Improving of Diabetes Diagnosis using Ensembles and Machine Learning Methods

Razieh Asgarnezhad[1*], Karrar Ali Mohsin Alhameedawi[1, 2]
1- Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran.
Email: razyehan@gmail.com (Corresponding author)
2- Department of Computer Engineering, Al-Rafidain University of Baghdad, Baghdad, Iraq.
Email: karraralimohsin125@gmail.com

Received: October 2021 Revised: November 2021 Accepted: January 2022

ABSTRACT:
Diabetes is one of the most common metabolic diseases, and diagnosis of it is a classification problem. The most challenge is this area is missing value problem. Artificial Intelligence techniques have been successfully implemented over medical disease diagnoses. Classification systems aim clinicians to predict the risk factors that cause diabetes. To address this challenge, we introduce a novel model to investigate the role of pre-processing and data reduction for classification problems in the diagnosis of diabetes. The model has four stages consists of Pre-processing, Feature sub-selection, Classification, and Performance. In the classification technique, ensemble techniques such as bagging, boosting, stacking, and voting were used. We considered both states with/without for pre-processing stage to reveal the high performance of our model. Two experiments were conducted to reveal the performance of the model for the diagnosis of diabetics Mellitus. The results confirmed the superiority of the proposed method over the state-of-the-art systems, and the best accuracy and F1 achieved 97.12% and 97.40%, respectively.

KEYWORDS: Data Mining, Pre-processing, Diabetes Mellitus, Ensembles, Machine Learning

1. INTRODUCTION

Diabetes is one of the most irritating diseases that affect the human body, as this disease affects all age groups, children, adults, and the elderly, but the majority who have a high rate of infection are the elderly, as they are infected with diabetes through shocks and crises. Psychological and others, where there is no treatment for this disease that heals completely, but it is treated without a complete cure only, except in some cases rarely, where diabetes has insulin treatment and other treatment that does not cure it completely. But this treatment is considered as soon as giving an analgesic or reducing high sugar for some. One of the most main problems caused by this disease, as it causes chronic damage, also makes the body less immune, as we see people who suffer from this disease lose their weight. It is caused by the lack of commitment to the prevention of this disease [1]. An unpleasant and chronic disease, and being infected with it leads to human psychological frustration and tension, as most of those who suffer from it suffer from stress.

In 2021, the authors built two separate groups, whereby the cases were classified as non-diabetic patients, and this proves that this technique introduced by the authors will improve the performance of the classification. They reached a recall of 80.86%, the precision of 80.95%, the accuracy of 80.86%, and F1 of 80.83%. These results appeared in proportion to the authors' work, satisfactory results, but our work far exceeded these results, as its accuracy reached 97.12%. Our result will improve the performance of diabetes [2]. In 2017, the authors suggested a study approach to predicting diabetes and improving classification performance. They used external detection as a step for diabetes prediction and pre-processing, where the results showed well. Our work is also superior to this work and has excellent forecasting, improving the performance of the classifier and handling missing values [3].

The current researchers present a strong predictive model that advances classification performance and works to improve its quality. We downloaded data from the UCI website, which is data for people with diabetes that contain symptoms of diabetes and registered cases with several 520 cases. Then, we applied techniques to gain success with these techniques gives distinct values than others. They used the Rapid Miner tool and applied ensemble operations including bagging, boosting, voting, and stacking [4]. We applied these techniques without pre-processing through Decision Tree (DT), Random Forest (RF), and K-nearest neighbor (KNN)

Paper type: Research paper
DOI: https://doi.org/10.52547/mjtd.11.1.33
How to cite this paper: R. Asgarnezhad and K. A. Mohsin Alhameedawi, "**Improving of Diabetes Diagnosis using Ensembles and Machine Learning Methods**", Majlesi Journal of Telecommunication Devices, Vol. 11, No. 1, pp. 33-41, 2022.

algorithms [5]. The accuracies of bagging, boosting, voting, and stacking reached 97.12, 97.12, 97.12, 96.15%, respectively. These values, which we obtained from applying techniques and without pre-processing are good for better prediction and satisfactory results, but it is not enough. Then, the current authors carried out other experiments, which is the application of the ensemble algorithms with pre-processing through KNN, RF, and DT. We obtained impressive results for improving classification performance, where the accuracy of bagging, boosting, voting, and stacking reached high results through which we can improve classification performance and give valuable and high values. The accuracy in four cases reached 97.12.,97.12, 97.12, 97.12%, and these values are considered the highest values in this article. We gave such results to predict well, as these results showed a high success that surpasses the previous works, and this indicates that the research presented in this paper by the current authors is solid. It is good and gives better results and will excellently predict diabetes and improve classification performance.

Our innovation in this paper is to develop a technical model for predicting good results and improving classification performance to predicting chronic diabetes. We have provided a high-accuracy model that improves and processes missing data and advances its performance. Ensemble techniques have been developed including bagging, boosting, voting, and stacking. The obtained results have been reached in some tables have extracted satisfactory and convincing results, and through this innovation, we will improve the performance of classification and predict diabetes significantly.

This paper is organized as follows: The proposed method is presented in next Section and followed by the experiment. Finally, the paper presents a conclusion in last Section.

2. LITERATURE REVIEW

Several works suggested predicting the autism diagnosis in the years between 2015 to 2021. We summarized some of the significant works herein.

Authors in 2015 suggested a method for the diagnosis of diabetes mellitus. Diabetes is one of the most dangerous diseases that cause blindness in the eyes, impaired vision, and disturbance of the eye networks, where this disease is considered more affecting the eye. A model that reduces blindness with the way they presented it can reduce the incidence of blindness significantly as it showed good results with them, but our work outperforms this work because we predicted greatly and its accuracy showed high values to improve the behavior and performance of diabetes, as we were able in our work to obtain accuracy on it at an average rate 97.12%. This result that we obtained confirms that our work outperforms all previous work, as it improves the performance of the classification in a different and significant way [6].

Authors in 2016 applied Data Mining Classification Techniques for the diagnosis of diabetes. Because of the importance of diabetes and its spread in the recent period, this prompted many researchers in the world to search and work to find a distinctive way to enhance the performance of sugar. These authors presented three algorithms to improve diabetes and predict with great suspicion for data mining, where the algorithms presented are the Self-Organizing Map (SOM), C4.5, and RF. These algorithms employed to the national population data gave good results with them where their highest accuracy of recall and precision with RF reached 90%, and these are good values, but our work outperforms this work by a large percentage, as in this paper we showed techniques that predict better than all previous work [7].

The current authors in 2017 suggested a model for pre-processing data that suffer from missing values and has stray values as well. They used the methods of replacing missing values and selecting the attribute where they worked to improve the performance of classification and prediction significantly in the troublesome chronic diabetes, which is one of the most prevalent diseases in the previous and current centuries. They presented techniques to improve From this disease, including Support Vector Machine (SVM), Naïve Bayes (NB), (latent Darrelle Allocation (LDA), and SVM where they reached satisfactory values, where the accuracy in SVM reached 84.35%, and this is the highest value in their submitted research, and this leads to improving the performance of classification, but in our work in this paper, we excel in this work, as our highest accuracy in this article reached 97.12% in their advanced techniques, and this outperforms all previous work, as we made a good prediction in improving sugar performance [8].

Authors in 2018, used classification algorithms to predict the diagnosis of diabetes. Techniques to predict the best values and improve the level of diabetes, where the authors applied classification techniques including DT, NB, and SVM, where the results showed with the highest accuracy in their paper, where the accuracy with the NB technique reached 76.30%, where this value is considered the highest accuracy in their paper which was presented in 2018 to improve diabetes, and this confirms the superiority of our work over the previous works, as the accuracy reached in this article 97.12% and this indicates that our work showed high results in improving classification performance and outperforming all previous work, as through our work we will make an excellent prediction and improve the performance of diabetes [9].

Authors in 2019 suggested a high predictive model

for improving diabetes performance. They introduced a hybrid model using Machine Learning (ML) to detect diabetes with improving the performance of NB, Bayes Net (BN), RF, and KNN. The results in their article showed excellent results, which are close to the results of our work. The recall, precision, accuracy, and F1 were 99.10, 99.10, 99.06, 99.10%, respectively. These values are considered among the best values and results in previous work. They predicted highly, and the authors were able to improve the performance of diabetes and improve its condition. Compared to our work, these results are good and gave excellent results, and this is a good thing for researchers to provide research and search for good ways to improve the performance of diabetes [10].

In 2019, the authors suggested a model investigate performance analysis of ML Techniques for diabetes mellitus. They analyzed early diabetes mellitus to improve the classification performance. They applied algorithms including SVM, NB, and KNN. They got little results compared to our work and the results we got where the results of these authors reached the accuracy of 74%, the precision of 72%, the recall of 74%, and the F1 of 72%. These are considered satisfactory values, but our work has surpassed it [11].

The authors suggested in 2020 a high model of machine learning, where their experiments showed good and satisfactory results with accuracy of 88.1%, recall of 87.8%, precision of 87.9%, and F1 of 87.83%. These values are good and germinate well, but our work exceeded it and showed that our work outperforms all previous work to improve the performance of sugar and better forecasting [12]. Authors in 2020 suggested a hybrid ML model predict diabetes mellitus. They proposed a hybrid model that teaches the prediction model to improve the performance of sugar and predict it and detect this disease, where diabetes is considered a chronic disease and affects the human body and has negative effects if it does not commit to taking medications and other care. They provided the mixers with techniques including boost, RF, BN, and KNN. They applied these techniques to get good results, and they got what they wanted. They got excellent values compared to its value, where the value of recall equal 99.10%, accuracy equal 99.06%, precision equal 99.10%, and F1 equal 99.10%. These values are considered good values for better prediction, and they are similar to the results in this article where our work showed good work and predicts abnormally and excellently [10].

In 2021, the authors proposed a model for improving diabetes. They suggested two separate data sets for diabetes improvement accuracy of 82.10%, recall of 82.10%, precision of 82.10%, and F1 of 82.10%. These results are considered satisfactory, but they are less accurate than our work, so it has been proven that our work is good and predicted greatly to improve the performance of sugar [2]. Authors in 2021 suggested a model using the map-reduce based optimally gradient boosted tree classification algorithm for diabetes mellitus diagnosis system. Development of a new map reducing technology of the Data Mining (DM) diabetes data classifier. They applied important techniques to obtain satisfactory results and obtained excellent results predicting diabetes and improving its performance. Introducing KNN through their presented work, they reached satisfactory results as recall of 97.48%, the precision of 99.23%, the accuracy of 97.79%, and F1 98.34%. These values are good compared to our work. They predict well and improve the performance of the classification, as this confirms that our work is excellent and is superior to some works, as well as we predict it in an excellent manner, which makes us excel over some works and outperform them in terms of accuracy [13].

To sum up, we studied and displayed an excellent model for better prediction, where we used ensemble algorithms that include bagging, boosting, stacking, and voting with/without pre-processing. It reaches 97.12%, and we have proven it through our work. The algorithms used in our paper gave good results and high accuracy, as they predicted well and improved classification performance. Our model is considered a distinguished work that outperforms many previous works and has wonderfully proven its quality and results. Results with some work and this are what makes us a result.

3. THE PROPOSED MODEL

Here, we have done two experiments to investigate the influence of ensemble and ML methods before/after pre-processing stages. For pre-processing, we used a method to address the outlier problem with KNN. We also used the feature to replace missing values with the mean for solving missing values. The employed ensembles and ML methods are bagging, boosting, voting, and stacking [14]. Fig. 1 shows the stages of the proposed model.

A brief description of the applied methods is of concern: The steps of our model are depicted in this section. We processed the data downloaded from the UCI website. The data on cases of diabetes. In this paper, we worked in divided into the participants of the first track, applying the ensemble techniques that include bagging, boosting, voting, and stacking in conjunction with RF, DT, and KNN classifiers without pre-processing with the Rapid Miner tool. Then we used the ensemble techniques with pre-processing which gave the highest values that can be obtained and outperformed the counterparts.

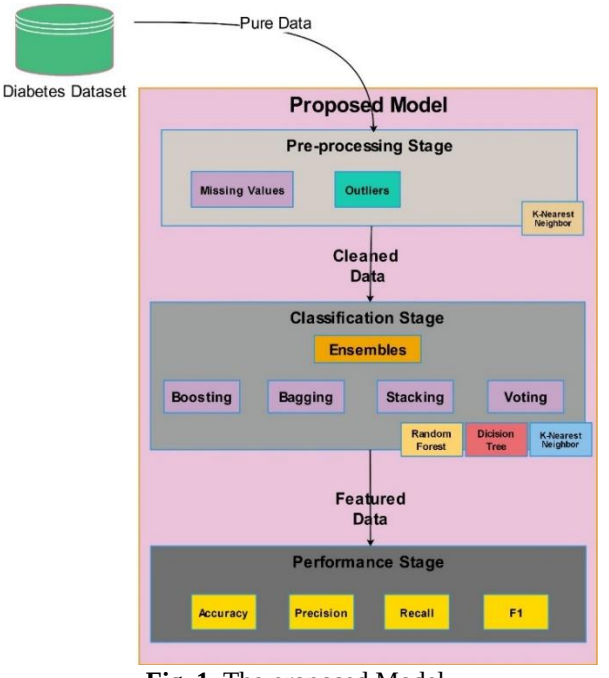


Fig. 1. The proposed Model.

3.1. Data collection

We collected the data for this paper by searching for diabetic records, where we downloaded the data and cases of the 720 diabetic patients and 17 attributes. There are missing values in this data, as diabetes is a highly prevalent and chronic disease, so large data was collected and downloaded from this site UCI to work on it and improve it and work on processing the lost data in it as well. We will apply ensemble algorithms including bagging, boosting, voting, stacking. To address the problem of this dataset pre-processing is used for missing values by replacing the missing values with values by mean. Also, we applied methods to detect outlier where these data containing all suspected cases and some cases and predictions for diabetes will improve and be addressed.

3.2. Pre-processing Stage for Missing Values

In this paper, we have the data of the number of records 720 record and 17 attributes, where this data contains missing values and stray values, and we applied the pre-processing techniques to it to treat this case and predict quickly, comfortably, smoothly, and far from complexity. We applied the process of replacing the missing values with the mean. We applied the detection outlier process. We develop these data to decrease the missing values by replacing them with values, and predicting high levels of diabetes, and improving performance. We work to promote predicting diabetes, and affects the structure and the human body. Through this pre-processing that we have done, we will get all the results Good, smooth, focused, and high values to improve the performance of the workbook. To solve current data problems that suffer from missing values, the Mean measurement is used.

$$Mean(A_i) = \frac{A_1 + A_2 + ... + A_i}{N} \quad (1)$$

where A_i is the values of the i^{th} column and N is the number of records in the dataset.

3.3. Classification Stage through Ensembles and ML Methods

In the classification stage, ensemble algorithms such as bagging, boosting, voting, and stacking are used. In this stage, we apply these techniques to improve the classification performance and obtain a high-accuracy result with good features [15] [16]. This stage is the classification stage, where it gave good results by implementing this stage with the pre-processing. The highest accuracy reached 97.12%, which this value is considered the highest accuracy in this research. It is a high value that surpasses its precedents, and high accuracy is considered.

Boosting: It is one of the operations of the ensemble algorithms, where it divides the data in the form of aggregates for better prediction. This stage is divided into the training section, in which other algorithms are applied together, and it is layered with reinforcement by the decision tree algorithm, RF, and KNN. We applied this technique in this paper by dividing it into a test and training, where good results were obtained with pre-processing. The highest accuracy reached 97.12%. These values are the best among the obtained values in this article. This technique showed success equal to and superior to the algorithms used in this article rather than the previous works. This technique obtained better prediction and giving good values, and improve the performance of diabetes.

Voting: At this stage, the voting method works with the Rapid Miner tool. This technique is applied with/without pre-processing. It consists of two classifiers for training and testing, where it was applied with the decision tree algorithm. Then, it was implemented to give very high values with pre-processing, which gave high accuracy. It reached 12.97%, and this is a good value for improving the performance of classification and prediction in diabetes. It is considered a sub-process consisting of two main classifiers. This stage is considered one of the most vital stages for predicting in an imaginary and excellent way and giving high values. Voting divides data into groups, and each group is a classification with recall, precision, accuracy, F1, where their values reached without pre-processing 93.62, 94.37, 94.23, 93.99%, respectively.

Bagging: In this section, we use the process of mobilization, where this method divides the data into

two groups, the training group by 70%, and the second group, the test section, which amounts to 30%. The DT algorithms, RF and KNN, where the highest peak reached 97.12%. This value is considered good and one of the best values extracted in this paper as it predicts better predictors and proved to be better than others.

Stacking: At this stage, the stacking technique was applied, as it is one of the groups of algorithms that predicts well. This method divides the data into two parts, the part of the basic trainees and the part of the process of testing the learners, where this technique was applied with pre-processing and without pre-processing. The high values were received that predict disease diabetes, and this process is vital to improve classification performance. Then we were able to apply this technique and get good results for predicting diabetes. This technique trains learners and gives them special directions. The high results also exhibited our model reached 97.12% with pre-processing. Our model employed DT algorithms, KNN, and RF, which gave high values for predicting diabetes, improving classifier performance, and better prediction.

4. EXPERIMENTS AND RESULTS

Experiment I: In the first experience, we researched the effect of summation algorithm operations, which are considered one of the most important operations including bagging, stacking, boosting, and voting operations were applied. These were applied without pre-processing via the Rapid Miner tool. Table 1 shows good results for bagging, boosting, and stacking reached 97.12, 97.12, 97.12%, respectively. These reflected the best values because these were considered equal values, and the stacking value, as a result, reached 96.15% accuracy. It considered less than the rest of the ensemble operations, but its accuracy is also good and high, and it is widely predicted as well These results give good results and through them, the performance of annoying diabetes, which is now considered one of the annoying and harmful diseases, will be improved. In previous years, diabetes was considered a serious disease, and even now because it makes the patient less energy and less immune and weakens the strength of his body. Through our results in Table 1, we will predict better results.

Table 1 The obtained results through ensembles without pre-processing in conjunction with DT and RF through Rapid Miner tool.

	Precision	Recall	Accuracy	F1
Bagging	97.18%	96.72%	97.12%	97.40%
Boosting	97.18%	96.72%	97.12%	97.40%
Voting	97.18%	96.72%	97.12%	97.40%
Stacking	95.94%	95.94%	96.15%	92.04%

According to Table 1, we worked on it without pre-processing. Then we analyzed and worked to improve it through the use of ensemble algorithms including bagging, boosting, voting, and stacking was applied. The results in the Table 1 showed very high and an improvement of the prediction of this disease with the highest accuracy in Table 1 reaching 97.12%.

Experiment II: In the second process, we applied ensemble algorithms on missing-valued data and possessing stray values, where we used the techniques of bagging, boosting, voting, and stacking. We conducted these experiments in our second experiment to improve the performance of classification and predict diabetes well, where the results showed in the Table 2 without pre-processing results. The highest accuracy in the Table 2 reached 95.19%, where this value with reinforcement is considered. The best among the values for predicting diabetes and improving the performance gives us excellent results because diabetes is an incurable and chronic disease, and affects the general groups and is your concern because. It has no cure to remove it permanently, but its effect remains with you, and rarely do we see cases that have been completely cured.

Table 2. The obtained results through ensembles without pre-processing with DT and KNN with Rapid Miner tool.

	Precision	Recall	Accuracy	F1
Bagging	93.62%	94.37 %	94.23 %	93.99%
Boosting	94.75%	95.15%	95.19%	95.91%
Voting	93.62%	94.37 %	94.23 %	93.99%
Stacking	95.94%	95.94%	96.15%	92.04%

Experiment III: In this process, we explain our third experience of applying high-precision algorithms at work. It does not show good results at work and often gives good results to predict the best possible results. We worked on the Rapid Miner tool, working on the downloaded data and studying it related to diabetes, where we used filling, reinforcement, stacking, and voting. Table 3 show us the values of recall and precision and accuracy and f-measure with ensemble and RF, DT.

Table 3. The obtained results through ensembles with pre-processing with RF and DT with Rapid Miner tool.

	Precision	Recall	Accuracy	F1
Bagging	97.18%	96.72%	97.12%	97.14%
Boosting	97.18%	96.72%	97.12%	97.41%
Voting	97.18%	96.72%	97.12%	97.41%
Stacking	97.18%	96.72%	97.12%	97.41%

Our work showed excellent results. We obtained a

high accuracy of 97.12, 97.12, 97.12, 97.12% for bagging, boosting, stacking, and voting. These processes are well predicted and are among the most common techniques that improve classification performance.

According to Table 3, we used algorithms to give high values of accuracy, recall, precision, and F1, which showed high and distinct results. The highest accuracy values reached 97.12, 97.12, 97.12, and 97.12% for bagging, boosting, stacking, and voting, respectively. These are considered high values obtained from our work that was applied in this paper, where high-precision touches were made to improve and predict the annoying and widespread diabetes in previous centuries. It also indicates that our work is good and will give satisfactory results compared to others. It will also outperform them and give good results in terms of accuracy.

Experiment IV: In Table 4, we applied an experiment with pre-processing to improve and predict chronic diabetes. We implemented the collection operations on the data downloaded from the UCI, where this is a data set for diabetes, which is incomplete data and suffers from missing values and stray values. We used the techniques of enhancement, filling, voting, and stacking. Table 4 shows high-accuracy values with each of recall and precision, and accuracy and F1, where we obtained good values, as the highest value of accuracy reached 97.12% with stacking, and this value is considered the best value for improving diabetes and predicting the best results with such accuracy.

Table 4. The obtained results through ensembles with pre-processing with DT and KNN with Rapid Miner Tool.

	Precision	Recall	Accuracy	F1
Bagging	93.62%	94.37 %	94.23 %	93.99%
Boosting	94.75%	95.15%	95.19%	95.91%
Voting	93.62%	94.37 %	94.23 %	93.99%
Stacking	97.18%	96.72%	97.12%	97.25%

5. EVALUATION METRICS AND DISCUSSION

To evaluate, accuracy, precision, recall, and F1 were applied. These measures define in Table 5.

Table 5. Parameters definitions.

Criteria	Evaluation
Accuracy	$(TP + TN) / (P + N)$
Precision	$(TP) / (TP + FP)$
Recall	TP / P
F1	$\frac{2\times precision\times recall}{precision+recall}$

Here, we worked with the Rapid Miner tool with/without pre-processing with ensemble algorithms including bagging, boosting, stacking, and voting. These techniques show overwhelming success in achieving results as well with a great experience. We divided our work into two parts which each part was divided into two parts as well. In the first part, we worked by applying the ensemble techniques without pre-processing in conjunction with DT, RF. The Table 1 showed the values for recall, precision, accuracy, and F1 equal 97.18, 96.72, 97.12, 97.41%, respectively. This proves that our work is good, predicts well, and gives good results. In the Table 2, we applied the second branch of the first part, where it was applied with bagging, boosting, stacking, and voting without aggravating treatment as shown Good results. We reached the highest value in Table 2 ensemble algorithm where the accuracy with stacking reached 96.15%, and this is a good value that improves the performance of the classifier, and the prediction is better. In the Table 3, we use the second part, but at this stage, we used the ensemble techniques with pre-processing and with DT, RF. The results in the Table 3 gave high and satisfactory results, as the highest accuracy in the Table 3 reached 97.12% with bagging, boosting, stacking, and voting. It indicates that the highest value in our business and will significantly predict and improve rating performance. In the Table 4, the second branch of the second part, we have applied the techniques in the ensemble with pre-processing and with DT, KNN. The results showed our excellent results predicted very well. It reached the highest accuracy in 97.12% through stacking and reflected the best option because it gave better accuracy than the rest of the algorithms and predicted excellently.

In the Table 6, we compared our work with other works. Our work outperformed others and was equal to others in terms of accuracy, 97.12%. It is a high value for better prediction, as our work has proven that it is towards the right action and good results because the application of such algorithms means that you are doing a great task with good results and better prediction. It is confirmed that our work made great progress in showing the results.

Here, the fees range from one to ten. The ROC is for the best classification of our model without pre-processing. The ROC is for the best classification of our work with pre-processing. These graphs represent the best results that we obtained from applying the ensemble algorithms with/without preprocessing in conjunction with Dt, RF, and KNN classifiers. It showed our results with the Rapid Miner tool are good enough. These graphs show how we work and the results we obtained. They represent the lifeline of our work based on the data removed from the UCI website on diabetes. We obtained high values of accuracy and prediction significantly and improved the performance of the classifier.

Table 6. A comparison among the obtained results through ensembles with pre-processing and other works.

Work	Precision	Recall	Accuracy	F1
[2]	82.30%	82.10%	82.10%	82.05%
[9]	75.9%	76.3%	76.30%	76%
[11]	72%	74%	74%	72%
[12]	87.9%	87.8%	88.1%	87.83%
Our work	97.18%	96.72%	97.12%	97.40%

In the fourth table, we have a comparison among our work and previous works in this content. It will be significant to improve the performance, as the highest accuracy in our article reached 97.12%. Compared with others, our model is good enough and improved classification performance and works on its development and prediction more largely. Figure 2 shows a comparison among our results and other counterparts. Here, Figs 3-5 show the ROC for the best classification of the current authors with pre-processing.

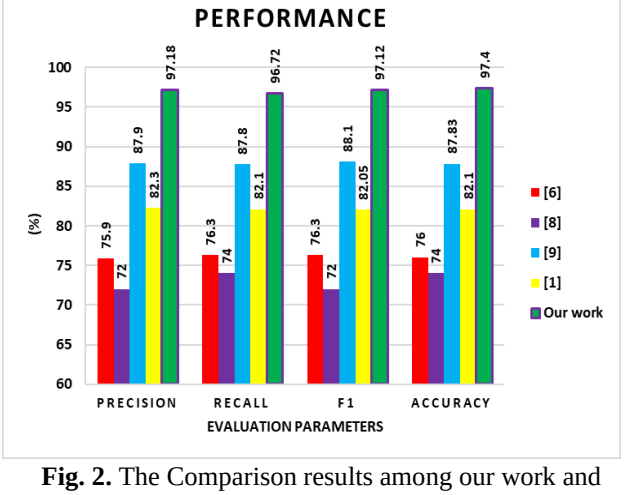


Fig. 2. The Comparison results among our work and others.

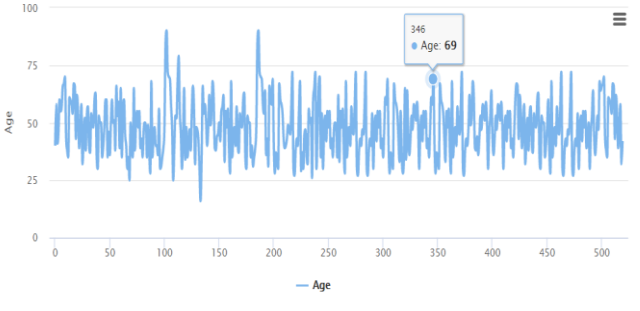
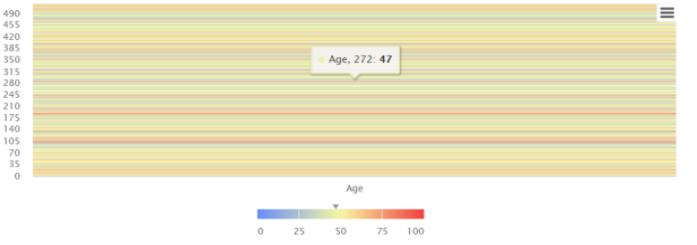
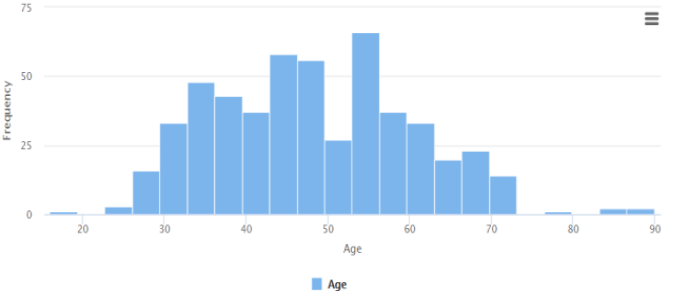


Fig. 3. The ROC for Bagging method with pre-processing spline.



Heat map



Histogram

Fig. 3. The ROC for Boosting method with pre-processing heat map and Histogram.

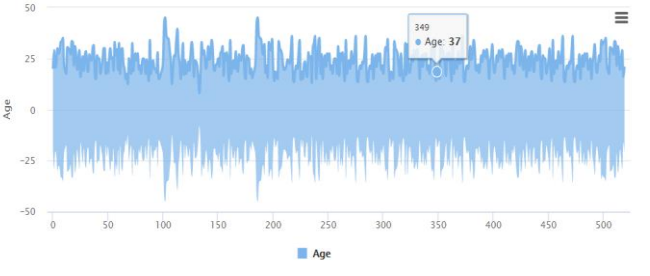


Fig. 4. The ROC for voting method with pre-processing stream graph.

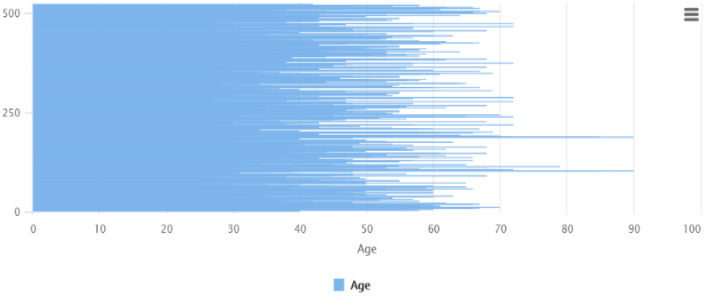


Fig. 5. The ROC for stacking method with pre-processing Bar (horizontal).

6. CONCLUSION

Diabetes is one of the most prevalent diseases, as it causes many chronic diseases and side effects. It influences all groups and does not depend on anyone. It is also recognized as a harmful disease if you do not

commit to complete prevention with it. It must take medicines continuously hitting insulin needles because these are important for disease diabetes. For this reason, this disease is considered annoying and chronic and affects all groups. It prompted researchers to research and work on high technologies and applications of the most prominent algorithms to obtain good results and forecast excellently.

In this paper, we proposed a model that improves the performance of diabetes, and this model depends on applying ensemble algorithms consist of boosting, bagging, stacking, and voting via the Rapid Miner. We applied these techniques with the following algorithms to obtain better results and predict well. These algorithms have relied on DT, RF, KNN with the ensemble. Our work is divided into two parts, each part consists of two branches, where we will explain this in detail. In the first part, we use ensemble without pre-processing via the Rapid Miner in conjunction with DT, RF. These techniques were used and the results showed good and better germination as recall, precision, accuracy, and F1 the measurement of these parts was 97.18, 96.72, 97.12, 97.42%, respectively. In the second part, we use ensemble without pre-processing via the Rapid Miner and in conjunction with DT, KNN. The results showed that the highest value belonged to stacking 95.94, 95.94, 96.15, 92.04% for recall, precision, accuracy, and F1, respectively. This is the highest value that predicts well and excellently and improves the performance of diabetes. Then in the third part, we use ensemble with pre-processing via the Rapid Miner and in conjunction with DT, RF. We got very high values, which is good for better prediction the highest accuracy. The highest results were 97.12, 97.12, 97.12, 97.12% for bagging, boosting, voting, and stacking. These values are considered the highest in our article, through which we will significantly predict and improve the performance of the classification. Then in the fourth part, we use ensemble with pre-processing with bagging, boosting, stacking, and voting via the Rapid Miner and in conjunction with DT, KNN. The high values were also obtained, and these values outperformed the previous works, and through them, we were able to reach a solution that satisfies everyone, predicts quickly, and works with high accuracy. The accuracy in the fourth table was 94.23, 95.19, 94.23 97.12% for bagging, stacking, voting, and boosting, respectively. Here, it turns out that the stacking value is the highest, and it predicts excellently and will give great results, while the rest is also good, but less than stacking. The techniques that we applied in this research showed an excellent model and gave wonderful results, and through which the performance of classification and prediction of diabetes was quickly improved. This model proved its superiority over the rest of the used techniques and previous works.

For future work, we will work on other algorithms and classifiers to acquire excellent results. Clustering is another alternative to work in conjunction with association rule and linear regression on the Hadoop platform.

REFERENCES

[1] R. Asgarnezhad and K. Ali Mohsin Alhameedawi, **"MVO-Autism: An Effective Pre-treatment with High Performance for Improving Diagnosis of Autism Mellitus**," Journal of Electrical and Computer Engineering Innovations (JECEI), 2021. https://doi.org/10.22061/jecei.2021.8109.480

[2] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, and A. Alhumam, **"Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning,"** Applied Sciences, vol. 11, no. 3, pp. 1173-1189, 2021.

[3] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz, **"An expert system for diabetes prediction using auto tuned multi-layer perceptron,"** in 2017 Intelligent systems conference (IntelliSys), 2017, pp. 722-728.

[4] R. Asgarnezhad, A. Monadjemi, and M. Soltanaghaei, **"NSE-PSO: Toward an Effective Model Using Optimization Algorithm and Sampling Methods for Text Classification,"** Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 8, pp. 183-192, 2020.

[5] R. Asgarnezhad, S. A. Monadjemi, and M. S. Aghaei, **"A new hierarchy framework for feature engineering through multi-objective evolutionary algorithm in text classification,"** Concurrency and Computation: Practice and Experience, 2021. https://doi.org/10.1002/cpe.6594

[6] M. M. Nentwich and M. W. Ulbig, **"Diabetic retinopathy-ocular complications of diabetes mellitus,"** World journal of diabetes, vol. 6, no. 3, pp. 489-532, 2015.

[7] T. Daghistani and R. Alshammari, **"Diagnosis of diabetes by applying data mining classification techniques,"** International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, pp. 329-332, 2016.

[8] R. Asgarnezhad, S. A. Monadjemi, and M. Soltanaghaei, **"FAHPBEP: A fuzzy Analytic Hierarchy Process framework in text classification,"** Majlesi Journal of Electrical Engineering, vol. 14, pp. 111-123, 2020.

[9] D. Sisodia and D. S. Sisodia, **"Prediction of diabetes using classification algorithms,"** Procedia computer science, vol. 132, pp. 1578-1585, 2018.

[10] M. S. Satu, S. T. Atik, and M. A. Moni, **"A novel hybrid machine learning model to predict diabetes mellitus,"** in Proceedings of International Joint Conference on Computational Intelligence, pp. 453-465, 2020.

[11] M. F. Faruque and I. H. Sarker, **"Performance analysis of machine learning techniques to predict diabetes mellitus,"** in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-4.

[12] R. Alshammari, N. Atiyah, T. Daghistani, and A. Alshammari, **"Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet,"** Online Journal of Public Health Informatics, vol. 12, 2020.

[13] R. T. Selvi and I. Muthulakshmi, **"Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system,"** Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 1717-1730, 2021.

[14] R. Asgarnezhad, A. Monadjemi, and M. Soltanaghaei, **"A High-Performance Model based on Ensembles for Twitter Sentiment Classification,"** Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 8, pp. 41-52, 2020.

[15] R. Asgarnezhad, S. A. Monadjemi, and M. Soltanaghaei, **"An application of MOGW optimization for feature selection in text classification,"** The Journal of Supercomputing, vol. 77, pp. 5806-5839, 2021.

[16] R. Asgarnezhad and S. A. Monadjemi, **"NB vs. SVM: A contrastive study for sentiment classification on two text domains,"** Journal of Applied Intelligent Systems & Information Sciences, 2021. https://doi.org/10.22034/JAISIS.2021.279225.1025.