

Diabetes Estimation using Machine Learning Techniques

Arkan Adnan Imran¹, Aya Khalid Hadi², Noor Najah Ali³, Diana S. Obaid³

^{1,2,3}Department of computer engineering techniques, University of Dijlah, Baghdad, Iraq

Arkan.adnan@duc.edu.iq, aya.khalid@duc.edu.iq, noor.najah@duc.edu.iq

Article Info

Article history:

Received :Sep., 15, 2024

Revised : Nov.,10, 2024

Accepted Dec., 2, 2024

Keywords:

Diabetes
Estimation
Accuracy
machine learning techniques

ABSTRACT

Machine learning techniques play an increasingly prominent role in medical diagnosis. With the use of these techniques, patients' data can be analyzed to find patterns or facts that are difficult to explain, making diagnoses more reliable and convenient. The purpose of this research was to compare the efficiency of diabetic classification models using four machine learning techniques: decision trees, random forests, support vector machines, and K-nearest neighbors. In addition, new diabetic classification models are proposed that incorporate hyper parameter tuning and the addition of some interaction terms into the models. These models were evaluated based on accuracy, precision, recall, and the F1-score. The results of this study show that the proposed models with interaction terms have better classification performance than those without interaction terms for all four machine learning techniques. Among the proposed models with interaction terms, random forest classifiers had the best performance, with 97.5% accuracy. The findings from this study can be further developed into a program that can effectively screen potential diabetes patients.

Corresponding Author: ARKAN A. AL-YASARI

Department of Computer Engineering Techniques, University of Dijlah

Al-masafi street, Baghdad, Iraq

Email: Arkan.adnan@duc.edu.iq

1. INTRODUCTION

Diabetes is a metabolic disease characterized by elevated blood sugar levels and organ damage such as kidney failure. when there is insufficient insulin production or utilization, which causes diabetes, computer programs make it easier to develop an IT system for identifying diseases based on clinical data . Based on genetics and diagnostic criteria as well as the region's epidemics, the most common type of diabetes in adults is mellitus. Pre-determined factors such as IFG or IGT are used to diagnose diabetes according to the American Diabetes Association (ADA). It can also be diagnosed if the blood glucose level is more than 200 mg/dL based on HbA1c, OGTT, and FPG [1].

Juvenile type 1 diabetes refers to insulin-dependence, and it occurs when the beta cells in the body cannot produce enough insulin. Beta cells vary from infants to adults (4–8). Other complications such as heart problems, nerve disorders, and kidney related complications can occur if glucose levels are not properly regulated in the body [9]. Usually, type 1 diabetes occurs under the age of 30, and patients are insulin-dependent, pumping insulin into their bodies. Type 2 diabetes mellitus is non-insulin-dependent and is otherwise determined as adult diabetes. It occurs due to the loss of secretion of insulin by β cells. Genetics, obesity, and an unhealthy lifestyle will contribute to the increase. Furthermore, type 2 diabetes occurs in middle age[2].

It can also occur prior to gestational diabetes in females and ethnic populations. Additionally, it has an impact on adolescents and children. Based on maternal characteristics during the second and third trimesters of pregnancy as well as biomarkers during the gestation period, gestational diabetes is anticipated. Recent studies have found that pregnant women of diverse ethnicities are at high risk for diabetes with modern technology, a massive amount of data is recorded, which facilitates the use of machine learning. Doctors are able to analyze a patient's disease using clinical metrics such as blood pressure and temperature and follow treatments after iterative analysis and refinement. Furthermore, artificial intelligence plays a significant role in fuzzy-based classification and diagnosis of disease using neural networks. Computer-aided comprehension is a factor that makes the situation worse, and the artificial NN ensemble is used for better disease diagnosis [3].

Although there is no cure for diabetes, early diagnosis can help people with both types of diabetes manage it and its health complications. People with prediabetes can take charge to help prevent it from becoming type II diabetes. The chances of developing each type of diabetes depends on a combination of risk factors. Until now, it has not been clear what causes type 1 diabetes, and how to prevent it is still unknown. One of the known risk factors is having a parent, brother, or sister with type 1 diabetes. Although people can be diagnosed with type 1 diabetes at any age, it is usually found in children, teens, or young adults. Unlike type 1 diabetes, there is more information about the risk factors for type 2 diabetes. They include having prediabetes, overweight or obesity, being aged 45 years or older, having a family history of diabetes, and being physically active less than three times a week [4]. The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2022) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization.[4] Aiswarya Iyer (2019) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result.[5] K. Rajesh and V. Sangeetha (2023) used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently.[8] Humar Kahramanli and Novruz Allahverdi (2021) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes.[9] B.M. Patil, R.C. Joshi and Durga Toshniwal (2022) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used.[10] Mani Butwall and Shraddha Kumar (2019) proposed a model using Random Forest Classifier to forecast diabetes behaviour.[7] Nawaz Mohamudally1 and Dost Muhammad (2021) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes.[11] Machine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised learning, Semi-supervised learning.

In this paper we solve the problem of the difficulty of distinguishing diseases, especially people with diabetes, as well as not knowing and distinguishing the affected person by analyzing his data previously registered in the database. The system also solves the problem of delay in detecting diabetes, as the algorithms are fast and highly efficient in distinguishing diseases.

2. METHOD

Figure 1 shows the main steps in the project, from entering information to classifying diabetes using the algorithms that have been programmed.

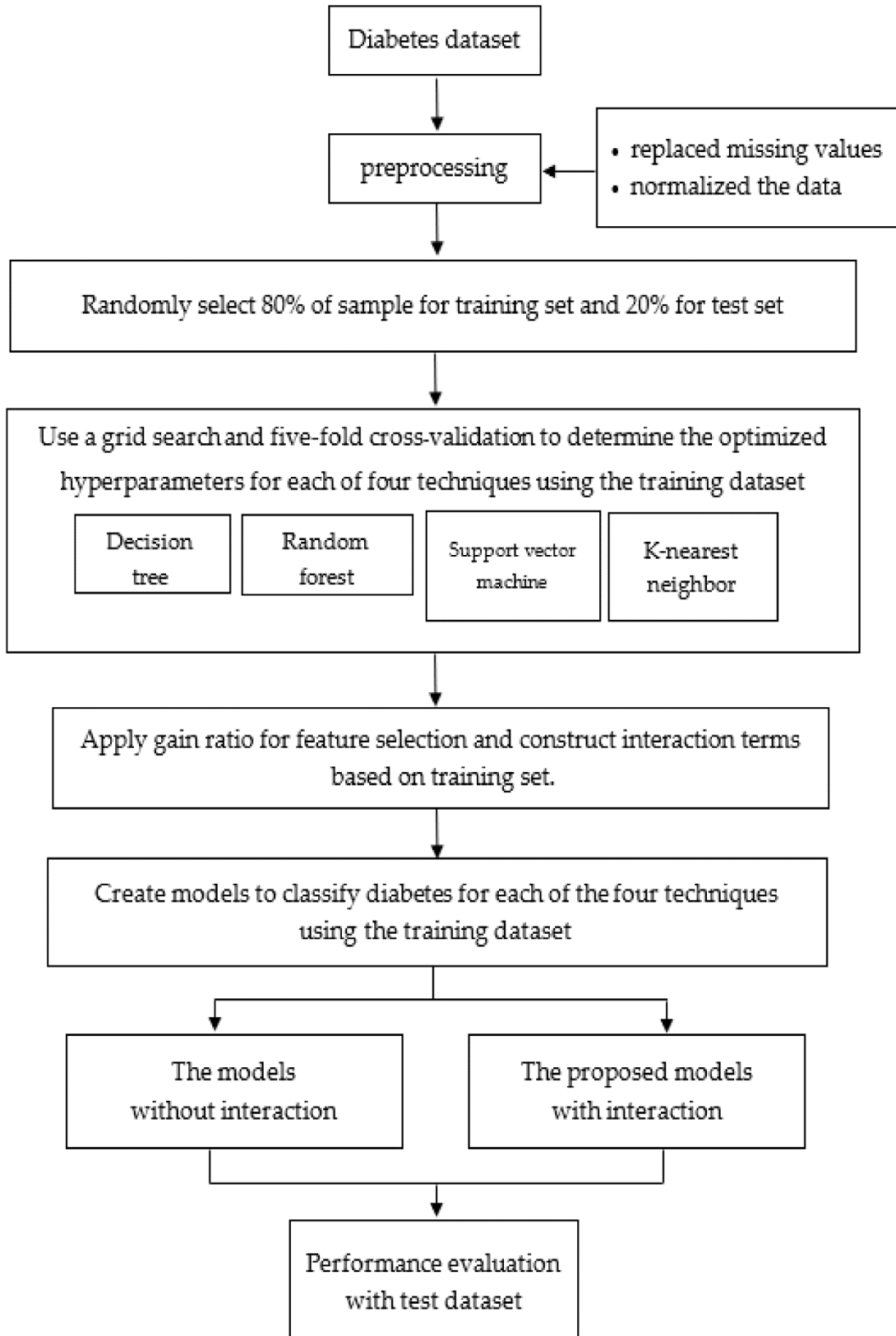


Figure 1: Proposed Flowchart of Diabetes estimation using machine learning techniques

A dataset is a collection of data from [16]. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes.

After opening the Matlab program and implementing the project, The algorithm will analyze the data at the level of males and females and give a result as shown in the following figure 2.

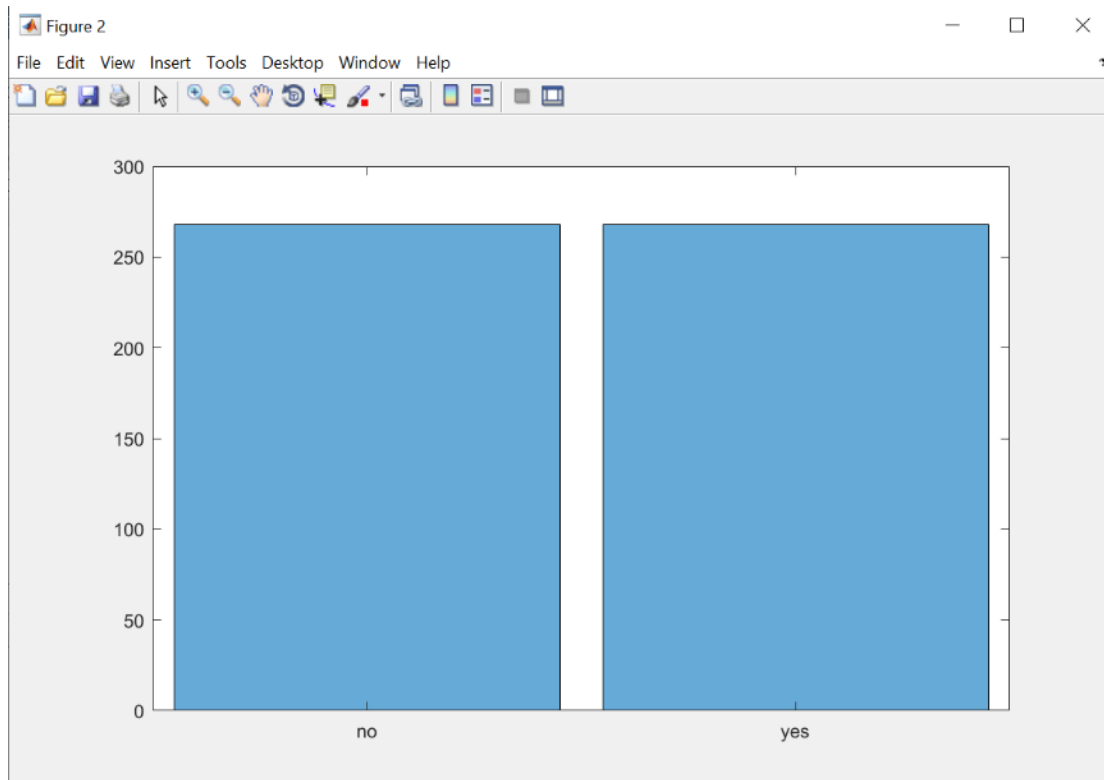


Figure 2: Dataset from [16]

The classification of data using machine learning algorithms is of high accuracy, as the data of people with diabetes has been classified through basic information to determine who has the disease or not, as show in Figure 2.

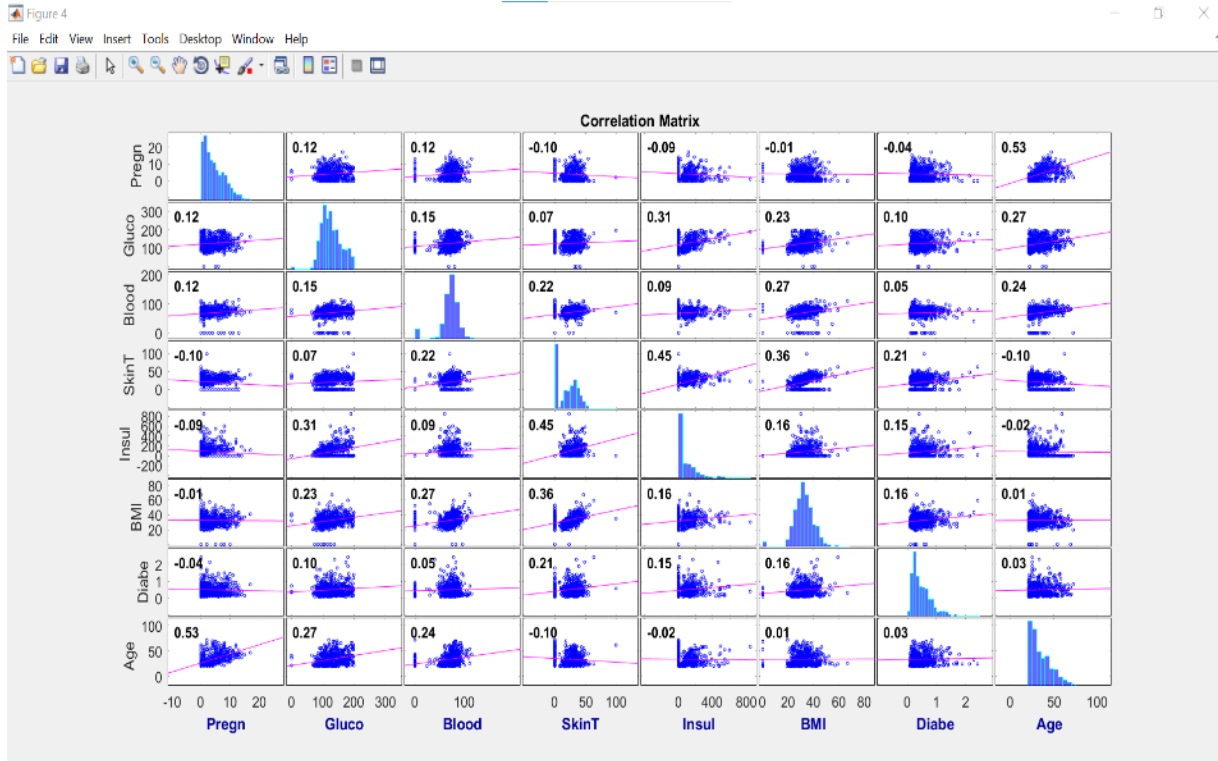


Figure 3: dataset Classification

After the analysis of the data is done, the algorithm reads the data set and gives the percentage of the data that are infected with diabetes and those that are not infected, as show in Figure 4.

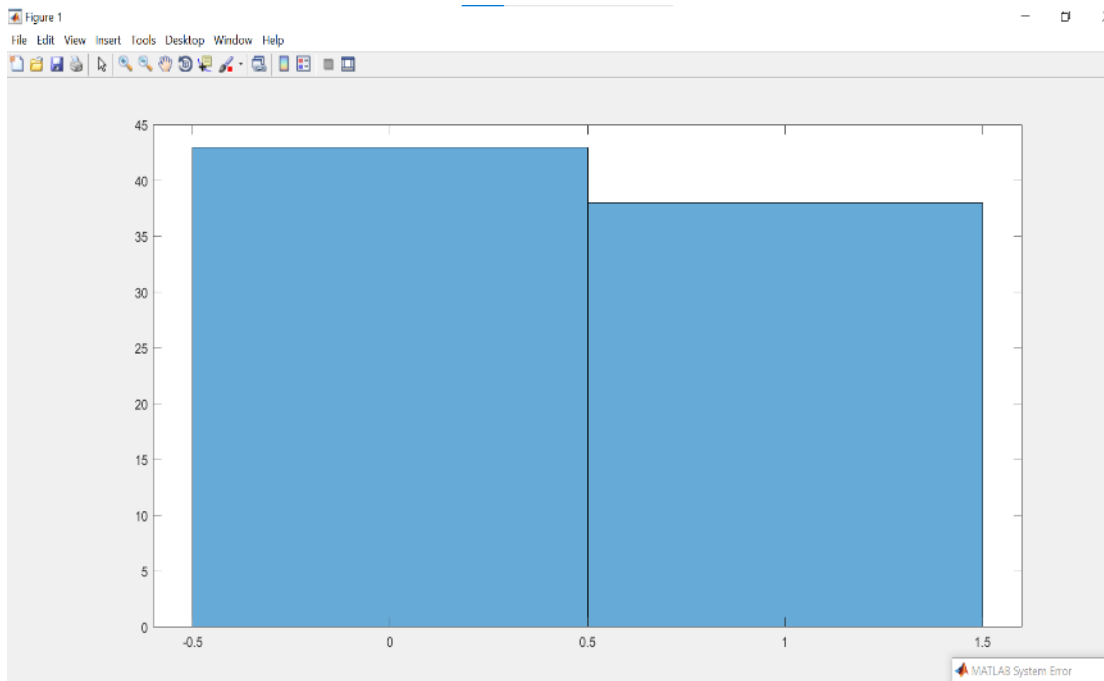


Figure 4: percentage of the dataset

3. RESULTS AND DISCUSSION

3.1 Hyper parameters for Machine Learning Techniques

In order to determine the hyper parameters for all machine learning techniques, we applied a grid search technique and five-fold cross-validation to the training dataset and compared the classification results based on accuracy, precision, recall, and the F1-score.

The hyper parameters used in the decision tree are as follows:

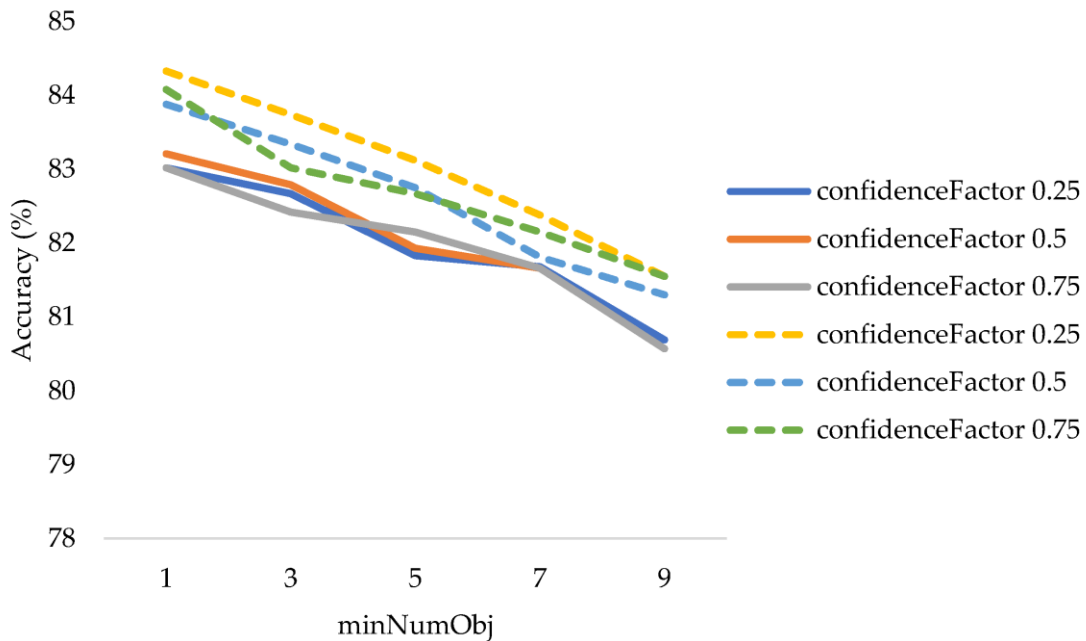
- Confidence Factor refers to the confidence intervals used in branching.
- MinNumObj refers to the minimum amount of learned information in the leaf node.

The hyper parameters for the models with and without interaction terms are shown in Table 1.

Table 1:The hyper parameters for the decision tree model.

Hyperparameter	Hyperparameter Value
confidenceFactor	0.25, 0.5, 0.75
minNumObj	1, 3, 5, 7, 9

Based on the result of five-fold cross-validation of the models without interaction, the hyperparameters that yielded the highest accuracy were confidenceFactor = 0.5 and minNumObj = 1, which provided an accuracy of 83.02%. Regarding the proposed models with interaction, the values were confidenceFactor = 0.25 and minNumObj = 1, which provided an accuracy of 97.08%, as shown in Figure 5.



----- : proposed models with interaction ——— : models without interaction

Figure 5: The accuracy of the models

3.2 Support Vector Machine

The hyper parameters used in the support vector machine are as follows:

- C refers to the regularization parameter.
- Kernel refers to the different types of mathematical functions, such as linear, polynomial, and RBF (radial basis function).
- Exponent refers to the exponent of the polykernel.
- Gamma refers to the hyper parameter that influences the learning dataset of the RBF kernel.

The hyper parameters for the models with and without interaction terms are shown in Table 2.

Table 2. The hyper parameters for the support vector machine model.

Hyperparameter	Hyperparameter Value
C	5, 10, 15, ..., 50
kernel	polykernel (exponent = 1), polykernel (exponent = 2, ..., 5), RBF
exponent	2, 3, 4, 5
gamma	0.05, 0.1, 0.2, 0.5, 1

According to the grid search, different kernels yield different optimal C values as follows:

- kernel = polykernel (exponent = 1); C = 5.
- kernel = polykernel (exponent = 2); C = 5.
- kernel = RBF, C = 10, gamma = 0.1.

Based on the five-fold cross-validation results of the models with and without interaction terms, the hyperparameters that yield the highest accuracy were kernel = polykernel (exponent = 2) and C = 5. The case in which the proposed models with interaction terms provided an accuracy of 78.11% and the models without interaction terms provided an accuracy of 77.79% is shown in Figure 6.

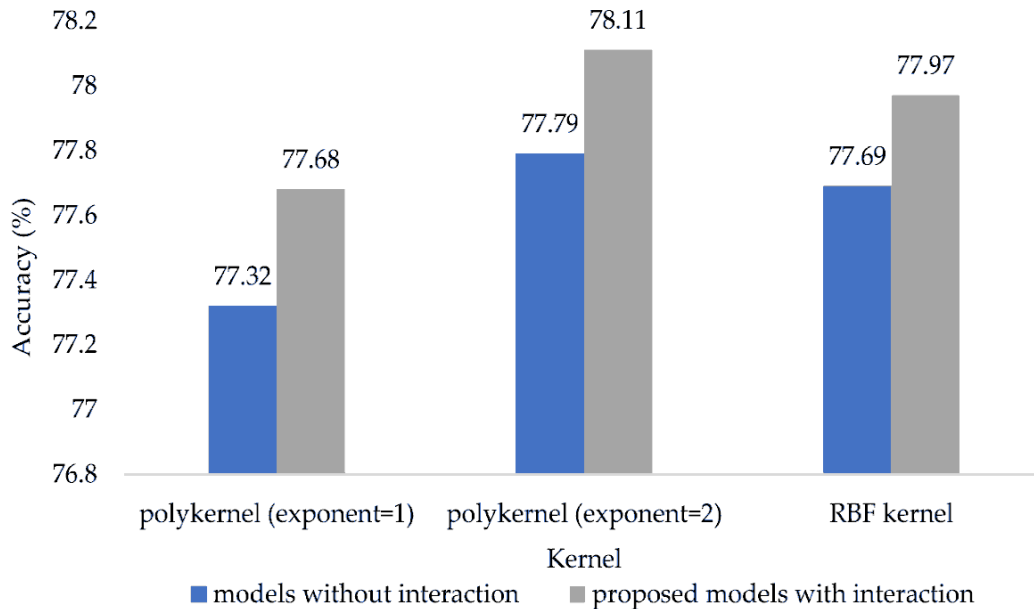


Figure 6: The accuracy of the models with and without interaction terms

4. CONCLUSION

Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results. The result, which only used fasting glucose, has a better performance especially in Luzhou dataset. It means that the fasting glucose is the most important index for predict, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes. In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree and neural network, but random forests are obviously better than the another classifiers in some methods. The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important. Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes.

ACKNOWLEDGEMENTS (10 PT)




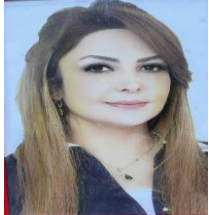
Author thanks Dijlah University for all type of support in this work.

REFERENCES

- [1] Griffin, P.; Rodgers, M.D. Type 1 Diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. Available online: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes> (accessed on 14 April 2023).
- [2] Griffin, P.; Rodgers, M.D. Risk Factors for Type 2 Diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. Available online: <https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes> (accessed on 14 April 2023).
- [3] Wei, S.; Zhao, X.; Miao, C. A comprehensive exploration to the machine learning techniques for diabetes identification. In Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 5–8 February 2018; pp. 291–295.

- [4] Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus with Machine Learning Techniques. *Front Genet.* 2018, 9, 515. [Google Scholar] [CrossRef] [PubMed]
- [5] Sneha, N.; Tarun, G. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* 2019, 6, 13.
- [6] International Statistical Classification of Diseases and Related Health Problems 10th Revision. Available online: <https://icd.who.int/browse10/2019/en#/E10-E14> (accessed on 29 April 2023).
- [7] Changpetch, P.; Pitpeng, A.; Hiriote, S.; Yuangyai, C. Integrating Data Mining Techniques for Naïve Bayes Classification: Applications to Medical Datasets. *Computation* 2021, 9, 99.
- [8] Laiteerapong, N.; Karter, A.J.; Liu, J.Y.; Moffet, H.H.; Sudore, R.; Schillinger, D.; John, P.M.; Huang, E.S. Correlates of quality of life in older adults with diabetes: The Diabetes & Aging Study. *Diabetes Care* 2011, 34, 1749–1753.
- [9] Davidson, K.W.; Barry, M.J.; Mangione, C.M.; Cabana, M.; Caughey, A.B.; Davis, E.M.; Donahue, K.E.; Doubeni, C.A.; Krist, A.H.; Kubik, M.; et al. Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Stateme. *JAMA* 2021, 326, 736–743.
- [10] Deepti, S.; Dilip, S.S. Prediction of Diabetes using Classification Algorithms. In Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Gurugram, India, 7–8 April 2018; pp. 1578–1585.
- [11] Hafeez, M.A.; Rashid, M.; Tariq, H.; Abideen, Z.U.; Alotaibi, S.S.; Sinky, M.H. Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm. *Appl. Sci.* 2021, 11, 6728.
- [12] Dimas, A.A.; Naqshauliza, D.K. Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. *Int. J. Emerg. Trends Eng. Res.* 2020, 8, 1689–1694.
- [13] Maneerat, P. WEKA Data Mining Program. 2012. Available online: <https://maneerat-paranan.blogspot.com/2012/02/weka.html> (accessed on 14 April 2023).
- [14] Yang, H.; Luo, Y.; Ren, X.; Wu, M.; He, X.; Peng, B.; Deng, K. Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 2021, 75, 140–149.
- [15] Guasch-Ferre, M.; Hruby, A.; Toledo, E.; Clish, C.B.; Martínez-González, M.A.; Salas-Salvado, J.; Hu, F.B. Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care* 2016, 39, 833–846.
- [16] [Open Datasets and Machine Learning Projects | Kaggle](#), 2024

BIOGRAPHIES OF AUTHORS

	<p>Asst. Lec. Arkan Adnan Imran. Received his MSc degree in Electrical and Computer Engineering from Altinbas University, Istanbul, Turkey, in 2022. He has been a full-time assist lecturer in the Computer Engineering Techniques Department at Dijlah University College, Baghdad, since March 2015. Currently, he serves as the Representative of the Optical Technologies Department at Dijlah University College, Baghdad, Iraq, a position he has held since 2023. He can be contacted via email: arkan.adnan@duc.edu.iq.</p>
	<p>Asst. Lec. Aya Khalid Hadi. Received her MSc degree in Computer Techniques Engineering from Middle Technical University, Electrical Engineering Technical College, Baghdad, Iraq, in 2023. She serves as a full-time lecturer in the Computer Engineering Techniques Department at Dijlah University College, Baghdad, since February 2024. She can be contacted via email: aya.khalid@duc.edu.iq</p>
	<p>Asst. Lec. Noor Najah Ali. Received her MSc degree in Computer Techniques Engineering from Middle Technical University, Electrical Engineering Technical College, Baghdad, Iraq, in 2023. She serves as a full-time lecturer in the Computer Engineering Techniques Department at Dijlah University College, Baghdad, since May 2024. She can be contacted via email: noor.najah@duc.edu.iq</p>
	<p>Diana Sabah Obaid is a lecturer in Computer Engineering Techniques Department - Dijlah University College – Baghdad – Iraq. She received B.Sc. degrees in Electrical Engineering Department, University of Baghdad, Baghdad, Iraq in 2005 and an M.Sc. degree in Electrical Engineering Department, University of Technology, Baghdad, Iraq in 2023. Mrs. Diana is interested in Renewable Energy, PV Systems, power electronics circuits, Optimization and Control of Microgrid system, and power management. Email: diana.sabah@duc.edu.iq</p>

الخلاصة

تلعب تقنيات التعلم الآلي دورًا بارزًا بشكل متزايد في التشخيص الطبي. باستخدام هذه التقنيات، يمكن تحليل بيانات المرضى للعثور على أنماط أو حقائق يصعب تفسيرها، مما يجعل التشخيص أكثر موثوقية وملاءمة. كان الغرض من هذا البحث مقارنة كفاءة نماذج تصنيف مرضى السكري باستخدام أربع تقنيات للتعلم الآلي: أشجار القرار، والغابات العشوائية، وآلات المتجهات الداعمة، وأقرب جيران K. بالإضافة إلى ذلك، تم اقتراح نماذج تصنيف جديدة لمرضى السكري تتضمن ضبط المعلمات الفائقة وإضافة بعض مصطلحات التفاعل إلى النماذج. تم تقييم هذه النماذج على أساس الدقة والإحكام والتذكر ودرجة F1. تظهر نتائج هذه الدراسة أن النماذج المقترحة مع مصطلحات التفاعل لديها أداء تصنيف أفضل من تلك التي لا تحتوي على مصطلحات تفاعل لجميع تقنيات التعلم الآلي الأربع. من بين النماذج المقترحة مع مصطلحات التفاعل، كان لمصنفات الغابات العشوائية أفضل أداء، بدقة 97.5%. يمكن تطوير النتائج من هذه الدراسة إلى برنامج يمكنه فحص مرضى السكري المحتملين بشكل فعال.