SciNexuses

Journal Homepage: sciencesforce.com/scin

SciNexuses Vol. 1 (2024) 44–51

Paper Type: Original Article

Anticipating Diabetes using Fusion-Ensemble Machine Learning Techniques

Alber S. Aziz [1],*, Khaled Ibrahim [1], Ahmed Elsharkawy [1] and Nariman Khaliel [2]

[1] Faculty of Information Systems and Computer Science, October 6th University, Cairo, 12585, Egypt;
Emails: albershawky.csis@o6u.edu.eg; khaledahmed872003@gmail.com; ahmed.yaser.el.sharkawy@gmail.com.
[2] Assistant professor in ECCAT Suez University; narimankhaliel.eccat@suez.edu.eg.

Received: 15 Feb 2024 **Revised:** 28 Apr 2024 **Accepted:** 27 May 2024 **Published:** 29 May 2024

Abstract

Diagnosis of diabetes is still a complex task for most of the already existing machine learning methods. The objective of the study is to develop a decision support system for predicting the probability of diabetes disease. By using machine learning algorithms and considering fuzzy logic to handle uncertainty, a decision support system can be developed for predicting the probability of having diabetes in the given dataset. Diabetes patients will increase day by day; In current research, several algorithms have been used to predict diabetes. These are major issues for diabetes. Using these technologies, A model will be implemented to predict diabetes using different algorithms with proper comparison and find the best algorithm to predict diabetes. The prediction model is made using 11 classification algorithms from Ski learn, and their accuracies are compared. The expected result is that one of the best algorithms can be obtained for making a diabetes detection model. The fusion (ensemble) model is used for diabetes prediction, intended to improve the accuracy of classification. and use two algorithms to apply the fusion (ensemble) model, which picks the maximum accuracy of the list of classifiers with the rest of the classifiers. The source code is publicly accessible at https://github.com/diabetes-mellitus-implementation.

Keywords: Diabetes, Machine Learning, Risk Assessment, Decision Support System, Predictive Modeling, Fuzzy Logic.

1 | Introduction

Diabetes, or diabetes mellitus (DM), is a chronic metabolic disorder connected to impaired carbohydrate, protein, and fat metabolism. The most popular symptoms are frequent urination, increased thirst, and increased hunger. If diabetes is not treated, it can cause many complications. The highest rate is missed sick days, which ultimately reduces work productivity [1]. The reduced productivity will increase financial costs due to the expenditure on medical costs and drugs for the patient. The higher the diabetes rate in a country, the more the government must anticipate the complications. It needs a diabetes detection model that can detect when a person is suffering from or has the potential to have diabetes [2, 3]. A study from reference says that 33% of diabetes sufferers in a country reduce work productivity. The diabetes detection model will be useful for the government to prevent high complications. One way to make the detection model is by using classification data mining. Classification can be used to make a model that can predict the class or label of the attributes based on the training data, so the government can classify the members of civil society who

Corresponding Author: albershawky.csis@o6u.edu.eg

https://doi.org/10.61356/j.scin.2024.1307

have or have the potential to have diabetes [4, 5]. The diabetes data that was obtained from Kaggle is used by us for training. The attribute data consist of: plasma glucose concentration, a function that provides information on the level of glucose in the body; two hours of serum insulin; age; and class variable (1: tested positive for diabetes, 0: tested negative for diabetes) [6]. The prediction model is made using 12 classification algorithms from Ski learn, and their accuracies are compared. One of the best algorithms can be obtained for making a diabetes detection model. The fusion (ensemble) of more than one classification model was applied. The contribution of this paper is enhancing the accuracy by applying a fusion ensemble. The paper is organized as follows; Section two presents the motivation behind working on this topic. Section three gives the diabetes prediction proposed model is discussed. Section four gives the results of the experiment followed by concluding remarks.

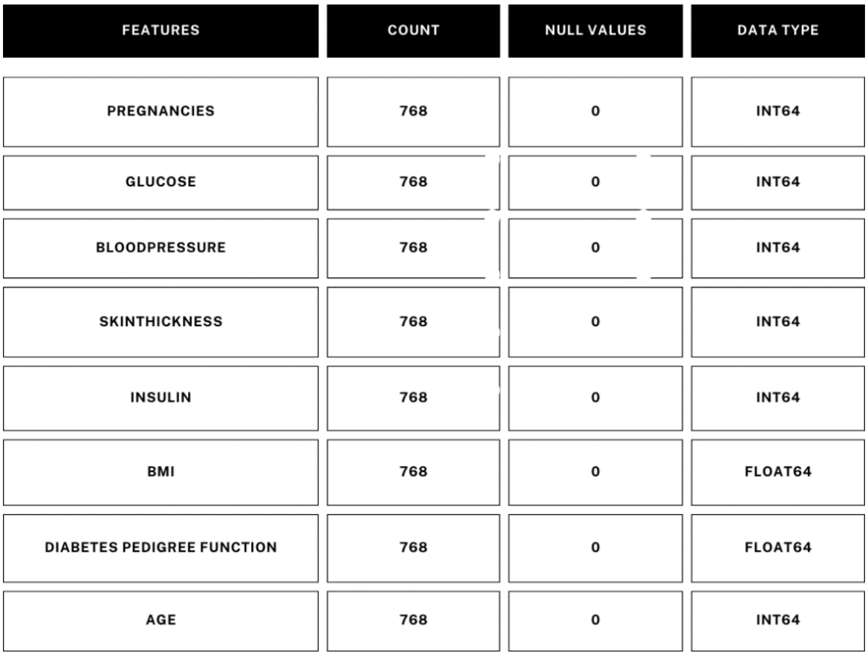
2 | Motivation

With the increase in the capabilities of machine learning, there have been many approaches to using machine learning to assist in diagnosis by classifying patients based on a set of attributes into the suspected diagnosis. This has been met with limited success as often the suspected diagnosis of a patient is difficult to confirm and the scope of possible diseases based on a set of attributes is too large, resulting in the machine simply choosing the most common disease. This occurs in the diagnosis of diabetes, where it is typically difficult to confirm a diagnosis and there are many different factors and types of diabetes which can lead to the wrong treatment if not identified correctly. This makes it an interesting field to apply machine learning and the instance of misclassification can have serious consequences.

3 | Proposed Model

3.1 | Data Collection and Preprocessing

The first step in our methodology involves gathering [7] a diverse dataset containing essential health parameters relevant to diabetes prediction. These parameters include but are not limited to glucose levels, blood pressure, BMI (Body Mass Index), insulin sensitivity, Skin Thickness, Pregnancy, Age, and Target Outcome [8]. To ensure data quality and consistency, The dataset will be preprocessed with missing values being handled, features normalized, and categorical variables encoded as needed. Figure 1 shows information about the used dataset while Figure 2 explains the number of positive and negative diabetes counts in the dataset.



FEATURES	COUNT	NULL VALUES	DATA TYPE
PREGNANCIES	768	0	INT64
GLUCOSE	768	0	INT64
BLOODPRESSURE	768	0	INT64
SKINTHICKNESS	768	0	INT64
INSULIN	768	0	INT64
BMI	768	0	FLOAT64
DIABETES PEDIGREE FUNCTION	768	0	FLOAT64
AGE	768	0	INT64

Figure 1. Information about the dataset.

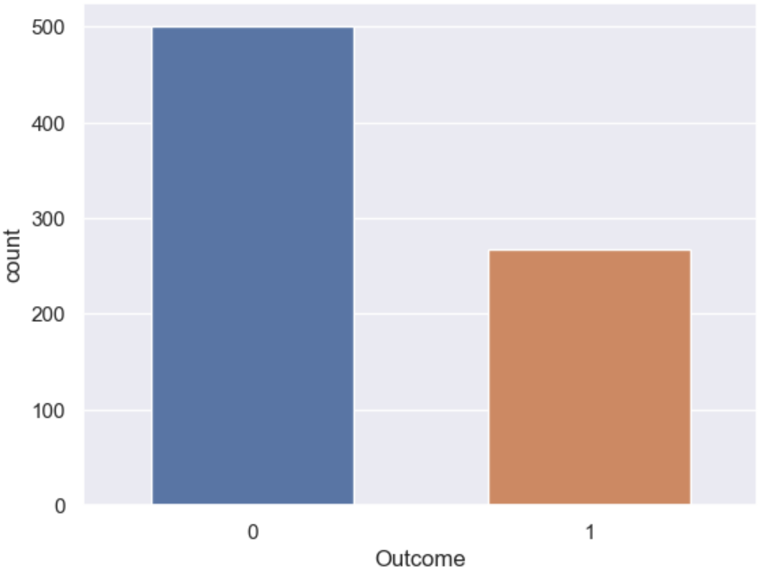
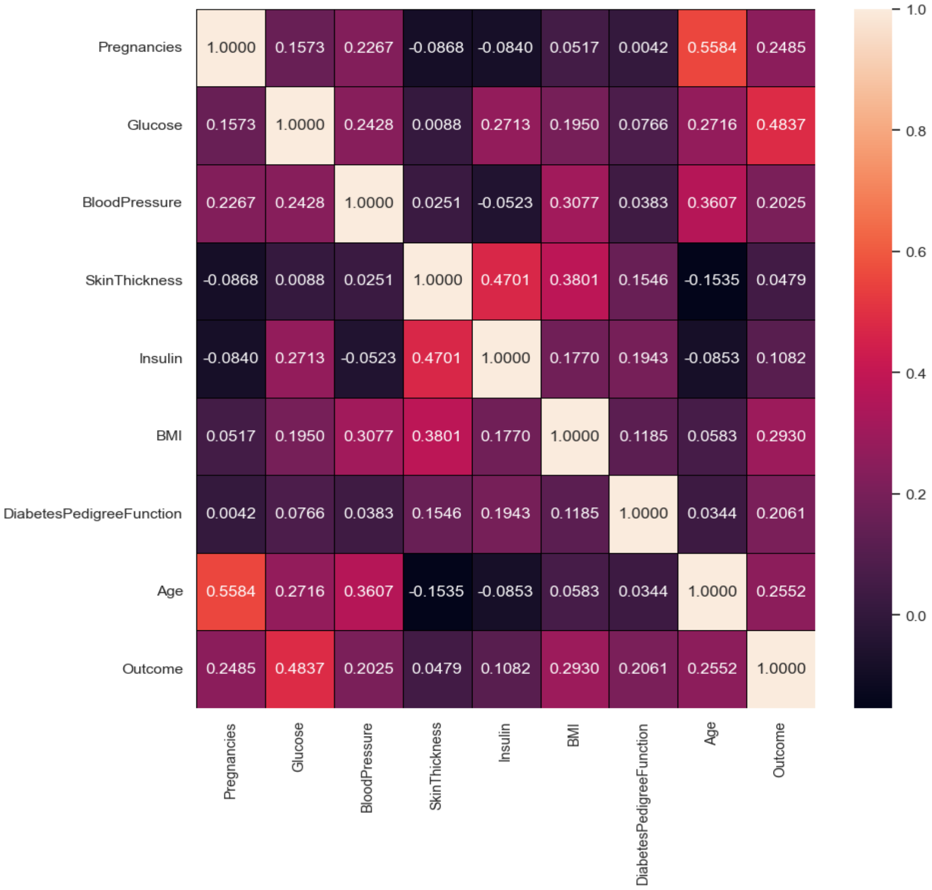


Figure 2. Count of positive and negative diabetes.

3.2 | Feature Selection

Techniques such as correlation analysis will be used [9]. The objective is to identify the most relevant variables that contribute significantly to our predictive model. This step is crucial for reducing dimensionality, enhancing efficiency, and improving the interpretability of the model. Figure 3 shows the correlation between features [10].



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.0000	0.1573	0.2267	-0.0868	-0.0840	0.0517	0.0042	0.5584	0.2485
Glucose	0.1573	1.0000	0.2428	0.0088	0.2713	0.1950	0.0766	0.2716	0.4837
BloodPressure	0.2267	0.2428	1.0000	0.0251	-0.0523	0.3077	0.0383	0.3607	0.2025
SkinThickness	-0.0868	0.0088	0.0251	1.0000	0.4701	0.3801	0.1546	-0.1535	0.0479
Insulin	-0.0840	0.2713	-0.0523	0.4701	1.0000	0.1770	0.1943	-0.0853	0.1082
BMI	0.0517	0.1950	0.3077	0.3801	0.1770	1.0000	0.1185	0.0583	0.2930
DiabetesPedigreeFunction	0.0042	0.0766	0.0383	0.1546	0.1943	0.1185	1.0000	0.0344	0.2061
Age	0.5584	0.2716	0.3607	-0.1535	-0.0853	0.0583	0.0344	1.0000	0.2552
Outcome	0.2485	0.4837	0.2025	0.0479	0.1082	0.2930	0.2061	0.2552	1.0000

Figure 3. The correlation between features.

3.3 | Feature Model Development

Machine learning algorithms effective in classification tasks will be explored. These algorithms include Logistic Regression, Random Forest, Decision Tree, SVC, and KNN. The preprocessed dataset will be used for training and fine-tuning these models, utilizing techniques such as a fusion classifier to optimize performance. The model will be developed using two main algorithms, single and fusion models as shown in the following flow charts. Figure 4 illustrates the pseudo-code for the single model while Figure 5 explains the flow chart for the same model. Figure 6 illustrates the pseudo-code for the ensemble model while Figure 7 explains the flow chart for the same model.

```
1- Define a list of classifiers, each containing a tuple with a name and a classifier object.
2- Define an empty dictionary to store classifier performance metrics.
3- For each classifier in the list of classifiers:
   3.1- Train the classifier on the training data.
   3.2- Make predictions on the test data.
   3.3- Calculate accuracy, precision, recall, and F1 score.
4- Display a chart showing the accuracy of each classifier.
```

Figure 4. Algorithm 1: Diabetes Prediction using single classification.

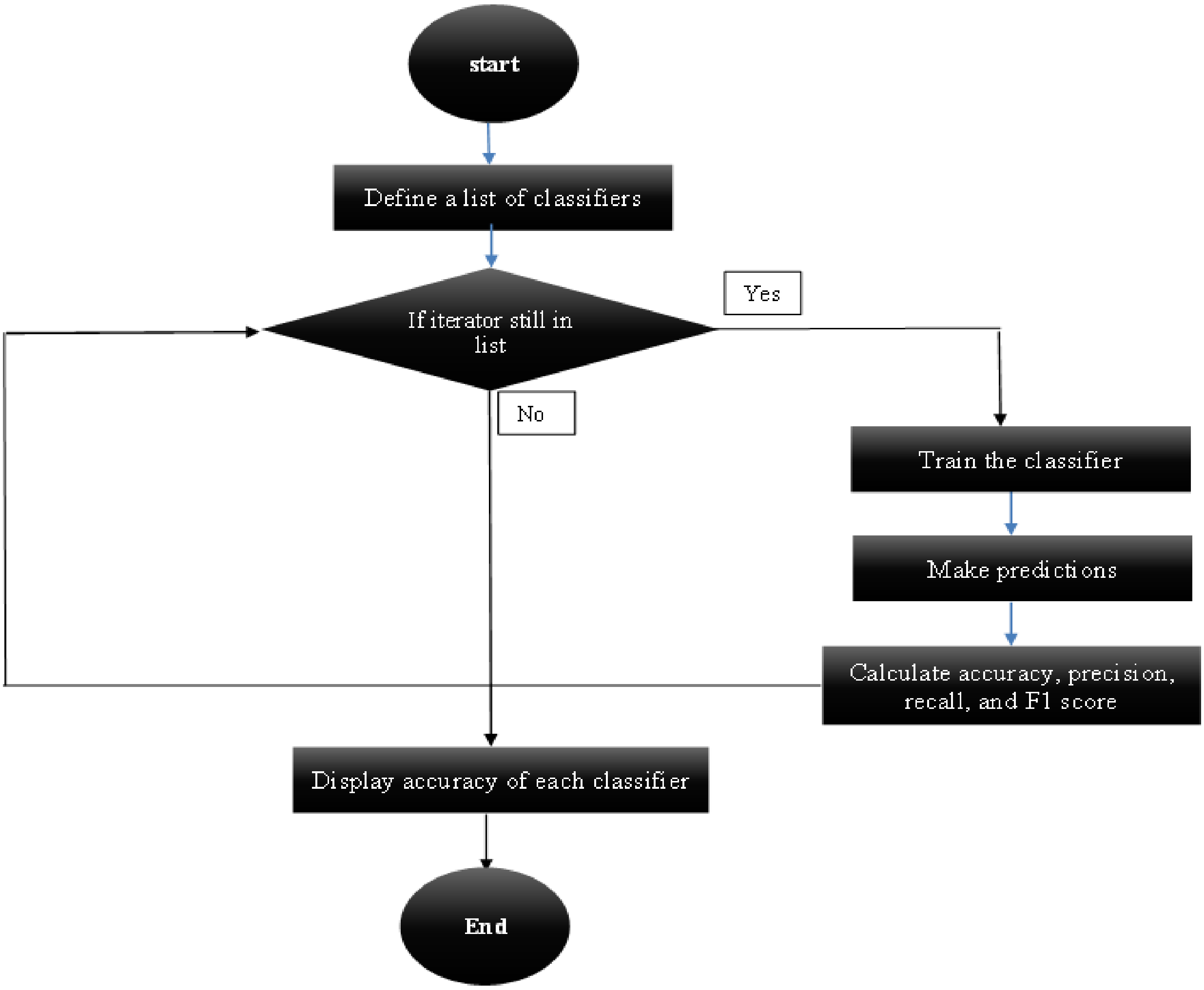

start
Define a list of classifiers
If iterator still in list
Yes
No
Train the classifier
Make predictions
Calculate accuracy, precision, recall, and F1 score
Display accuracy of each classifier
End

Figure 5. Single model classifiers.

```
1- Define a new dictionary to store the classifier names and their accuracies
2- Initialize variables to track maximum accuracy and the classifier(s) with maximum accuracy
3- For each classifier in the list of classifiers:
3.1- If the classifier is not the one with maximum accuracy:
  3.1.1- Create a list containing the current classifier and the classifier with maximum accuracy
  3.1.2- Combine a classifier with the maximum named ensemble classifier
  3.1.3- Train the ensemble classifier on the training data
  3.1.4- Make predictions on the test data
  3.1.5- Calculate accuracy
  3.1.6- Append the classifier name and accuracy to the dictionary
  3.1.7- Update maximum accuracy and the list of classifiers with maximum accuracy if applicable
4- Display a chart showing the accuracy of each ensemble classifier
```

Figure 6. Algorithm 2: Diabetes Prediction using fusion ensemble.

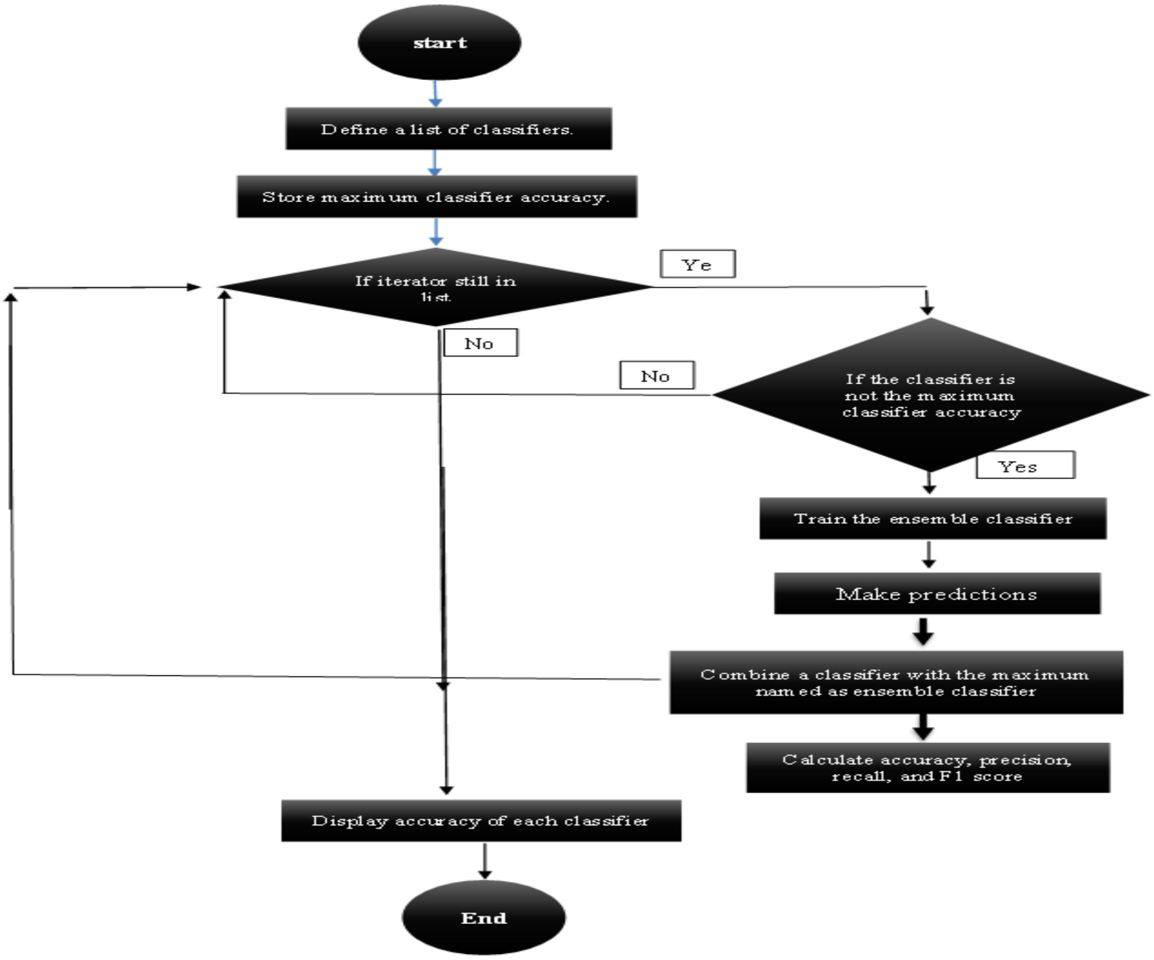

start
Define a list of classifiers.
Store maximum classifier accuracy.
If iterator still in list.
Ye
No
No
If the classifier is not the maximum classifier accuracy
Yes
Train the ensemble classifier
Make predictions
Combine a classifier with the maximum named as ensemble classifier
Calculate accuracy, precision, recall, and F1 score
Display accuracy of each classifier
End

Figure 7. Fusion model classifiers.

3.4 | Model Evaluation

To assess the performance of our predictive model, for binary classification tasks, A variety of evaluation metrics will be used. These metrics may include accuracy, precision, recall, F1-score, and confusion matrix analysis. By rigorously evaluating the model's performance, its reliability and effectiveness are ensured to identify individuals at risk of developing diabetes mellitus [11].

- Confusion Matrix: It gives us a matrix as output and describes the complete performance of the model.
- Precision: represents the ratio of accurate positive predictions to the total positive predictions made by the classifier.

- Recall: is the proportion of correct positive outcomes divided by the total number of relevant samples.
- F1 score is a metric for evaluating a test's accuracy, calculated as the harmonic mean of precision and recall. Ranging from 0 to 1, it indicates both the precision and robustness of the classifier.

4 | Experimental Results

4.1 | Result of Single-Model Classification

The machine learning algorithms were applied to the dataset, and the resulting accuracies are as follows: Random Forest gives the highest accuracy of 83%. Figure 8 shows the results of twelve distinct single classifiers from the most popular in this field of interest. The Random Forest classifier is the most superior one.

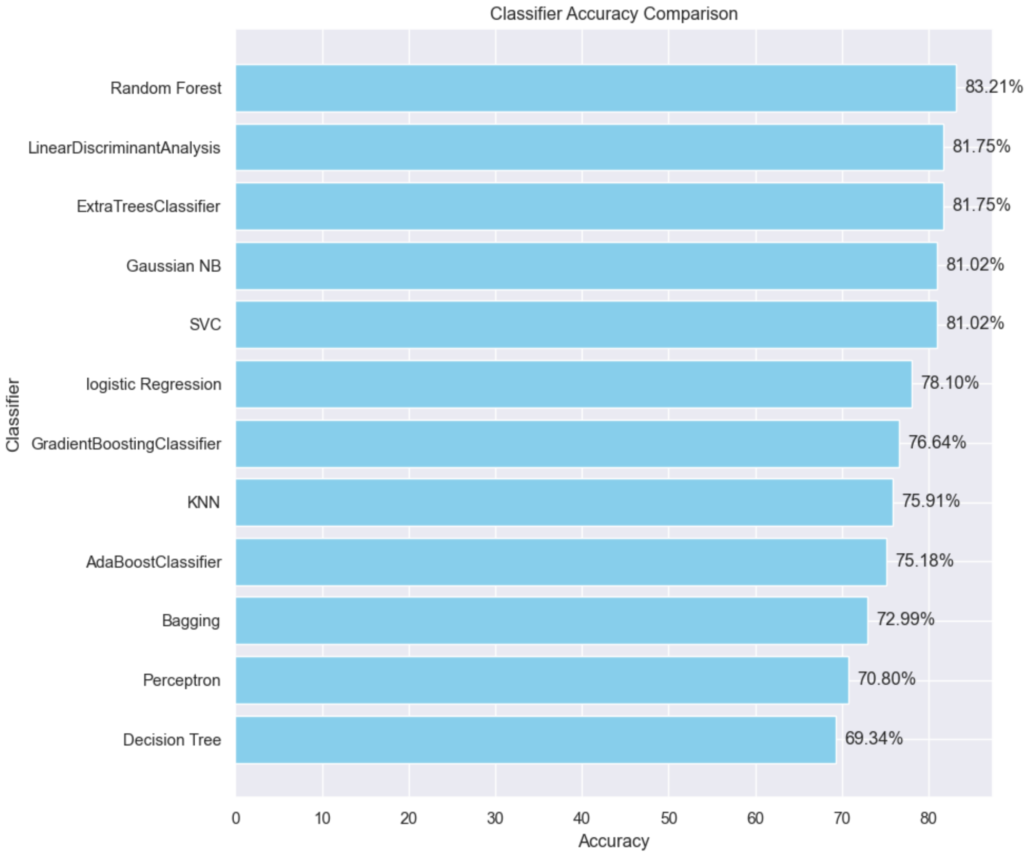


Figure 8. Accuracy results after applying a single classification.

4.2 | Result of Fusion Ensemble-Model Classification

Using the fusion (ensemble) method, the highest accuracy was achieved is 86% for Linear Discriminant Analysis and Random Forest Combines the highest classifier with the rest of the classifiers. Figure 9 shows the results of eleven distinct fused double classifiers with a Random Forest classifier.

This work contributed and applied the "fusion (ensemble)" method, which resulted in notable accuracies for Linear Discriminant Analysis and Random Forest at 86%. This ensemble method enhances the model's predictive capabilities by combining the strengths of multiple algorithms. And emphasizing the importance of various health parameters in diabetes prediction. Through data collection and preprocessing, we incorporated essential factors such as glucose levels, blood pressure, BMI, insulin sensitivity, Skin Thickness, Pregnancy, and Age into our model. Our research involved exploring and evaluating multiple machine learning algorithms known for their effectiveness in classification tasks. By comparing algorithms like Logistic Regression, Random Forest, Decision Tree, SVC, and KNN.

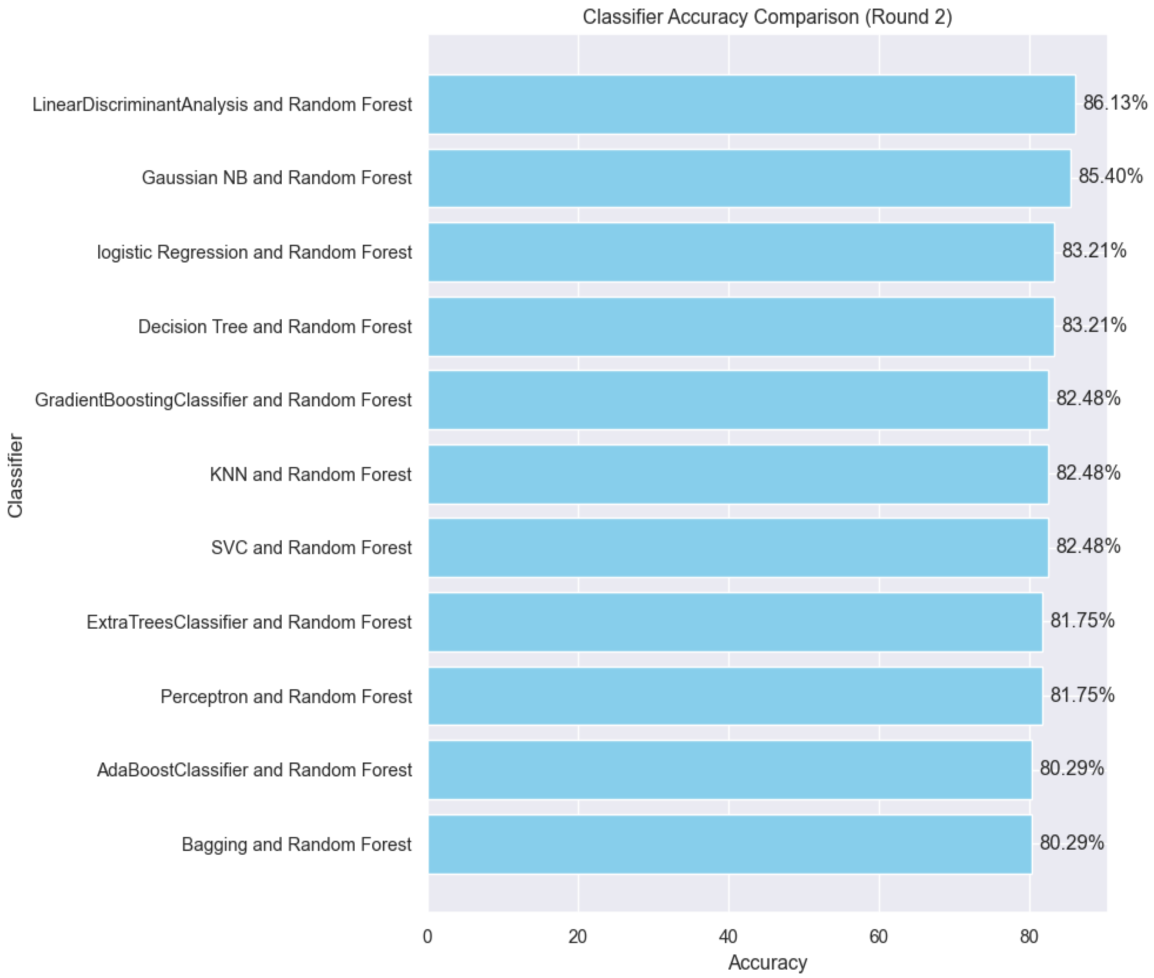


Figure 9. Accuracy results after applying fusion ensemble classification.

5 | Conclusion and Future Work

The paper presents a comprehensive approach to leveraging machine learning algorithms for the early detection and anticipation of Diabetes Mellitus. With the prevalence of diabetes escalating globally, innovative solutions are crucial for effective prevention and personalized management. Traditional diagnostic methods, while valuable, have limitations in accuracy and timeliness, highlighting the need for data-driven approaches like machine learning. The results of our model evaluation demonstrated promising accuracies, with Logistic Regression yielding the highest accuracy of 83%. Additionally, employing the fusion (ensemble) method led to an accuracy of 86% for Linear Discriminant Analysis and Random Forest. These findings underscore the potential of machine learning to revolutionize diabetes care, from early detection to personalized intervention strategies. As a future work, the fused classifiers may be combined with more counts than two.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

"Conceptualization, A.A. and N.K.; Methodology, A.E.; Software, K.I.; Validation, A.A., N.K. and A.E.; formal analysis, K.I.; investigation, A.E.; resources, A.A.; data maintenance, N.K.; writing-creating the initial design, N.K.; writing-reviewing and editing, K.I.; visualization, A.E.; monitoring, A.A.; project management, N.K.; funding procurement, A.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Gauri D. Kalyankar, Shivananda R. Poojara, and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Dr. Saravana Kumar N M, Eswari T, Sampath P, and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[9] Humar Kahramanli and Novruz Allahverdi," Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July 2008.

[10] B.M. Patil, R.C. Joshi, and Durga Toshniwal," Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[11] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm", Journal of Computing, Volume 3, Issue 12, December 2011.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.