generation Computing Technologies (NGCT-2015), Dehradun, India, 4-5 September, 2015

[5] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," in IEEE Access, vol. 8, pp. 120537-120547, 2020, doi: 10.1109/ACCESS.2020.3005540.

[6] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[7] Chen, P., Pan, C. Diabetes classification model based on boosting algorithms. BMC Bioinformatics, Vol. 19, March 2018.

[8] Birjais, R., Mourya, A.K., Chauhan, R. et al. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences. Vol. 1, 1112 September 2019, https://doi.org/10.1007/s42452-019-1117-9

[9] Yukai Li, Huling Li, Hua Yao, "Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017", Computational and Mathematical Methods in Medicine, vol. 2018, Article ID 7207151, 8 pages, 2018. https://doi.org/10.1155/2018/7207151

[10] D. Sisodia, D.S. Sisodia, “Prediction of Diabetes using Classification Algorithms”, International Conference on Computational Intelligence and Data Science (ICCIDS), Procedia Computer Science, Vol. 132, pp. 1578–1585, 2018.

[11] S Das, A Mishra, P Roy – 2019, “Automatic Diabetes Prediction Using Tree Based Ensemble Learners”, International Conference on Computational Intelligence & IoT(ICCIIoT), 2018.

[12] Wei S, Zhao X, Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. In Internet of Things (WF-IoT), 2018 IEEE 4th World Forum, pp. 291-295, 5 Feb, 2018.

[13] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 2020, pp. 1-6, doi: 10.1109/iSAI-NLP51646.2020.9376787.

[14] Ren, Q., Cheng, H., Han, H.: Research on machine learning framework based on random forest algorithm, AIP Conference Proceedings, vol. 1820, 2017.

[15] M. Srivenkatesh, “Prediction of Breast Cancer Disease Using Machine Learning Algorithms,” International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9 Issue. 4, February 2020.

[16] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, “Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors,” IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017.

[17] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, M. Saifuzzaman," Expert Cancer Model Using Supervised Algorithms with a LASSO Feature Selection Approach," International Journal of Electrical and Computer Engineering, Vol. 11, No. 03, pp. 2632~2640, 2021, DOI: 10.11591/ijece.v11i3.pp2632-2640.

[18] S. H. Ripon, ‘‘Rule induction and prediction of chronic kidney dis-ease using boosting classifiers, Ant-Miner and J48 Decision Tree,’’ inProc. Int. Conf. Elect., Comput. Commun. Eng. (ECCE), Cox’s Bazar, Bangladesh, 2019, pp. 1–6.

[19] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab,‘‘A comprehensive survey for intelligent spam email detection,’’IEEEAccess, vol. 7, pp. 168261–168295, 2019. DOI: 10.1109/ACCESS.2019.2954791.

[20] “An Overview of Gradient Boosting Algorithm. Accessed,” [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-Boosting-algorithm-machine-learning/, [Accessed: 24-12-2020].

[21] Gradient Boosting Algorithm. Accessed: Jun. 31, 2020. [Online]. Available: https://data-flair.training/blogs/gradient-Boosting-algorithm/

[22] T. Chen and C. Guestrin, ‘‘XGBOOST: A scalable tree boosting system,’’inProc. 22nd ACMSIGKDD Int. Conf. Knowl. Discovery Data Mining,2016, pp. 785–794.

[23] J. Cheng, G. Li, and X. Chen, ‘‘Research on travel time prediction modelof freeway based on gradient boosting decision tree,’’IEEE Access, vol. 7,pp. 7466–7480, 2019, doi: 10.1109/ACCESS.2018.2886549.

[24] “UCI Machine Learning Repository: Pima Indians Diabetes,” [Online]. Available: https://archive.ics.uci.edu/ml/machinelearning-databases/pima-indians-diabetes/. [Accessed: 28-12-2020].

[25] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework," in IEEE Access, vol. 8, pp. 154759-154788, 2020, doi: 10.1109/ACCESS.2020.3017082.

[26] P. Ghosh, S. Azam, A. Karim, M. Jonkman, M.Z. Hasan, "Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases," 5th ACM International Conference on Information System and Data Mining (ICISDM2021), 2021.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, maxrelevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[28] "Standard Scaler Technique," [online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, [Accessed: 27-12-2020].

[29] T. P. Latchoumi, J. Dayanika, G. Archana, "A Comparative Study of Machine Learning Algorithms using Quick-Witted Diabetic Prevention," Annals of R.S.C.B., vol. 25, Issue 4, pp. 4249 – 4259, 2021.

[30] W. Gou et al., "Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated with Type 2 Diabetes," Diabetes Care, vol. 44, pp. 358–366, 2021.