applied sciences

MDPI

Article

Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study

Luís Chaves and Gonçalo Marques *

Polytechnic of Coimbra, ESTGOH, Rua General Santos Costa, 3400-124 Oliveira do Hospital, Portugal; aa3luischaves@gmail.com

* Correspondence: goncalosantosmarques@gmail.com

Abstract: Diabetes is a life-long condition that is well-known in the 21st century. Once known as a disease of the West, the rise of diabetes has been fed by a nutrition shift, rapid urbanization and increasingly sedentary lifestyles. In late 2019, a new public health concern was emerging (COVID-19), with a particular hazard concerning people living with diabetes. Medical institutes have been collecting data for years. We expect to achieve predictions for pathological complications, which hopefully will prevent the loss of lives and improve the quality of life using data mining processes. This work proposes a comparative study of data mining techniques for early diagnosis of diabetes. We use a publicly accessible data set containing 520 instances, each with 17 attributes. Naive Bayes, Neural Network, AdaBoost, k-Nearest Neighbors, Random Forest and Support Vector Machine methods have been tested. The results suggest that Neural Networks should be used for diabetes prediction. The proposed model presents an AUC of 98.3% and 98.1% accuracy, an F1-Score, Precision and Sensitivity of 98.4% and a Specificity of 97.5%.

Keywords: diabetes; COVID-19; SARS-CoV-2; data mining; machine learning

Citation: Chaves, L.; Marques, G. Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. *Appl. Sci.* **2021**, *11*, 2218. https://doi.org/10.3390/app11052218

Received: 30 January 2021
Accepted: 26 February 2021
Published: 3 March 2021

1. Introduction

Diabetes mellitus is a chronic disease caused when the pancreas does not produce enough insulin, or when the body cannot use effectively the insulin it produces. Insulin is a hormone that moves the glucose from the bloodstream to the body cells where it will be used as energy. If not consumed by the cells, the excess sugar in the blood can lead to serious health problems [1]. Diabetes and its complications have a significant economic impact on individuals and their families, health systems and national economies. In 2019, the global expenditures for diabetes treatment were estimated at 760 billion U.S Dollars and are expected to rise to 845 billion U.S. Dollars by 2045 [2]. Diabetes can have life-threatening consequences for the cardiovascular, renal, and nervous system if it is not treated. In 2019, was estimated that 463 million people have diabetes worldwide. Moreover, it is predicted the prevalence to be 578 million and 700 million by 2030 and 2045, respectively [3].

Severe acute respiratory syndrome SARS-CoV-2 is the virus responsible for the coronavirus disease 2019 (COVID-19). On 30 January 2020, the World Health Organization announced the outbreak as a Public Health Emergency of International Concern and later on 11 March 2020, a pandemic crisis. The virus was first reported in Wuhan, China, in December 2019 and quickly spread worldwide. As of 3 January 2021, 84,985,054 confirmed cases along with 1,841,077 deaths had been reported by the Center for Systems Science and Engineering at Johns Hopkins University. In general, people with diabetes are more likely to have severe symptoms and complications when infected with any virus, combined with other chronic conditions such as heart disease, which increases their risk of getting those severe complications if infected with COVID-19 [4]. Recent studies have shown that "Elevated glucose levels increase SARS-CoV-2 replication, glycolysis sustains SARS-CoV-2 replication via the production of mitochondrial reactive oxygen species and activation of hypoxia-inducible factor 1α". Patients with diabetes mellitus typically fall into higher

categories of SARS-CoV-2 infection severity than those without, and poor glycaemic control predicts an increased need for medications, hospitalizations, and increased mortality [5].

Data Mining methods search for patterns and trends in large-scale data by using advanced mathematical algorithms to partition the data and evaluate future events' probability. Data Mining includes several disciplines such as statistics, probability, machine learning and artificial intelligence [6]. Digital databases combined with the ability to apply computationally intensive statistical methodology to these data, powered by fast computers, have increased the number of applications of Data Mining in several domains. The Healthcare industry is constantly generating and storing new data [7]. The effective use of this data can assist professionals in providing a fast and accurate diagnostic [8]. One in two people live with diabetes without being aware of it [3]. Their condition is not being monitored and kept under control. Consequently, they are more likely to experience additional complexities if infected by the coronavirus.

This study aims to analyse how different classification algorithms behave when applied to a training data set. The application of Data Mining can save lives in the future by providing a tool for an early diagnosis of diabetes. The experiment was conducted in Orange Data Mining, a machine learning and data mining suite for data analysis through Python scripting. On the other hand, the main contribution is to present the results of six different machine learning methods for early diagnosis of diabetes. Moreover, the results recommend the use of Neural Networks for early diagnosis of diabetes. Finally, all the model configuration details used are described, and the results are compared with similar research activities in this field. The authors use a public data set to test the proposed methods. The main reasons to use a public data set is to overpass the current challenges of data collection regarding the General Data Protection Regulation (GDPR) applied in Europe.

2. Materials and Methods

This study uses six different classifications algorithms, namely Naive Bayes, Neural Network, AdaBoost, k-nearest neighbors (kNN), Random Forest, and Support Vector Machine (SVM). Other studies present reliable results with this methodology [9–14]. Although several optimization methods are available in the literature, the authors do not have used any optimization method. Since the study's objective is to use low complexity methods which do not require high computational resources. The experimental setup is presented in Figure 1. The proposed methods' results are compared with the related work. This experiment was run using Orange v3.27.1 on a machine equipped with an Intel® Core™ i9-9880H @ 2.3–4.8 GHz, 16384 MB DDR4-2666 RAM, and an AMD Radeon Pro 5500 M–4096 MB GDDR6. All tests were performed using a 10-fold Cross-Validation technique to split the training and testing data set. Orange is an open-source project developed by Bioinformatics Lab at the University of Ljubljana, Slovenia, in collaboration with the open-source community. Data Mining algorithms are implemented through visual programming by use of widgets or Python scripting [15]. Its simplicity and ease of use are some of the features that make Orange so popular, especially in an educational environment. The implemented Orange workflow is presented as Supplementary File 1.

2.1. Naive Bayes

Based on Bayes' Theorem, the Naive Bayes is a probabilistic classifier used for classification problems. This algorithm assumes that predictors/features are independent, meaning that particular features do not affect the other. Consequently, it is called naive [16].

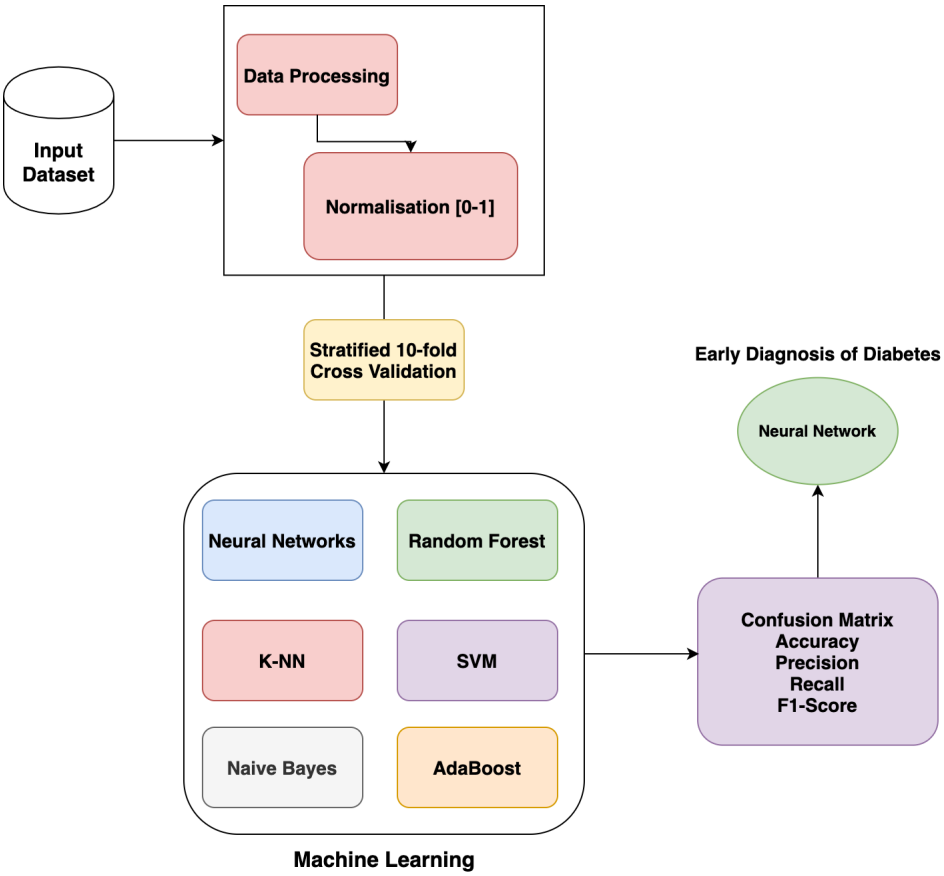

Figure 1. Experimental Setup.

2.2. Neural Networks

Inspired by the way biological nervous systems process information, Artificial Neural Networks (ANNs) are composed of interconnected elements named neurons, processing and cooperating to determine solutions to specific problems. Similar to humans, the learning process of ANNs is based on examples. Instead of a set of instructions for the accomplishment of a specific task, they are given examples to analyse and find a way to solve the problem [17]. The proposed Neural Network model parameters used are described. The number of neurons in hidden layers is 31. The activation function used is ReLu, the Solver is L-BFGS-B, and the alpha is 0.001. Finally, the maximum number of iterations is 30, and we use replicable training.

2.3. AdaBoost

Introduced in 1995 by Freund and Schapire, the Adaptive Boosting or AdaBoost is a machine learning meta-algorithm. It can be used both for classification and regression problems. It starts by predicting the original data set and gives equal weight to each observation from sequence learners on different weighted training data. If the prediction is erroneous, it gives a higher weight to the observation that has been predicted incorrectly. It then continues to learn until a limit is reached in the number of models or accuracy [18]. The proposed AdaBoost model uses a tree base estimator, and the number of estimators is 50. The learning rate is 1, and the classification algorithm implemented is SAMME.R. The regression loss function used is linear.

2.4. kNN

kNN is a supervised machine learning algorithm that can resolve both classification and regression tasks. It is one of the most fundamental and simple classification methods with low execution and computing time [19]. The process assumes that similar things exist

close to one another; the closer two samples are, the more likely they are to relate to the same category. First, the k parameter is determined, and this is the number of neighbours for a given point. Then, through distance functions, it calculates the distance of the new data that will be included in the sample data set. It is assigned to the class of k neighbors according to the attribute values. Finally, the data is labelled [20]. The proposed method uses three neighbours, the metric is Euclidean, and the weight is the distance.

2.5. Random Forest

Random Forest is a supervised learning algorithm, described as a combination of tree predictors. It is used both for classification and regression [21]. It is considered to be one of the most accurate general-purpose learning technique. The Random Forest is easy to implement and can manage a large number of input variables without over-fitting [22]. Random Forest adds additional randomness to the model while growing the trees. Instead of searching for the most critical feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model [23]. The proposed model uses 10 trees and do not split subsets smaller than 5.

2.6. SVM

Originated from statistical learning theory, SVM is a supervised machine learning model that uses classification algorithms for bi-category classification problems [24]. It is based on a linear division; however, not all data can be linear division. The two categories points may require a curve to divide their borders in the two-dimensional space. To linearly separate, a low-dimensional space point to the high-dimensional space is mapped over SVM and then use the principles of linear division to determine the border's classification. It is a linear division in the high-dimensional space, while in the original data space, it is a non-linear division. In a hyperplane, the objective of SVM is to find n number of features which distinctly classifies the data points. It then finds a space with the maximum margin, which means the maximum distance between both classes' data points. Maximizing the margin distance provides reinforcement. This results in an improvement of classification accuracy [25]. The proposed SVM model is defined with a cost of 1, and the ϵ is 0.1. The Kernel is Polynomial, and the g is defined as auto. The numerical tolerance is 0.001, and the iteration limit is 100.

2.7. Data set Analysis and Preprocessing

The data set contains information about newly diabetic or in the process of being diagnosed. A total of 520 instances classified by 16 attributes diabetes-related used as features and 1 class attribute that specifies the subject diagnoses as positive or negative. The data was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh [9]. The data is publicly available on the UCI Machine Learning Repository. The patients' ages range from 16 to 90 years old. Table 1 displays the data set attributes and their possible values. Data can have many irrelevant and missing parts which make the elimination of noisy instances necessary. To achieving better-quality analysis results, preprocessing is a crucial step. Consequently, the authors have applied normalization to the interval [0–1] since this is proposed as an effective method by previous studies [26,27]. Data imputation methods are not used since this data set does not have missing values.

Table 1. Dataset attributes.

Attribute Name	Answer		Type
Age	16–90		feature
Gender	1. Male	2. Female	feature
Polyuria	1. Yes	2. No	feature
Polydipsia	1. Yes	2. No	feature
Sudden weight loss	1. Yes	2. No	feature
Weakness	1. Yes	2. No	feature
Polyphagia	1. Yes	2. No	feature
Genital thrush	1. Yes	2. No	feature
Visual blurring	1. Yes	2. No	feature
Itching	1. Yes	2. No	feature
Irritability	1. Yes	2. No	feature
Delayed healing	1. Yes	2. No	feature
Partial paresis	1. Yes	2. No	feature
Muscle stiffness	1. Yes	2. No	feature
Alopecia	1. Yes	2. No	feature
Obesity	1. Yes	2. No	feature
Class	1. Positive	2. Negative	target

The information gain is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute. Gain ratio tries to overcome this bias by adjusting the information gain to each attribute, allowing for consistency of the attribute values [28]. In this study, we used all 16 features. Table 2 presents the feature ranking according to Information Ratio. The three most relevant features for this dataset are Polyuria, Polydipsia and Gender.

Table 2. Ranked Attributes.

Rank	Feature	Information Ratio
1	Polyuria	0.3623
2	Polydipsia	0.3619
3	Gender	0.1720
4	Sudden weight loss	0.1518
5	Partial paresis	0.1467
6	Polyphagia	0.0884
7	Irritability	0.0912
8	Alopecia	0.0551
9	Visual blurring	0.0470
10	Weakness	0.0436
11	Age	0.0113
12	Muscle stiffness	0.0115
13	Genital thrush	0.0118
14	Obesity	0.0059
15	Delayed healing	0.0016
16	Itching	0.0001

2.8. Performance Metrics

Performance metrics tell us how a data mining algorithm is performing on a given data set. Consequently, we can compare the results of different algorithms and decide which one performs better or worse. Therefore, Area Under the Curve (AUC), classification accuracy (CA), Precision, Recall/Sensitivity, Specificity and F1-Score have been used. These performance metrics are selected since they are used by most of the related work [9–14].

One of the essential metrics for the evaluation of any classification model's performance is ROC, which is a probability curve. AUC represents the degree or measure of separability. It tells the capacity of the model to distinguish between classes. The higher the AUC, the better the p model's performance at predicting the true positives and the true negatives [29].

Accuracy (CA) refers to the correct predictions rate. It is given by the division of total correct predictions by the total number of instances.

$$CA = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is also called positive predictive value and reports which of those who were predicted to be positive are actually positive. It is defined as the number of true positives divided by the number of true positives plus the false positives.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity calculates the true positive rate, and this is how many of the actual positives were correctly labelled. It is defined as the number of true positives divided by the number of true positives plus the number of false negatives.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity defines the true negative rate. This is the proportion of actual negatives which were correctly predicted. To obtain this metric, we divide the number of false positives by the number of true negatives plus the number of false positives.

$$Specificity = \frac{TN}{TN + FP}$$

F1-Score is the harmonic mean of Precision and Recall. It presents a better measure of the incorrectly classified cases than the CA.

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity}$$

2.9. K-Fold-Cross-Validation

Cross-Validation is a data re-sampling procedure used to evaluate machine learning models on a limited data sample. Divided into K partitions, or folds, the data set is then iterated over each fold, using it to test the model and the remaining k-1 portions for training. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. Using this method, the whole data set is used for both training and validation, which will produce a more accurate result [30]. This method has been selected since it was recommended in several previous studies [31,32].

3. Results

In this section, we present the experimental results. The authors have done six experiments using different machine learning methods. These include Naive Bayes, Neural Network, AdaBoost, kNN, Random Forest, and SVM. The confusion matrix of each experiment is presented to allow the readers to calculate other performance metrics, if necessary.

In the first experiment, we applied the Naive Bayes classifier, which correctly predicted 452 instances out of 520, a success rate of 86.92%. Table 3 presents 178 true negatives against 46 false negatives and 274 true positives against 22 false positives.

Table 3. Confusion Matrix of Naive Bayes Classification.

		Prediction		
		Negative	**Positive**	**Σ**
Actual	**Negative**	178	22	200
	Positive	46	274	320
	Σ	224	296	520

In the second experiment, we applied the Neural Network classifier, which correctly predicted 510 instances out of 520, a success rate of 98.08%. Table 4 presents 195 true negatives against 5 false negatives and 315 true positives against 5 false positives.

Table 4. Confusion Matrix of Neural Network Classification.

		Prediction		
		Negative	**Positive**	**Σ**
Actual	**Negative**	195	5	200
	Positive	5	315	320
	Σ	200	320	520

In the third experiment, we applied the AdaBoost classifier, which correctly predicted 506 instances out of 520, a success rate of 97.31%. Table 5 presents 194 true negatives against 8 false negatives and 312 true positives against 6 false positives.

Table 5. Confusion Matrix of AdaBoost Classification.

		Prediction		
		Negative	**Positive**	Σ
Actual	**Negative**	194	6	200
	Positive	8	312	320
	Σ	202	318	520

In the fourth experiment, we applied the kNN classifier, which correctly predicted 506 instances out of 520, a success rate of 97.31%. Table 6 presents 199 true negatives against 13 false negatives and 307 true positives against 1 false positive.

Table 6. Confusion Matrix of kNN Classification.

		Prediction		
		Negative	**Positive**	**Σ**
Actual	**Negative**	199	1	200
	Positive	13	307	320
	Σ	212	308	520

In the fifth experiment, we applied the Random Forest classifier, which correctly predicted 504 instances out of 520, a success rate of 96.92%. Table 7 presents 194 true negatives against 10 false negatives and 310 true positives against 6 false positives.

Table 7. Confusion Matrix of Random Forest Classification.

		Prediction		
		Negative	**Positive**	**Σ**
Actual	**Negative**	194	6	200
	Positive	10	310	320
	Σ	196	324	520

In the last experiment, we applied the SVM classifier, which correctly predicted 505 instances out of 520, a success rate of 97.12%. Table 8 presents 192 true negatives against 7 false negatives and 313 true positives against 8 false positives.

Table 8. Confusion Matrix of Support Vector Machine (SVM) Classification.

		Prediction		
		Negative	**Positive**	**Σ**
Actual	**Negative**	192	8	200
	Positive	7	313	320
	Σ	199	321	520

According to obtained results in Tables 9 and 10, we can state that Neural Networks presents the best classification accuracy of 98.1%. At the same time, Naive Bayes scored the lowest accuracy with 86.9%. Moreover, the F1-Score of the proposed Neural Networks is 98.4%.

Table 9. Results of Classification Algorithms.

Classifier	Correctly Classified	Misclassified
Naive Bayes	86.92%	13.08%
Neural Network	98.08%	1.92%
AdaBoost	97.31%	2.69%
kNN	97.31%	2.69%
Random Forest	96.92%	3.08%
SVM	97.12%	2.88%

Table 10. Results of Evaluation Measures for Area Under the Curve (AUC), classification accuracy (CA), F1-Score, Precision, Sensitivity, and Specificity.

Classifier	AUC	CA	F1-Score	Precision	Sensitivity	Specificity
Naive Bayes	0.946	0.869	0.890	0.926	0.856	0.890
Neural Network	0.983	0.981	0.984	0.984	0.984	0.975
AdaBoost	0.973	0.973	0.978	0.981	0.975	0.970
kNN	0.988	0.973	0.978	0.997	0.959	0.995
Random Forest	0.997	0.969	0.975	0.981	0.969	0.970
SVM	0.993	0.971	0.977	0.975	0.978	0.960

4. Discussion

When dealing with diseases such as diabetes it is essential to provide an early and accurate diagnosis. A delayed diagnosis of diabetes can lead to severe health consequences if vigilance is not applied. Therefore, when applying data mining techniques to predict someone's condition, they must be as highly accurate as possible. The cost of a false

negative is enormously higher than a false positive. If wrongly diagnosed, the subject might take a relaxed posture without knowing their real condition, and this may lead to severe health concerns.

Several similar works are available in the literature. In this section, the results of this study are compared with the state of art. In 2019, M. M. Faniqul Islam et al. analyzed a diabetes data set, implementing a Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm with 10-fold Cross-Validation. Their results exposed the highest accuracy of 97.40% using Random Forest classifier [9]. In 2020, K. Alpan and G. S. İlgi presented a study comparing data mining classification techniques for diabetes using WEKA Tool. In total, seven algorithms such as Bayes Network, Naive Bayes, Decision tree (J48), Random tree, Random forest, kNN and SVM have been used. The authors registered a 98.07% accuracy, being kNN the classifier with the best performance using 10-fold Cross-Validation [10]. H. Naz and S. Ahuja applied data mining techniques on the PIMA diabetes data set. The authors compared Deep Learning, ANN, Naive Bayes and Decision Tree. According to results, Deep Learning provided the best performance with 98.07% accuracy [11]. A shuffled sampling with 80/20% ratio into the training and validation set has been used. In 2020, N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia analysed a diabetes data set proposing a model based on ANNs. Their study proved the highest accuracy of 88% efficacy [12]. In 2018, M. Peker, O. Özkaraca, and A. Sasar conducted a study using Orange Tool to analyse a diabetes data set. Authors implemented and compared Random Forest, Feed-Forward Artificial Neural Networks, kNN, SVM and Decision Tree. The results revealed ANNs to be the algorithm with the highest accuracy. It correctly predicted 93.85% of cases [13]. Table 11 presents the comparative results between studies.

Table 11. Studies Results Comparison.

Study	N.B. [1]	ANN	kNN	R.F. [2]	SVM	AdaBoost
[9]	**87.4%**	-	-	97.4%	-	-
[10]	87.11%	-	98.07%	97.5%	92.11%	-
[11]	76.33%	90.34%	-	-	-	-
[12]	-	88%	-	-	-	-
[13]	-	93.85%	90.85%	91.34%	78.51%	-
[14]	79%	80.47%	**98.62%**	**98.8**%	75.29%	77.16%
This study	86.92%	**98.08%**	97.31%	96.9%	**97.12**%	**97.3**%

[1] Naive Bayes; [2] Random Forest. The bold numbers represent highest value.

The proposed Neural Network classifier presented better results than any of the other studies mentioned above. K. Alpan and G. S. İlgi reported higher accuracy of 98.07% using kNN classifier [10] and S. Malik et al. report 98.62% [14]. Our study showed a result of 98.08% employing Neural Network. The results show an AUC of 98.3%. The ROC curve for the negative and positive class is presented in Scalable Vector Graphics (SVG) format as Supplementary Files 2 and 3, respectively. kNN classifier also presents outstandingly results. It achieved 97.31% of accuracy, however, K.Alpan and G. S. İlgi [10] and S. Malik et al. report 98.62% [14] presented better results. With 97.12% of accuracy, the SVM algorithm is the third most accurate classifier. K. Alpan and G. S. İlgi reported 92.11% [10], and M. Peker, O. Özkaraca, and A. Sasar reported 78.51% [13]. Random Forest also presented reliable results when compared to the other studies. It achieved 96.90% of accuracy, presenting a better result than M. Peker, O. Özkaraca, and A. Sasar reporting 91.34% [13]. However, the authors of [9,10,14] reported 97.40%, 97.50% and 98.8%, respectively. Naive Bayes classifier has presented the lowest accuracy of 86.92%. However, it performed better than the H. Naz and S. Ahuja study, which presented 76.33% efficiency [11]. K. Alpan and G. S. İlgi study [10] and M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra [9] outperform our result. Their research reports values of 87.11 and 87.4%, respectively. Finally, the proposed AdaBoost method provides 97.3% accuracy, which is

higher than the results proposed by S. Malik et al. [14] (77.16%). Table 11 includes studies that have used different data sets. This is a limitation regarding the comparison of the results. We use a recent data set, and the number of studies that use the same data source is limited. Therefore, we have compared our results with works that use a different data source following the approach suggested by [14].

Using a 10-fold Cross-Validation technique, we ensure that all the data has been used for training and testing to prevent over-fitting and under-fitting. It is critical to note that some of the cited studies do not present a clear explanation of the parameters used. Consequently, we could not reproduce their results. Nevertheless, all the studies have their own limitations. After a systematic evaluation, we can confirm that the Neural Network is the classification technique that performed most competently. However, while it may fit best for this data set, it might not be the case when applied to another. The data set is also limited. It does not consider family history of diabetes, consumption of certain prescription drugs, smoking, and sleep deprivation. Although data collected grows each day, European regulation regarding general data protection delivers a challenge when obtaining data within EU countries. Hence, we could only use a publicly available data set. We believe that providing data to conduct these experiments can revolutionize modern medicine. The results presented will support future research activities in this domain. The use of machine learning for disease diagnosis is even more essential in the current times. SARS-CoV-2 represents a threat to people living with diabetes. The employment of these procedures may protect lives in a pandemic situation such as for COVID-19. A quick determination of someone's condition will prevent further complications and even avoid life losses. Furthermore, healthcare facilities are under high pressure due to COVID-19 patients and the access to treatment and diagnosis of other diseases is compromised. Therefore, this study suggests that machine learning methods are effective for the early diagnosis of diabetes. However, machine learning or other computer-aided systems will never replace human care since personal contact and medical experience play a crucial rule in clinical decision making.

5. Conclusions

Data Mining has the ability to support clinical decision support systems. A massive amount of data is being collected by medical institutions. These data can be used to support healthcare facilities and public health. Detecting diseases in their early stages might dramatically influence how a person will live to the rest of his days. In this study, we applied classifications methodologies such as Naive Bayes, Neural Network, AdaBoost, kNN, Random Forest, and SVM to a publicly accessible diabetes data set. All methods achieved above 86% efficiency. The authors used normalization on a public data set containing information of 520 patients between 16 and 90 years old. The results have shown that Neural Network is an effective method to diagnose the early stages of diabetes disease. It correctly predicted 510 out of 520 instances, which represents 98.1% accuracy. The experimental validation has been conducted using 10-fold Cross-Validation to avoid over-fitting. The data set has limitations since it does not consider family history of diabetes, consumption of certain prescription drugs, smoking, and sleep deprivation. In the near future, the authors want to conduct the same experiments with new data containing the above-mentioned features and compare the results.

Supplementary Materials: The following are available online at https://www.mdpi.com/2076-3417/11/5/2218/s1, File S1: Supplementary File 1, Figure S2: Supplementary File 2, Figure S3: Supplementary File 3.

Author Contributions: All authors have designed the study, developed the methodology, performed the analysis, and written the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: We choose to exclude this statement as the study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository (UCI Machine Learning Repository) that does not issue DOIs.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
AUC	Area Under the Curve
CA	classification accuracy
FN	False Negative
FP	False Positive
kNN	k-nearest neighbors
TP	True Positive
TN	True Negative
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SVM	support-vector machine

References

1. WHO. Diabetes. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 4 January 2021).
2. Williams, R.; Karuranga, S.; Malanda, B.; Saeedi, P.; Basit, A.; Besançon, S.; Bommer, C.; Esteghamati, A.; Ogurtsova, K.; Zhang, P.; et al. Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **2020**, *162*, 108072. [CrossRef] [PubMed]
3. Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **2019**, *157*, 107843. [CrossRef] [PubMed]
4. American Diabetes Association. How COVID-19 Impacts People with Diabetes. Available online: https://www.diabetes.org/coronavirus-covid-19/how-coronavirus-impacts-people-with-diabetes (accessed on 3 January 2021).
5. Lim, S.; Bae, J.H.; Kwon, H.-S.; Nauck, M.A. COVID-19 and diabetes mellitus: From pathophysiology to clinical management. *Nat. Rev. Endocrinol.* **2021**, *17*, 11–30. [CrossRef] [PubMed]
6. Jothi, N.; Rashid, N.A.; Husain, W. Data Mining in Healthcare—A Review. *Procedia Comput. Sci.* **2015**, *72*, 306–313. [CrossRef]
7. Sun, J.; Reddy, C.K. Big data analytics for healthcare. In Proceedings of the 19th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; p. 1525. [CrossRef]
8. Chitra, R.; Seenivasagam, V. Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques. *ICTACT J. Soft Comput.* **2013**, *3*, 605–609. [CrossRef]
9. Islam, M.M.F.; Ferdousi, R.; Rahman, S.; Bushra, H.Y. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*; Springer: Singapore, 2020; pp. 113–125._12. [CrossRef]
10. Alpan, K.; İlgi, G.S. Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. In Proceedings of the 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 22–24 Octber 2020; pp. 1–7. [CrossRef]
11. Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J. Diabetes Metab. Disord.* **2020**, *19*, 391–403. [CrossRef] [PubMed]
12. Pradhan, N.; Rani, G.; Dhaka, V.S.; Poonia, R.C. 14—Diabetes prediction using artificial neural network. In *Deep Learning Techniques for Biomedical and Health Informatics*; Agarwal, B., Balas, V.E., Jain, L.C., Poonia, R.C., Sharma, M., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 327–339. Available online: https://www.sciencedirect.com/science/article/pii/B9780128190616000148 (accessed on 3 February 2021).
13. Peker, M.; Özkaraca, O.; Sasar, A. Use of Orange Data Mining Toolbox for Data Analysis in Clinical Decision Making: The Diagnosis of Diabetes Disease. In *Expert System Techniques in Biomedical Science Practice*; 2018; pp. 143–167. Available online: https://www.researchgate.net/publication/329707993_Use_of_Orange_Data_Mining_Toolbox_for_Data_Analysis_in_Clinical_Decision_Making_The_Diagnosis_of_Diabetes_Disease (accessed on 3 February 2021).

14. Malik, S.; Harous, S.; El-Sayed, H. Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women. In Proceedings of the International Symposium on Modelling and Implementation of Complex Systems, Batna, Algeria, 24–26 October 2020; Springer: Cham, Switzerland, 2020.
15. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
16. Gandhi, R. Naive Bayes Classifier. Medium, 17 May 2018. Available online: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed on 7 January 2021).
17. Maind, S.B.; Wankar, P. Research paper on basic of artificial neural network. *Int. J. Recent Innov. Trends Comput. Commun.* **2014**, *2*, 96–100.
18. Boosting Algorithm. Boosting Algorithms in Machine Learning. Analytics Vidhya, 9 November 2015. Available online: https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/ (accessed on 6 January 2021).
19. Taunk, K.; De, S.; Verma, S.; Swetapadma, A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 15–17 May 2019; pp. 1255–1260. [CrossRef]
20. Li, S.; Zhang, K.; Chen, Q.; Wang, S.; Zhang, S. Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm. *IEEE Access* **2020**, *8*, 139512–139528. [CrossRef]
21. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.:1010933404324. [CrossRef]
22. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
23. A Complete Guide to the Random Forest Algorithm. Built In. Available online: https://builtin.com/data-science/random-forest-algorithm (accessed on 7 January 2021).
24. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
25. Yang, Y.; Li, J.; Yang, Y. The research of the fast SVM classifier method. In Proceedings of the 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 18–20 December 2015; pp. 121–124. [CrossRef]
26. Jo, J.M. Effectiveness of normalization pre-processing of big data to the machine learning performance. *J. Korea Inst. Electron. Commun. Sci.* **2019**, *14*, 547–552.
27. Dalwinder, S.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524.
28. Karegowda, A.G.; Manjunath, A.S.; Jayaram, M.A. Comparative study of attribute selection using gain ratio and correlation based feature selection. *Int. J. Inf. Technol. Knowl. Manag.* **2010**, *2*, 271–277.
29. Narkhede, S. Understanding AUC—ROC Curve. Medium, 26 May 2019. Available online: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (accessed on 7 January 2021).
30. Berrar, D. Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 542–545. Available online: https://www.researchgate.net/profile/Daniel_Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf (accessed on 3 February 2021).
31. Battineni, G.; Sagaro, G.G.; Nalini, C.; Amenta, F.; Tayebati, S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. *Machines* **2019**, *7*, 74. [CrossRef]
32. Ayon, S.I.; Islam, M. Diabetes Prediction: A Deep Learning Approach. *Int. J. Inf. Eng. Electron. Bus.* **2019**, *11*. [CrossRef]