Turkish Journal of Engineering
https://dergipark.org.tr/en/pub/tuje
e-ISSN 2587-1366

Effect of dimension reduction with PCA and machine learning algorithms on diabetes diagnosis performance

Yavuz Bahadır Koca [*1], **Elif Aktepe** [2]

[1]*Afyon Kocatepe University, Department of Electrical Engineering, Türkiye, ybkoca@gmail.com*
[2]*Afyon Kocatepe University, Department of Electronics and Automation, Türkiye, eaktepe@aku.edu.tr*

Cite this study: Koca, Y. B., & Aktepe, E. (2024). Effect of dimension reduction with PCA and machine learning algorithms on diabetes diagnosis performance. Turkish Journal of Engineering, 8 (3), 447-456

https://doi.org/10.31127/tuje.1413087

Keywords
Diabetes
Data analysis
Machine learning
PCA
Random forest

Research Article

Received: 01.01.2024
Revised: 23.01.2024
Accepted: 29.01.2024
Published:05.07.2024

Abstract

Diabetes, a long-term metabolic disorder, causes persistently high blood sugar and presents a significant global health challenge. Early diagnosis is of vital importance in mitigating the effects of diabetes. This study aims to investigate diabetes diagnosis and risk prediction using a comprehensive diabetes dataset created in 2023. The dataset contains clinical and anthropometric data of patients. Data simplification was successfully applied to clean unnecessary information and reduce data dimensionality. Additionally, methods like Principal Component Analysis were applied to decrease the number of variables in the dataset. These analyses rendered the dataset more manageable and improved its performance. In this study, a dataset encompassing health data of a total of 100,000 individuals was utilized. This dataset consists of 8 input features and 1 output feature. The primary objective is to determine the algorithm that exhibits the best performance for diabetes diagnosis. There was no missing data during the data preprocessing stage, and the necessary transformations were carried out successfully. Nine different machine learning algorithms were applied to the dataset in this study. Each algorithm employed various modelling approaches to evaluate its performance in diagnosing diabetes. The results demonstrate that machine learning models are successful in predicting the presence of diabetes and the risk of developing it in healthy individuals. Particularly, the random forest model provided superior results across all performance metrics. This study provides significant findings that can shed light on future research in diabetes diagnosis and risk prediction. Dimensionality reduction techniques have proven to be valuable in data analysis and have highlighted the potential to facilitate diabetes diagnosis, thereby enhancing the quality of life for patients.

1. Introduction

High blood sugar, the most important characteristic marker of diabetes causes severe damage, especially to kidneys, eyes, nerves and heart [1,2]. Type 2 diabetes, usually seen in adults, occurs when the body becomes resistant to insulin or cannot produce enough insulin. Presently, prevalence of type 2 diabetes has increased significantly [1,3]. Approximately 422 million people worldwide have diabetes, and the majority live in low- and middle-income countries [4]. Every year, 1.5 million people die due to diabetes-related causes. Diabetes cases and prevalence have been increasing steadily in recent years [5]. For people living with diabetes, access to affordable treatment and medications such as insulin is vital. There is a global target to stop the increase in diabetes and obesity by 2025 [6,7].

The key to living a good life with diabetes is early diagnosis. People living with undiagnosed and untreated diabetes often have worsened health outcomes. Therefore, it is important that basic diagnostics, such as blood sugar testing are easily accessible. By analysing diabetes related data Machine Learning (ML) can offer advanced methods for early diagnosis [8]. Additionally, by using the knowledge that patients with diabetes need periodic expert evaluations and treatment a ML supported system can play an important role in monitoring and managing patients. In recent years, ML based approaches have come to the fore in many health and medical application fields.

Different approaches are employed in ML to process data and derive meaningful insights. These ae supervised, unsupervised, semi-supervised and reinforcement. Supervised learning involves training

models using labelled datasets, associating input features with specific outcomes. For instance, a dataset could use individual attributes like weight and height to predict the onset of health conditions such as diabetes. On the other hand, unsupervised learning focuses on pattern recognition and exploration without predefined target variables. In this context, it is effective for identification to the detection of new disease mechanisms, genotypic variations or phenotypic patterns. Semi-supervised learning utilizes datasets containing both labelled and unlabelled information. It aims to strike a balance between the two. This approach is particularly effective in heterogeneous data structures, leveraging the advantages of both supervised and unsupervised techniques. Reinforcement learning stands apart by combining concepts from both supervised and unsupervised learning methodologies. And it is based on trial-error exploration rather than solely relying on structured data, reflects the human learning experience. Through iterative learning and feedback mechanisms, reinforcement learning optimizes decisions and actions to achieve desired outcomes. This process results in outputs closely resembling human learning through adaptation and interaction.

Machine learning offers an approach to creating and evaluating learning ability based on data obtained regarding diabetes [9]. In recent years, the use of ML algorithms for analysis and prediction in the field of healthcare has attracted great attention among researchers. ML algorithms increase diagnostic accuracy thanks to their ability to handle large data sets by combining information from various data sources. These algorithms have the potential to provide more accurate, faster and more economical results in health diagnosis [8]. With the combination of wearable technologies and embedded systems, real-time monitoring opportunities in the field of healthcare have increased significantly. In this way, human health parameters can be monitored instantly instead of periodic follow-ups. The integration of wearable technologies and management of diseases such as blood pressure and diabetes offer the opportunity to provide a more comfortable life for individuals. For the field of diabetes, instantaneous changes in glucose levels and other health parameters have become available with wearable sensors placed on the body instead of blood tests [10,11]. These applications offer significant opportunities to improve patients' quality of life by using artificial intelligence techniques such as ML.

Many studies on diabetes diagnosis are carried out to predict the presence of progression of the disease using the obtained data sets. These studies provide important insights into diabetes diagnosis using data analytics methods and ML algorithms.

Mujumdar and Vahidehi [12] conducted a study using various machine learning algorithms for diabetes diagnosis. In this study, a large data set was analysed and predictions were made for the diagnosis of diabetes using patients' clinical and anthropometric data. Researchers have determined that the logistic regression model achieved a high accuracy rate of 96% in diagnosing diabetes.

Kopitar et al. [13] compared regression models, which are commonly used in the prediction of undiagnosed Type 2 diabetes and ML-based prediction models (LightGBM, XGBoost, RF, Glment). Researchers have observed that no clinically meaningful improvement is achieved when more complex prediction models are used. Also, Kumar et al. [14] employed an unsupervised learning approach, the Deep Neural Network (DNN) classifier, for diagnosing type 2 diabetes. Additionally, they utilized a feature importance model packaged with extra trees and random forest for feature selection. Another study aimed to create a prediction model to better identify individuals at risk of diabetes. Predictive models have been developed using ML techniques such as Gradient Boosting Machine (GBM) and Logistic Regression (LR). Additionally in the study, the discriminatory ability of the models was measured by evaluating the area under the Receiver Operating Characteristic Curve (AROC). Researchers have emphasized that GBM and LogR models exhibit superior performance than Random Forest (RF) and Decision Tree (DT) models [15].

Sowah et al. [1] combined multiple artificial intelligence algorithms to address various factors affecting the health status of individuals with diabetes. Soni and Varma [16], various ML algorithms such as SVM, KNN, RF, DT, LR and GBM were used. The results showed that the RF classifier had the highest performance with 77% classification accuracy. Tasin et al. [17] studied machine learning classification models in the training and testing stages. The results revealed that the XGBoost classifier with the ADASYN approach exhibited the highest performance with an accuracy rate of 81%. In other studies, using different machine learning methods on diabetes, researchers have conducted studies on advanced prediction and approaches [18–22].

This study uses various machine learning algorithms by performing data simplification to predict the presence of diabetes and the risk of developing it in healthy individuals in the future, using 100,000 patient data taken from the Kaggle database. The dataset includes patients' clinical and anthropometric data. Data simplification was carried out to reduce the size of the dataset and purify it from unnecessary information. Later, various machine learning algorithms such as LogR, GBM, RF, DT and SVM were used to predict the presence of diabetes and the risk of developing it in the future in healthy individuals. The results obtained show that machine learning models are successful in predicting the presence of diabetes and the risk of developing it in the future in healthy individuals. In particularly, it was determined that the random forest classifier exhibited the highest performance. These findings reveal the potential to improve patients' quality of life by increasing early diagnosis and intervention opportunities for diabetes.

The results highlight the significant opportunities provided by using artificial intelligence techniques on diabetes and provide healthcare professionals with valuable information on the diagnosis of diabetes and its likelihood of progression. Analyses and prediction models show that it can be an effective tool in the management and treatment of diabetic patients.

2. Materials and Method

Algorithms used in fields such as machine learning and artificial intelligence perform predictive modelling processes to forecast future outcomes by utilizing data and statistics. Diabetes commonly presents with abnormal metabolism and elevated blood sugar levels. It can lead to specific complications in different body parts such as the eyes, kidneys, and nervous system. Such symptoms can be employed for data collection purposes, followed by a modelling process based on factors such as age and gender. Data analytics plays a significant role in the management and diagnosis of chronic diseases like diabetes within healthcare. For instance, medical records of past diabetes patients, data on blood sugar levels, treatment methods, and age can be utilized to predict future diabetes risk. By processing these data with machine learning algorithms, personalized risk profiles can be constructed, enabling earlier diagnosis and the development of more effective treatment plans for patients. Therefore, in healthcare issues such as diabetes diagnosis and management, data analytics and predictive modelling serve as powerful tools to enhance patient care and treatment processes. These data play a critical role in the development of algorithms used to achieve improved health outcomes and enhance the quality of life for patients.

2.1. Dataset and data preprocessing

In this study, the diabetes dataset used was obtained from the Kaggle data sharing platform [23]. The dataset contains a total of 100,000 records belonging to individuals. Out of these records, 8,500 represent diabetes patients, while 91,500 do not have diabetes. The dataset consists of a total of 9 features that include symptoms related to diabetes. Eight of these features are used as input parameters and one is the output parameter. The attributes and their values for the dataset are detailed in Table 1.

During the preprocessing stage of the dataset, checks were conducted to identify any missing, scattered or mis-defined data. However, no missing data was detected in the dataset. The necessary transformations to convert the data into a processable form were carried out based on the information provided in Table 1. In the subsequent stage, 20% of the dataset were set aside to evaluate the model's performance. The remaining data points were utilized to train the model. This division of the data allowed for the assessment of the model's accuracy and generalization ability.

Table 1. Attributes and values in the dataset.

No	Attribute	Values
1	Gender	0 – Male 1 – Female 2 - Other
2	Age	[0-80]
3	Hypertension	0-No hypertension 1- There is hypertension
4	Heart disease	0- No heart disease 1- There is heart disease
5	Smoking history	.....
6	Body mass index	[10-95,7]
7	HbA1c level	[3,5-9]
8	Blood glucose level	[80-300]
9	Diabetes	0-No diabetes 1-Have diabetes

The statistics for the features in the dataset are presented in Table 2. This table provides statistical summaries of different attributes in a dataset, including measures such as mean, standard deviation, minimum, and maximum values.

Figure 1 displays the correlation matrix, which is employed to assess the relationship between diabetes and other parameters. Correlation coefficients depict the direction and strength of a linear relationship between two variables. A positive correlation signifies a scenario where both variables increase or decrease concurrently. At the same time, a negative correlation indicates that one variable increases as the other decreases.

2.2. Machine learning algorithms

Machine learning offers various approaches and solution methods for different types of problems. The selection of the most suitable algorithm for each problem type depends on the structure of the dataset and the nature of the intended solution. The algorithms provided in Figure 2 offer different approaches for tasks such as data classification, regression, clustering, dimensionality reduction, and other types of problems.

This study was developed using the Python programming language within the Anaconda platform, specifically in the Spyder environment. Scikit-learn and TensorFlow libraries were utilized for programming. A comprehensive analysis was conducted for the diagnosis of diabetes, involving nine different supervised machine

Table 2. Dataset analysis.

Characteristic	Average	Standard deviation	Minimum Value	Maximum Value
Gender	0.59	0.49	0.00	2.00
Age	41.89	22.52	0.08	80.00
Hypertension	0.07	0.26	0.00	1.00
Heart disease	0.04	0.19	0.00	1.00
Smoking history	1.39	1.59	0.00	5.00
Body mass index	27.32	6.64	10.01	95.69
HbA1c level	5.53	1.07	3.50	9.00
Blood glucose level	138.06	40.71	80.00	300.00
Diabetes	0.09	0.28	0.00	1.00

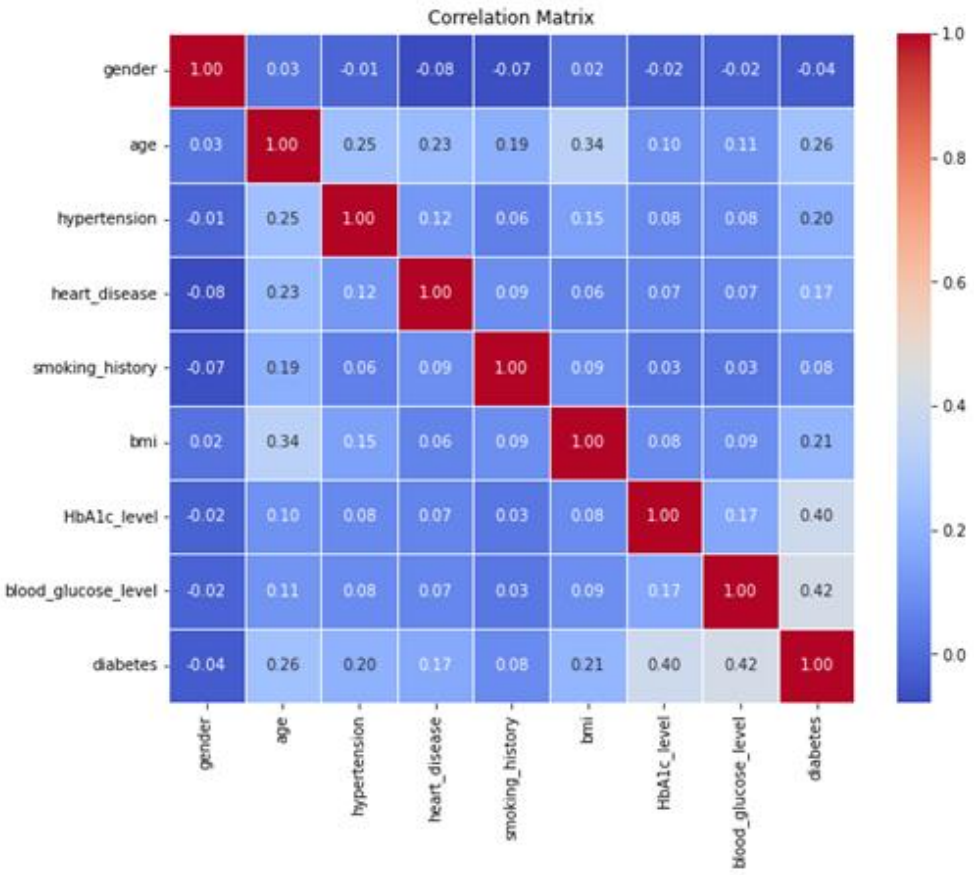

Figure 1. Correlation matrix.

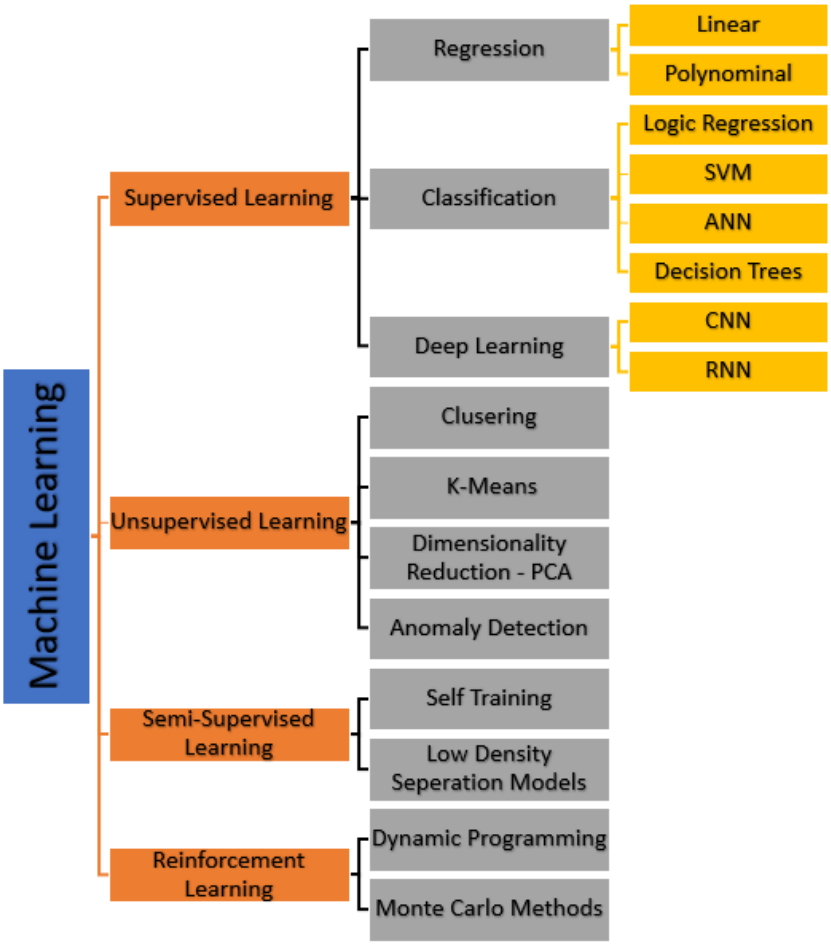

Figure 2. Machine learning algorithms tree.

learning algorithms along with the results of Principal Component Analysis (PCA). The purpose of this analysis was to enhance the accuracy of diabetes diagnosis and determine the best-performing algorithm. The process employed for the development and evaluation of predictive models is illustrated in Figure 3.

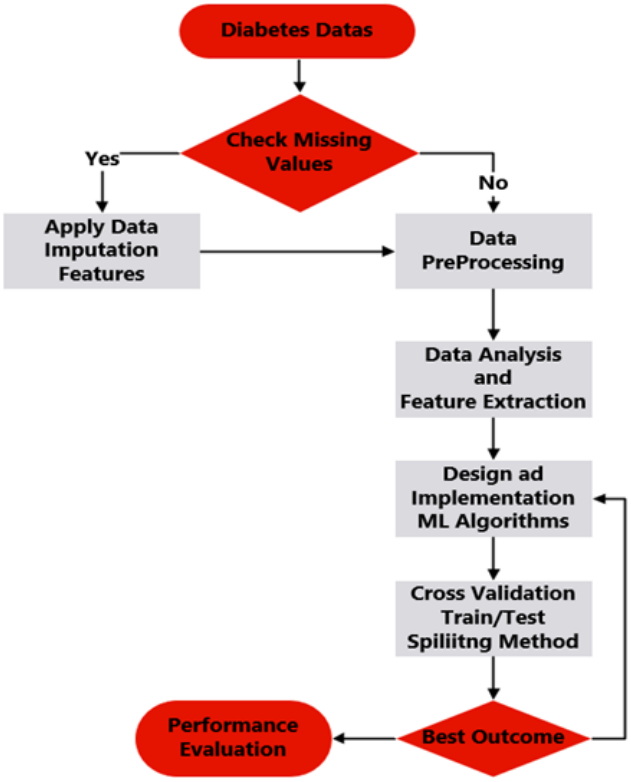

Figure 3. Algorithms process.

The dataset was analysed using various machine learning algorithms, including Linear Regression (LR), Polynomial Regression (PR), Logistic Regression (LogR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), K-Nearest Neighbors (K-NN), Random Forest (RF), and Recurrent Neural Networks (RNN) for diabetes diagnosis. Each algorithm learned from the dataset and made predictions to evaluate its performance in diagnosing diabetes.

PCA (Principal Component Analysis) method was applied to reduce the dimensionality of the dataset, and the performance of algorithms was measured by experimenting with different numbers of components. This helped simplify the dataset, making it more manageable. Changes in the performance of algorithms with different numbers of components from PCA were analysed in detail. The results underwent a comprehensive evaluation to compare the performance of each algorithm and PCA with different numbers of components. This analysis determined which algorithm performed best and provided the most accurate predictions for diabetes diagnosis under different component scenarios.

2.2.1. Linear regression

Linear Regression (LR), predicting the dependent variable when the values of independent variables are given. Using the LR model, data is modelled with the best-fitting linear line [24,25]. LR model named *lr_model* is defined using the *LinearRegression()* function this dataset. The model is trained with training data using the *fit()* function. Then, the trained model makes predictions on the test data, and these predictions are rounded to 0 or 1. The accuracy value is calculated by comparing the rounded predictions with the actual class labels.

2.2.2. Polynomial regression

Polynomial Regression (PR) expresses the relationship between dependent and independent variables with a non-linear, polynomial equation [26]. PR provides more flexibility and can better fit the true structure of the data. Polynomial features were created by increasing the degree of the data in this study. The data is split into training and test data. Then, a PR model is created and fitted to the training data. When training is completed, predictions are made on the test data. The predictions are rounded to the nearest integer, and the accuracy value is calculated by comparing these rounded predictions with the actual test data.

2.2.3. Logistic regression

Logistic Regression (LogR) is a technique applied to classification problems. This method attempts to express the correlation between dependent and independent variables with a curve that is similar to a straight line. This line-like curve is used to predict the categorical values of the dependent variable. Independent variables are measured with binary or multiple values [19]. A LogR model is defined using the *LogisticRegression()* function in this study. The model is trained with the training data. Predictions are made, and the accuracy value is calculated by comparing these predictions with the actual class labels.

2.2.4. Support vector machines

Support Vector Machines (SVMs) are a supervised machine learning algorithm that can be used for classification or regression tasks. SVMs work by finding a hyperplane that separates two classes of data points in a high-dimensional space. The hyperplane is chosen to maximize the margin between the two classes, which helps to ensure that the model is robust to noise and outliers [16]. In this study, an SVM model was trained on a dataset of labeled data. The model was then used to make predictions on a test dataset. The accuracy of the model was evaluated by comparing the predicted labels to the true labels.

2.2.5. Artificial neural networks

Artificial Neural Networks (ANN) are known for their ability to learn complex data structures and relationships. An artificial neural network consists of layers organized in a structured manner [27]. The first layer takes input data, while the last produces predictions or classification results. The hidden layers in between process the data to capture higher-level features. In this study, the model consists of two layers.

In this architecture, data from the input layer is processed in the first layer, and the output is obtained as a probability value between 0 and 1 at the output layer. Training is done for ten epochs (cycles) and in each epoch. Once training is complete, the model makes predictions on the test data. These predictions are then rounded to obtain 0 or 1 values. After making predictions, the accuracy value is calculated by comparing them with the actual class labels.

2.2.6. Decision trees

Decision Trees (DT) start with a root node initially. This root node splits the data into child nodes under specific conditions. These child nodes, in turn, further split into sub-nodes based on conditions and decisions. This process continues until it reaches leaf nodes, where the data is classified or predicted [27,28] . Decision Trees are highly effective in understanding patterns and relationships in a dataset. Each node has the ability to classify or make predictions by evaluating data features. In this study, the model is trained with the training data. After making predictions, the accuracy value is calculated by comparing these predictions with the actual class labels.

2.2.7. K-Nearest neighbors

K-Nearest Neighbors (K-NN) algorithm's fundamental principle is to determine the class or value of a data point by using the majority of its closest neighbors. Data points are positioned in space, and when a new point arrives, its closest neighbors are examined. The K-NN algorithm makes predictions using the labels or values of these neighbors [16]. The model is trained with the training data. After making predictions, the accuracy value is calculated by comparing these predictions with the actual class labels.

2.2.8. Random forest

Random Forest (RF) is a powerful machine learning algorithm used for classification and regression problems. This algorithm is constructed by combining multiple decision trees for accurate predictions and classifications. It achieves this through a voting process where each tree votes based on its own analysis and the final result reflects the consensus of the majority [17,29]. Each decision tree is trained with different subsets of data and random feature selection. This allows each tree to learn from different features and data points. The RF algorithm aims to obtain more reliable and stable results by combining the predictions of these trees. In this study, a Random Forest model is defined using *RandomForestClassifier()*. The model is trained with the training data. After making predictions, the accuracy value is calculated by comparing these predictions with the actual class labels.

2.2.9. Recurrent neural network

Recurrent Neural Network (RNN) is a type of artificial neural network designed to process data such as time series and sequential data. RNNs have the ability to internally store information from previous time steps, making them suitable for the analysis of sequential data. In this study, an RNN model was employed to process time series data [30]. The model is trained with training data. During training, the model utilizes information from previous time steps to understand and learn the temporal dependencies within the data. Once training is completed, the model makes predictions on test data. These predictions are then compared to the actual class labels, and an accuracy value is calculated. This process is employed to assess how the model processes and predicts time series data.

2.2.10. Principal component analysis

Principal Component Analysis (PCA) is a statistical technique employed for dimensionality reduction in multi-dimensional datasets, aiming to encapsulate the core information within the data [31]. Its fundamental aim is to express the data with a reduced number of eigenvalue components, simplifying the dataset. The main idea of PCA is to create new transformed principal components. It aims maximize the variance in the dataset. These new components are linear combinations of the original variables. PCA achieves this by selecting the components that carry the highest variance, effectively filtering out unnecessary or low-variance variables, thereby expressing the dataset in a lower dimension for improved interpretation and analysis. PCA finds applications in data compression, visualization, noise reduction, and enhancing the performance of certain models. In this study, the impact of PCA is investigated by selecting different numbers of components (ranging from 8 to 1). Consequently, the dimensionality of the dataset is sequentially reduced from 8 to 1, and the outcomes of this dimension reduction process are evaluated. The experiments conducted with various numbers of components aim to provide a detailed analysis of PCA's effect on the dataset.

3. Results

The study evaluated nine supervised ML algorithms. Evaluating their accuracy, precision, recall and F1 scores. According to Table 3, most models demonstrated notably high levels of these metrics when applied to classify the dataset. These results indicate that the machine learning algorithms used successfully classified symptoms related to diabetes in the dataset.

The high values of F1 score, recall, accuracy and precision indicate that the models were effective in identifying and classifying cases related to diabetes. This suggests that the selected algorithms, such as LogR, PR, SVM, ANN, DT, KNN, RF and RNN all showed promising performance in diabetes diagnosis.

Furthermore, it's important to note that these algorithms were evaluated using various performance metrics to ensure a comprehensive assessment of their effectiveness in diabetes diagnosis. The combination of these metrics provides a well-rounded view of each algorithm's performance, highlighting their strengths and suitability for different aspects of diabetes diagnosis.

Table 3. Performance evaluation of machine learning algorithms.

ML Algorithms	Accuracy	Precision	Recall	F1 Score
LR	0,938	0,941	0,938	0,922
PR	0,952	0,955	0,952	0,944
LogR	0,952	0,948	0,952	0,947
SVM	0,947	0,950	0,947	0,936
ANN	0,959	0,959	0,959	0,955
DT	0,951	0,952	0,951	0,952
KNN	0,952	0,950	0,952	0,947
RF	0,969	0,969	0,969	0,968
RNN	0,963	0,963	0,963	0,959

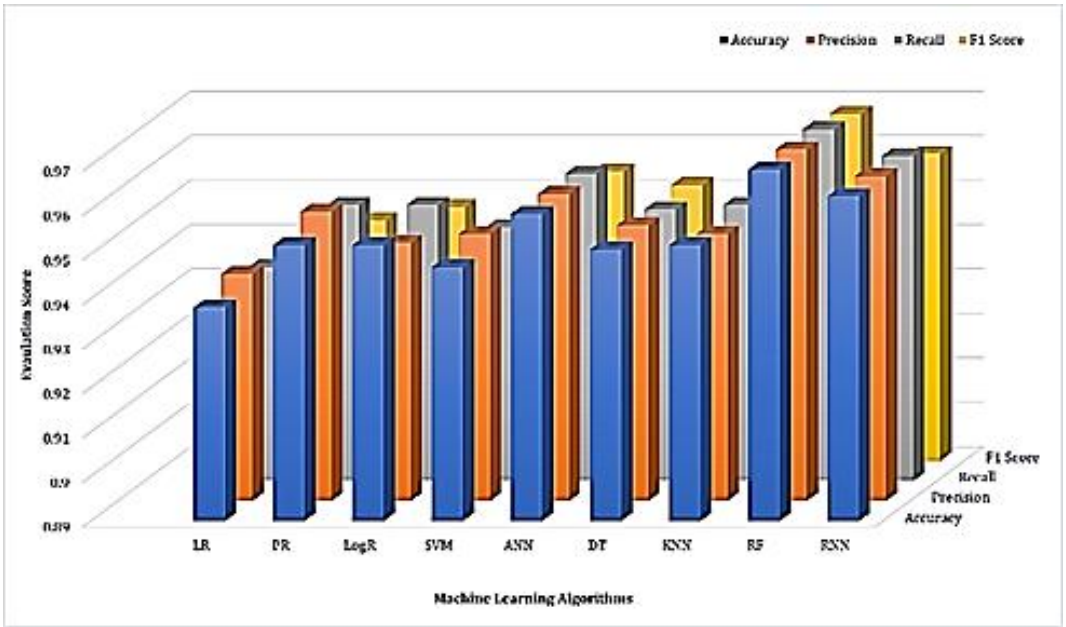

Figure 4. Machine learning algorithms performance.

Figure 4 shows the performance analysis of machine learning algorithms.

In conclusion, the study's findings demonstrate that the application of supervised machine learning algorithms, along with the evaluation of multiple performance metrics can significantly contribute to accurate diabetes diagnosis. This has the potential to improve understanding in the field and enhance access to correct diagnosis and treatment for patients.

Especially the RF model has demonstrated the highest performance in all metrics. This means that this model achieved high values not only in accuracy but also in precision, recall, and F1 score, indicating both a high rate of correct classification and a high rate of correctly classified positive cases. Additionally, it outperformed in terms of the rate of detecting all true positives (recall) and F1 score. This indicates that the RF algorithm can be effectively used in critical medical conditions such as diabetes diagnosis. However, the performance of other algorithms is also notably high. For example, ANN stands out with its high accuracy and classification success. Similarly, DT and KNN algorithms have also yielded good results. These results emphasize the effectiveness of various machine learning algorithms in diagnosing diabetes. The choice of the most suitable algorithm may depend on specific requirements, such as the balance between precision and recall or computational efficiency. Therefore, this study provides valuable insights into the selection and application of machine learning algorithms in the field of medical diagnosis, particularly for diabetes.

Table 4 displays the test accuracies achieved by each model on the PCA-transformed dataset for different numbers of components. Generally, PCA is used with the aim of dimensionality reduction, intending to represent the dataset in a lower dimension. This may lead to some information loss, but it can reduce noise in the dataset and potentially improve the performance of some models. As seen in Table 4, as the number of components decreases, the accuracy value generally tends to decrease. This indicates that fewer components result in more significant information loss. However, in some cases, reducing very high-dimensional datasets to smaller dimensions can accelerate model training and help with better generalization. Therefore, the choice of the number of components to use should be made carefully, depending on the specific application and dataset.

The accuracy rate of machine learning algorithms given in Table 4 increases as the number of components increases, as shown in the graph in Figure 5. When working with the real dataset with 8 components, the accuracy values on the PCA-transformed dataset are generally low or similar. However, the PCA transformation for the RNN model produces results that are almost the same as the real dataset's accuracy. When dimensionality reduction is performed with PCA for component numbers 7, 6, 5, 4, and 3, the accuracy values increase or remain almost the same for most models. In this case, it has been observed that performance can be improved by reducing the complexity of the model and

representing the dataset in a lower dimension. When reducing the number of components to 2 or lower levels, the performance of the models significantly decreases. This leads to the loss of important features of the dataset and makes it difficult for models to learn the data effectively.

In conclusion, when PCA transformation is used by carefully selecting the right number of components, it can improve the performance of some models and reduce computation time by representing the dataset in a lower dimension. However, the choice of the correct number of components should be made carefully, depending on the structure of the dataset and model performance.

Table 4. Effect of component number on accuracy.

Number of Components	LR	PR	LogR	SVM	ANN	DT	KNN	RF	RNN
Real Data (n=8)	0,938	0,952	0,952	0,947	0,953	0,951	0,952	0,970	0,966
PCA (n=7)	0,937	0,952	0,958	0,960	0,967	0,953	0,952	0,969	0,971
PCA (n=6)	0,936	0,952	0,957	0,960	0,968	0,953	0,952	0,96	0,970
PCA (n=5)	0,936	0,952	0,958	0,960	0,967	0,953	0,952	0,968	0,968
PCA (n=4)	0,929	0,952	0,939	0,948	0,945	0,923	0,941	0,944	0,947
PCA (n=3)	0,929	0,952	0,939	0,948	0,947	0,921	0,941	0,941	0,947
PCA (n=2)	0,928	0,952	0,938	0,948	0,947	0,923	0,943	0,938	0,948
PCA (n=1)	0,928	0,952	0,940	0,947	0,947	0,912	0,943	0,912	0,944

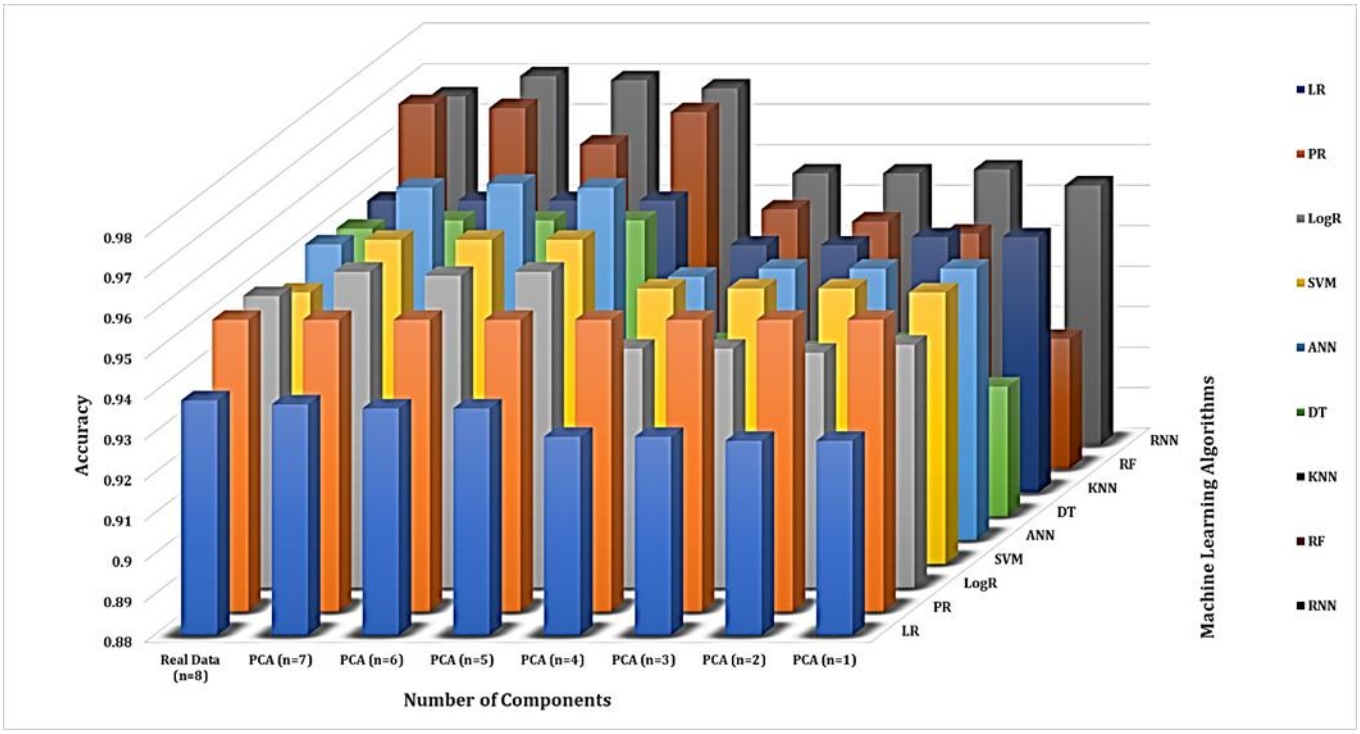

Figure 5. Accuracy of machine learning algorithms according to PCA

4. Conclusion

In this study, data mining approaches were used to evaluate the performance of different machine learning algorithms for diabetes diagnosis. Additionally, the impact of reducing the dataset's dimensionality using the PCA method was examined. The performance of diabetes diagnosis was assessed with 9 different supervised machine learning algorithms applied to the dataset. The analyses revealed that the RF algorithm had the highest accuracy, precision, recall, and F1 score values compared to other algorithms. This result indicates that the RF algorithm can be successfully used in critical medical situations like diabetes diagnosis. Furthermore, ANN, DT, and KNN algorithms also demonstrated good performance. However, the dataset's dimensionality was reduced using the PCA method, and the performance of algorithms was measured with different numbers of components. It was shown that when the PCA transformation is used by selecting the correct number of components, it can enhance the performance of some models and reduce computation time by representing the dataset in a lower dimension. Nevertheless, the choice of the right number of components should be made carefully, considering the dataset's structure and model performance.

Among the limitations of the study are the limited number of features in the dataset and its lack of representation of different types of diabetes. Additionally, the dataset's imbalance should be taken into account, as it may have some impact on classification performance. Future studies are recommended to use larger and balanced datasets and include different diabetes types. Moreover, exploring combinations of different machine learning algorithms and more advanced feature selection methods could further enhance performance.

Author contributions

Yavuz Bahadır Koca: Visualization, Conceptualization, Software, Writing-Reviewing and Editing, Validation.
Elif Aktepe: Methodology, Data curation, Writing-Original draft preparation, Software.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Sowah, R. A., Bampoe-Addo, A. A., Armoo, S. K., Saalia, F. K., Gatsi, F., & Sarkodie-Mensah, B. (2020). Design and development of diabetes management system using machine learning. International Journal of Telemedicine and Applications, 2020(1), 8870141. https://doi.org/10.1155/2020/8870141
2. Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International Informatics and Software Engineering Conference (UBMYK), 1-4. https://doi.org/10.1109/UBMYK48245.2019.8965556
3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005
4. Abudnejad, N., Salehpour, M., & Saadati, Z. (2023). Theoretical evaluation of boron carbide nanotubes as non-enzymatic glucose sensors. Chemical Physics Letters, 823, 140510. https://doi.org/10.1016/j.cplett.2023.140510
5. Başer, B. Ö., Yangın, M., & Sarıdaş, E. S. (2021). Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112-120. https://doi.org/10.19113/sdufenbed.842460
6. World Health Organization WHO European Regional Obesity Report (2022). World Health Organization. Regional Office for Europe. ISBN 9289057734.
7. Sun, J., Ren, J., Hu, X., Hou, Y., & Yang, Y. (2021). Therapeutic effects of Chinese herbal medicines and their extracts on diabetes. Biomedicine & Pharmacotherapy, 142, 111977. https://doi.org/10.1016/j.biopha.2021.111977
8. Hasanzad, M., Aghaei Meybodi, H. R., Sarhangi, N., & Larijani, B. (2022). Artificial intelligence perspective in the future of endocrine diseases. Journal of Diabetes & Metabolic Disorders, 21(1), 971-978. https://doi.org/10.1007/s40200-021-00949-2
9. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005
10. Cappon, G., Acciaroli, G., Vettoretti, M., Facchinetti, A., & Sparacino, G. (2017). Wearable continuous glucose monitoring sensors: a revolution in diabetes treatment. Electronics, 6(3), 65. https://doi.org/10.3390/electronics6030065
11. Zherebtsov, E. A., Zharkikh, E. V., Kozlov, I. O., Loktionova, Y. I., Zherebtsova, A. I., Rafailov, I. E., ... & Rafailov, E. U. (2019, June). Wearable sensor system for multipoint measurements of blood perfusion: pilot studies in patients with diabetes mellitus. In European Conference on Biomedical Optics, 11079_62. https://doi.org/10.1117/12.2526966
12. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299. https://doi.org/10.1016/j.procs.2020.01.047
13. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10(1), 11981. https://doi.org/10.1038/s41598-020-68771-z
14. Nadesh, R. K., & Arivuselvan, K. (2020). Type 2: diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering, 1, 55-61. https://doi.org/10.1016/j.ijcce.2020.10.002
15. Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19, 1-9. https://doi.org/10.1186/s12902-019-0436-6
16. Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(09), 921-925.
17. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1-2), 1-10. https://doi.org/10.1049/htl2.12039
18. Cahn, A., Shoshan, A., Sagiv, T., Yesharim, R., Goshen, R., Shalev, V., & Raz, I. (2020). Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. Diabetes/metabolism Research and Reviews, 36(2), e3252. https://doi.org/10.1002/dmrr.3252
19. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19(1), 1-15.

https://doi.org/10.1186/s12911-019-0918-5
20. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531. https://doi.org/10.1109/ACCESS.2020.2989857
21. Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 18(1/2), 90-100. https://doi.org/10.1016/j.aci.2018.12.004
22. Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1, 1-8. https://doi.org/10.1007/s42452-019-1117-9
23. Nandy, S. (2023). Kaggle. https://www.kaggle.com/datasets/sharmisthanandy/diabetes
24. Abdelaziz, A., Elhoseny, M., Salama, A. S., & Riad, A. M. (2018). A machine learning model for improving healthcare services on cloud computing environment. Measurement, 119, 117-128. https://doi.org/10.1016/j.measurement.2018.01.022
25. Schober, P., & Vetter, T. R. (2021). Statistical Minute Logistic Regression in Medical Research.
26. Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(2), 140-147. https://doi.org/10.38094/jastt1457
27. Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In Recent Developments in Machine Learning and Data Analytics: IC3 2018, 740, 67-78. https://doi.org/10.1007/978-981-13-1280-9_6
28. Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 122-127. https://doi.org/10.1109/RAICS.2015.7488400
29. Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. Procedia Computer Science, 192, 467-477. https://doi.org/10.1016/j.procs.2021.08.048
30. Al-Askar, H., Radi, N., & MacDermott, Á. (2016). Recurrent neural networks in medical data analysis and classifications. In Applied Computing in Medicine and Health, 147-165. https://doi.org/10.1016/B978-0-12-803468-2.00007-2
31. Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). Computers & Geosciences, 19(3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R