Jundishapur J Chronic Dis Care. 2019 July; 8(3):e94173. doi: 10.5812/jjcdc.94173.
Published online 2019 July 14. **Research Article**

Early Diagnosis of Diabetes Mellitus Using Data Mining and Classification Techniques

Seyed Ataaldin Mahmoudinejad Dezfuli [1,*], Seyedeh Razieh Mahmoudinejad Dezfuli [2], Seyed Vafaaldin Mahmoudinejad Dezfuli [1] and Younes Kiani [2]

[1] Technology Development Center, Dezful University of Medical Sciences, Dezful, Iran
[2] Islamic Azad University of Dezful, Dezful, Iran

[*] *Corresponding author*: School of Medicine, Dezful University of Medical Sciences, Dezful, Iran. Tel: +98-9370776019, Email: s.atamahmoudi@gmail.com

Received 2019 May 21; **Revised** 2019 June 26; **Accepted** 2019 July 05.

Abstract

Background: According to the World Health Organization, the seventh major cause of human death in 2030 will be diabetes, which of course is a very severe disease and if not treated thoroughly and on time, can lead to critical problems, including death. Accordingly, diabetes is one of the main priorities in medical science researches, which usually produce lots of information. The role of data mining methods in diabetes research is critical, which is considered as one of the optimum procedures of extracting knowledge from a large amount of diabetes-related data.
Objectives: This research has focused on developing an ensemble system using data-mining methods based on three classification methods, namely, weighted k-nearest neighbor, simple decision tree and logistic regression algorithms to detect diabetes mellitus of the human.
Methods: The proposed ensemble method algorithm applies votes given by each of the classifiers to attain the final result. This voting mechanism considers each estimation of the classifiers as an input to the ensemble system and then computes the statistical mode for its output to get the majority vote.
Results: Apparently, these classifiers give the accuracy of 77.00%, 77.30%, 79.30%, and 80.60% for decision tree, weighted k-nearest neighbor, logistic regression, and the ensemble method, respectively.
Conclusions: The results of the proposed method illustrate an acceptable improvement of accuracy compared to other methods. Consequently, it supports the idea that hybrid approaches are more effective in comparison with the simple classification methods that use classifiers separately.

Keywords: Chronic Disease, Diabetes Mellitus, Early Diagnosis, Data Mining, Classification, Decision Tree, Logistic Regression, Weighted K-Nearest Neighbor, Cross Validation

1. Background

1.1. Diabetes Mellitus

With regard to the increased prevalence of diabetes, the high cost of treatment for diabetic patients, along with many deaths and medical errors, diabetic prevention is more essential than its treatment (1). The number of diabetic people has increased from 108 million in 1980 and now affects over 422 million people globally (2). Furthermore, the global prevalence of adults' diabetes who were over 18 years old has increased from 4.7% to 8.5% by 1980 to 2014, and it has been growing more quickly in the low and middle -income countries (2). Correspondingly, diabetes is a chronic disease and the main cause of kidney failure, blindness, heart attacks, lower limb amputation, and stroke (2). In 2012, 1.5 million people passed away directly from diabetes, and there were 2.2 million deaths due to high levels of blood glucose (BGL) (2). The BGL is a statistical concept, defined as the distribution of fasting plasma glucose (FPG) that uses in the diagnosis of type 2 diabetes mellitus (T2DM), when it is exceeding 130 (mg/dl) (3). Accordingly, the world health organization (WHO) declares that in 2030 diabetes will be the 7th leading cause of death (4). Thus, diabetic prevention with early diagnosis of diabetes is our main research priority.

Diabetes mellitus is a metabolic disease that affects the body's ability to adjust BGL (5). In diabetes, there will be an abnormal increase in blood glucose level (hyperglycemia) which leads to significant medical conditions, including ischemic heart disease, stroke, nephropathy, neuropathy, retinopathy, and peripheral vascular disease (5, 6).

One of the reasons for hyperglycemia in humans is insulin deficiency, which occurs when beta cells in the pancreas are no longer able to produce insulin. This condition is known as the type one diabetes mellitus (T1DM) (5, 7, 8). People with T1DM need a daily injection of insulin to regulate their blood glucose levels, and if they don't access to insulin, they cannot survive. The cause of T1DM is unknown and also is not currently preventable, but its symptoms include urination and persistent hunger, excessive thirst, visual changes, weight loss, and fatigue (2).

The other type and the most common form of diabetes mellitus, which is caused by insufficiency of insulin secretion, is that the body does not produce sufficient insulin and insulin does not affect the cells (type two diabetes mellitus) (T2DM) (2, 9). Although, the T2DM symptoms are similar to T1DM but are often less common or totally absent. As a result, the disease may not be detected for several years until its effects exhibit in the present (2). Approximately 50% - 80% of T2DM cases are not diagnosed (7, 10).

The interaction of genetic and metabolic factors will determine the risk of T2DM. Family history of diabetes, overweightness, obesity, ethnicity, previous gestational diabetes compound with older age, physical inactivity, unhealthy diet, and smoking will increase the risk of diabetes (2). Both genetics and environmental factors, such as race, obesity, age, gender, and lack of exercise, evidently play significant roles in diabetes diagnosis whereby overweightness and obesity are the influential risk factors for T2DM (2, 5, 7, 9, 11-13).

Correspondingly, Marinov et al. (14) investigated 17 articles describing different data-mining methods utilized for diabetes research. They expressed that data mining can play a dominant role in diabetes research and ultimately improve the quality of health care for diabetes patients. Likewise, the significance of our study is that the early diagnosis of diabetes reduces the cost of treatment.

If diabetes is not treated in time and suitably, it will lead to very severe complications including death; as a result of which the disease is one of the main priorities in medical research that generates a wealth of clinical data (14). Data mining is the process of extraction of useful knowledge from a large amount of clinical data to predict using techniques such as classification, clustering, and association, which make it one of the effective methods in diabetes research (12). Data mining can significantly help diabetes research and ultimately improve the quality of health care (14, 15). Data mining methods in disease diagnosis using many complex machine-learning algorithms to discover a hidden pattern, increase the accuracy rate of detection which such identified patterns utilized to predict upcoming events (12, 15).

This paper aims at developing an ensemble diabetes early diagnosis system, which can forecast whether the patient has diabetes or not. Moreover, this system using classification data mining methods, namely weighted k-nearest neighbor, simple decision tree, and logistic regression, which can extract knowledge from clinical patient data.

In section 1 we will review an introduction on DM and the importance of applying data mining techniques in the medical field. Furthermore, we will have a quick review on the related research background. Section 2 consists of data acquisition, pre-processing and data normalization, data analysis, reviewing different classifiers, cross-validation technique, confusion matrix and the proposed method. Section 3 will portray the result obtained by the proposed method. In section 4 we will discuss about comparative analysis of the result of different studies with our proposed method. Finally, the section 5 is the conclusion section of this research. Section 6 is about suggestions and future works of this study and the last section 7 is about the compliance with ethical standards.

1.2. Research Background

Most diabetes researches focus on two general approaches using data mining for early diagnosis of diabetes. The first approach is to use a specific classification algorithm to estimate the risk of diabetes on the patient's diabetes data, and the second approach is to apply hybrid algorithms. For example, Thirumal and Nagarajan (12) discussed different data mining algorithms namely, decision tree, k-nearest neighbor, naïve Bayes, and SVM tested with Pima Indian diabetes dataset. The main goal is to get the best algorithms that with given data provides higher accuracy. Likewise, Lee and Wang (9) presented a novel fuzzy expert system for diabetes diagnosis support application which could give a semantic description of diabetes. Conversely, Tafa et al. (16) proposed the joint implementation of two algorithms, namely support vector machine SVM and naïve Bayes to minimize their specific weakness such that the accuracy of the joint ensemble method was improved up to 97.6%. This methodology could decrease the false negative answers, which is a crucial issue in medical diagnoses. Accordingly, Han et al. (17) used an SVM along with an ensemble learning module which turns the "black box" of SVM decisions into logical rules to monitor diabetes. Furthermore, the study illustrates that the hybrid system is efficient and can provide a tool for diabetes diagnosis. Likewise, Barakat et al. (7) used the same technique and represented that the intelligible SVMs provide a promising tool for the prediction of diabetes. Furthermore, De Silva et al. (15) developed an ensemble system which used a voting process between three classification

methods namely, decision tree, naïve Bayes, and SVM algorithm to get the final result.

Lee and Kim (11) evaluated the predictive power of different phenotypes using hypertriglyceridemic waist (HW) and specific anthropometric measurements such as waist circumference (WC) and triglyceride (TG) levels as the constituents of the HW phenotype. The study showed that the relationship between WC and T2DM was higher than TG association with T2DM. However, the results of this study cannot be generalized to other populations, because the study population was only Korean women and men (11). Likewise, Lee et al. (3) employed an arrangement of anthropometric measures as the input of two different machine learning algorithms to predict the fasting plasma glucose (FPG) status and the results indicate that using normalized data of high and normal FPG groups can enhance the estimation and decrease the intrinsic bias of the model toward the majority class. Moreover, Simon et al. (5) aimed to discover sets of diabetes risk factors by applying association rule mining to electronic medical records (EMR). Similarly, Purushottam et al. (18) designed a system which efficiently discovers the rules to estimate the risk level of diabetes using C45 rules and partial tree to evaluate the system.

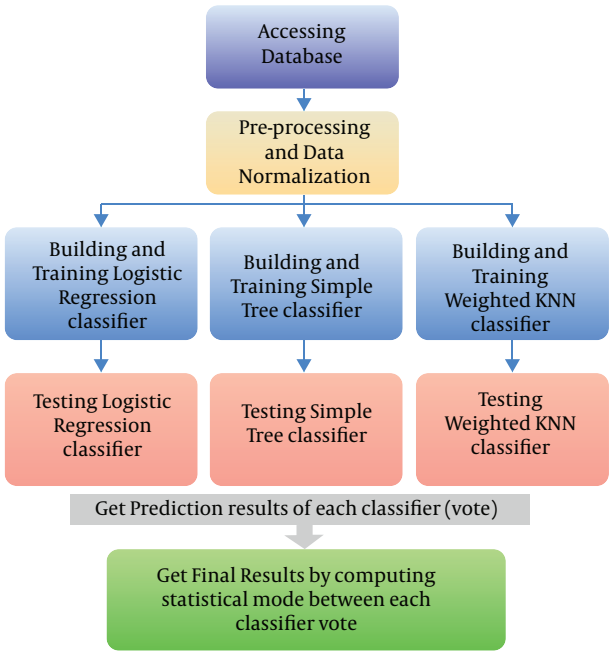

Figure 1. Proposed ensemble method algorithm

2. Methods

2.1. Data Acquisition

The principle of this study is to develop an ensemble system using several data-mining methods and evaluate different diabetes risk factors for DM diagnosis. We used three different classifiers, namely weighted KNN, decision tree, and logistic regression after the preprocessing stage and trained them with the diabetes database. Furthermore, we considered the prediction of each classifier as a vote to develop an ensemble algorithm such that the voting system brings us the majority predicting of classifiers using statistical mode between their predictions. The algorithm of our proposed method shown in Figure 1.

The research data extracted from the Pima Indian diabetes dataset available at the UCI machine learning repository (19). Indeed, the original owner of this database is the National Institute of Diabetes and Digestive and Kidney Diseases. In particular, all patients here are females and at least 21 years old of Pima Indian heritage. Let's noted that we have eliminated cases that have missing data. Consequently, we used 392 cases from 768 cases. The final data set constituted from 33.2% of instances with class value 1 and 66.8% of instances with class value 0 that encompassed 160 cases and 262 cases of total 392 cases respectively.

2.1.1. Attribute Information

1. Number of times pregnant = NPreg
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test = PGOGTT
3. Diastolic blood pressure (mm Hg) = DBS
4. Triceps skin fold thickness (mm) = TSFT
5. Two-hour serum insulin (mu U/mL) = SI
6. Body mass index (weight in kg/(height in m)^2) = BMI
7. Diabetes pedigree function = DPF
8. Age (years) = Age
9. Class variable (0 or 1)
10. Class distribution: (class value 1 is defined as "tested positive for diabetes" and class value 0 is defined as "tested negative for diabetes").

2.2. Pre-Processing and Data Normalization

Initially, we employed a personal laptop which includes a windows 10 with 64-bit operating system constituted of hardware with an Intel(R) Core(TM) i7-4710HQ CPU, and 12 GB of memory. Furthermore, we utilized the Matlab engineering software version 2016 to implement the study procedure. The statistical analysis of Pima Indian diabetes dataset is presented in Tables 1 and 2. As shown in Table 1, the range of values differs widely. Therefore, a normalization method implemented and normalized the data between 0 and 1 value to control the data scattering, and consequently improving data classification results (3).

Table 1. Data Statistics Analysis Before Normalization

Number	Attribute	Mean	Standard Deviation	Coefficient of Variation (CV)	Min	Max
1	NPreg	3.30	3.21	0.97	0	17
2	PGOGTT	122.62	30.86	0.25	56	198
3	DBS	70.66	12.49	0.17	24	110
4	TSFT	29.14	10.51	0.36	7	63
5	SI	156.05	118.84	0.76	14	846
6	BMI	33.08	7.02	0.21	18.20	67.10
7	DPF	0.52	0.34	0.66	0.08	2.42
8	Age	30.86	10.20	0.33	21	81

Table 2. Data Statistics Analysis After Normalization

Number	Attribute	Mean	Standard Deviation	Coefficient of Variation (CV)	Min	Max
1	NPreg	0.19	0.18	0.97	0	1
2	PGOGTT	0.46	0.21	0.46	0	1
3	DBS	0.54	0.14	0.26	0	1
4	TSFT	0.39	0.18	0.47	0	1
5	SI	0.17	0.14	0.76	0	1
6	BMI	0.30	0.14	0.21	0	1
7	DPF	0.18	0.14	0.66	0	1
8	Age	0.16	0.17	0.33	0	1

2.3. Data Analysis

The distribution of attribute values before and after data normalization is shown in Figures 2 and 3. As shown, during the preprocessing stage, the scattering of data controlled and scaled between interval 0 to 1 successfully. Generally, in this study, we did complete data analysis along with graphical summaries. We used histogram since it is a precise graphical demonstration of the distribution of numerical data.

2.4. Decision Tree

Decision tree learning is known as a predictive modeling approach used in statistics which maps observations about an item to decide about the target value of the item. In tree structures, branches describe conjunctions of features and leaves signify class labels (20).

2.5. K-Nearest Neighbors' Algorithm

The k-nearest neighbors' algorithm (KNN) as a non-parametric approach used for classification in which the input includes the k adjacent training samples in the feature space, where the output is a class member. In KNN, consider k as a small positive integer, the object assigns to the class most common among its k adjacent neighbors.

1. Consider $L = \{(y_i, x_i), i = 1, ..., nL\}$ as a learning set of observations x_i along with class membership y_i. In addition, consider x be a new observation, whose class label y must be predicted.

2. Discover the k + 1 adjacent neighbors to x using a distance function d (x, x(i)).

3. Use the (k + 1) th neighbor in order to standardize the k minimum distances via

$$D(i) = D(x, x(i)) = \frac{d(x, x(i))}{d(x, x(k+1))} \quad (1)$$

4. Apply any kernel function K(.) to transform the normalized distances $D_{(i)}$ into weights $w_{(i)} = K(D_{(i)})$.

5. Choose the class, which shows a weighted majority of the k adjacent neighbors to estimate the class membership y of observation x (21):

$$\hat{y} = max_r \left(k \sum_{i=1}^{k} \omega_{(i)} I\left(y_{(i)} = r\right) \right) \quad (2)$$

2.6. Logistic Regression

Since the logistic regression uses the standard logistic function interpreted as a probability which takes any real input t, ($t \epsilon R$), whereas the output always lay between zero and one (22, 23). The logistic function σ (t) is:

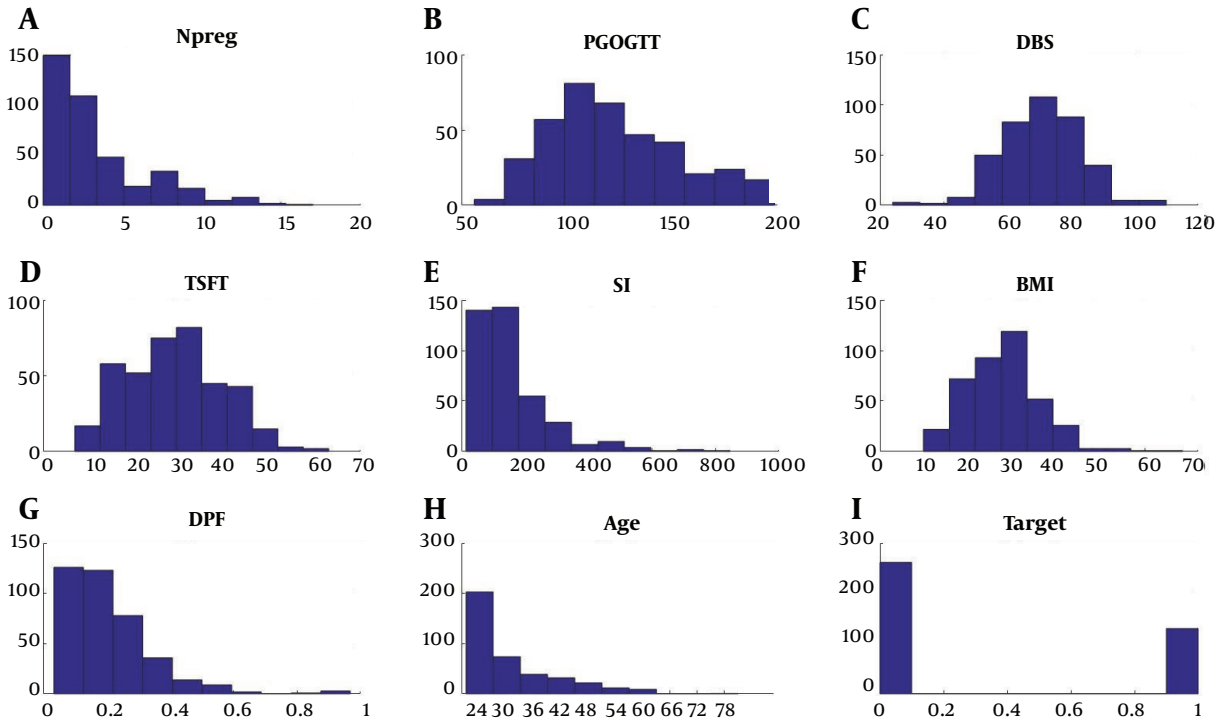

Figure 2. Data distribution before normalization

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \quad (3)$$

Consider t as a linear function of a single explanatory variable x. so, t equals:

$$t = \beta_0 + \beta_1 x \quad (4)$$

And the logistic function is:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (5)$$

Note that F(x) interprets as the probability of the dependent variable (23).

2.7. Cross Validation Technique

Cross-validation is a data mining method utilized to estimate the error rate and the performance of classification algorithms. Here, the dataset partitions into n-folds which comprises the testing and training sets. This process repeats n times for the testing and training data sets and, finally, the error rates for n sets are averages to yield an overall error rate (12).

We applied 10-fold cross-validation technique for each of the classifiers using Matlab engineering software to estimate the predictive accuracy of the model trained with all the data. Consequently, we obtained the confusion matrix toward each of the classifiers as well. However, the method requires multiple epochs, but it makes efficient use of all the data.

2.8. Confusions Matrix

The confusion matrix is applicable to perform the accuracy of classifiers and to show the relationship between outcomes and predicted classes (12). According to the Matlab, the confusion matrix scheme includes the rows and the columns which are equal to the predicted class (Output Class) and the correct class (Target Class), respectively. Moreover, the diagonal cells indicate the percentage of predicted classes correctness while the off-diagonal cells represent the classifier mistakes. Furthermore, the right column and the lowest row of confusion matrix determines the accuracy of each predicted and correct class, respectively and, the bottom right cell defines the overall precision.

3. Results

According to Matlab engineering software, in Figures 4 to 7, the first two diagonal cells display the percentage and number of correct classifications by the trained classifier.

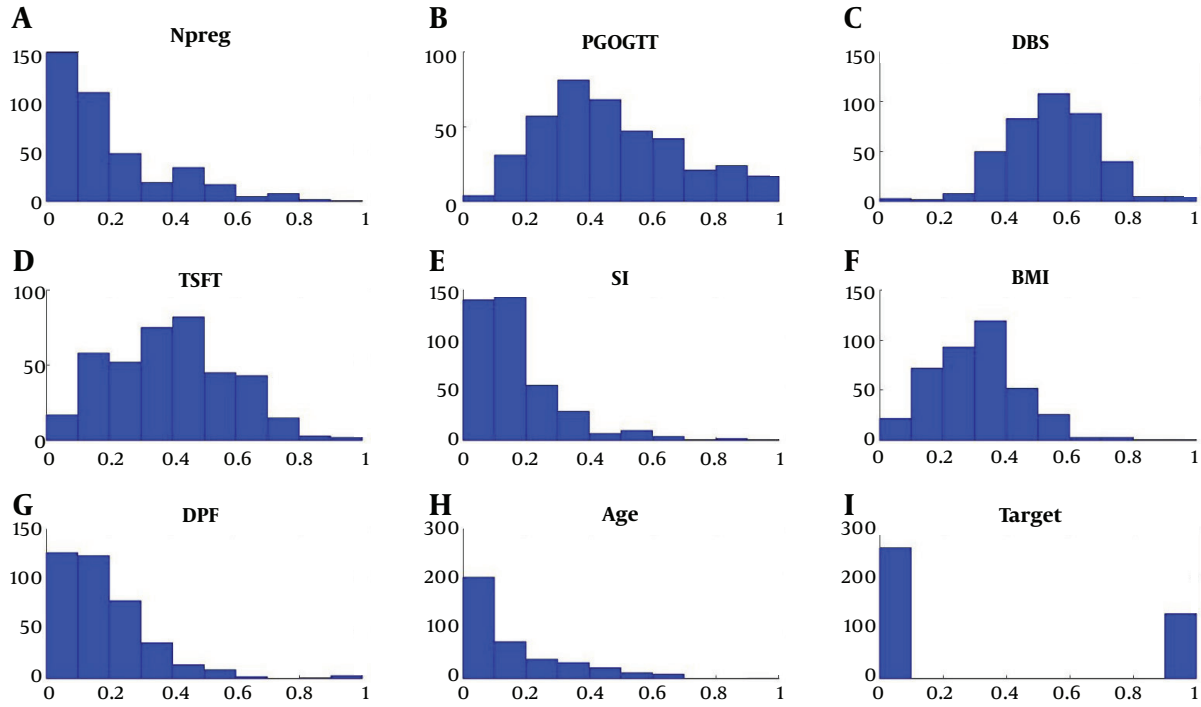

Figure 3. Data distribution after normalization

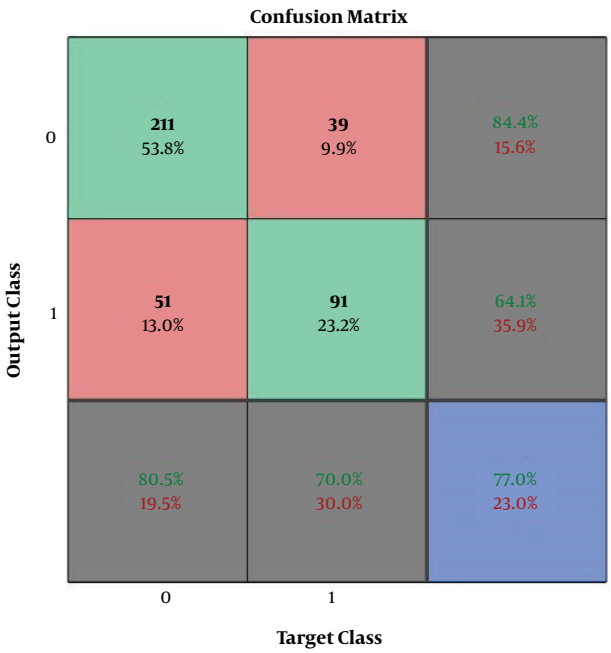

Figure 4. Confusion matrix for simple tree classifier

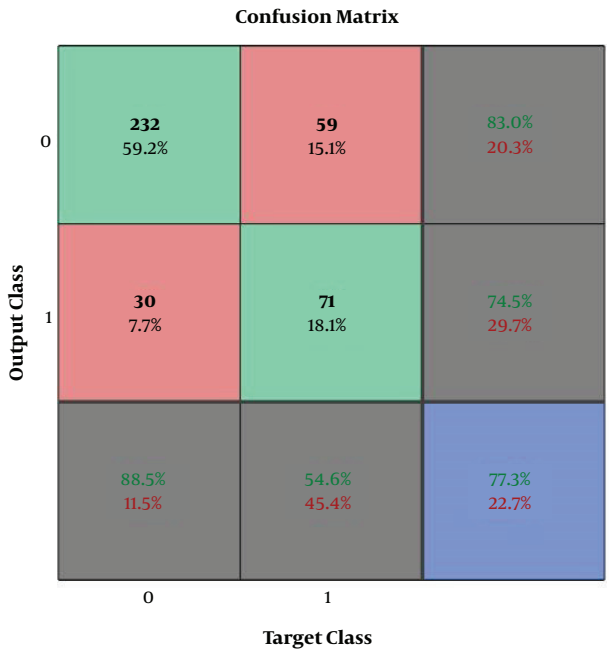

Figure 5. Confusion matrix for weighted KNN classifier

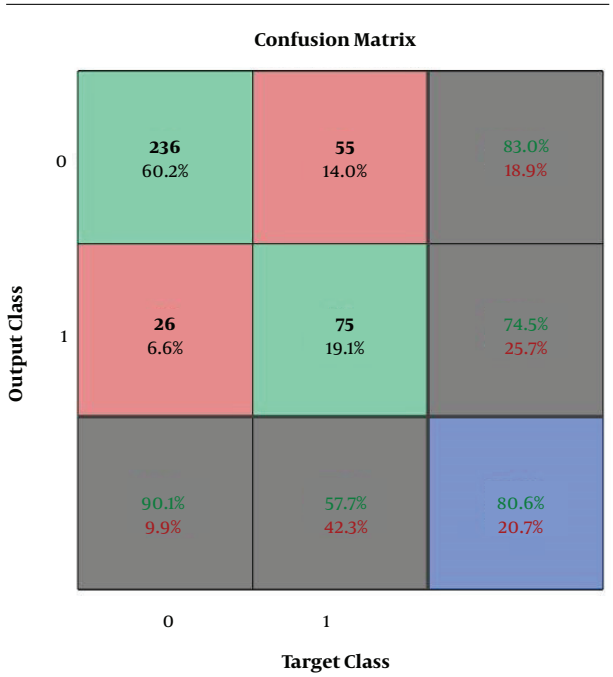

Figure 6. Confusion matrix for logistic regression classifier

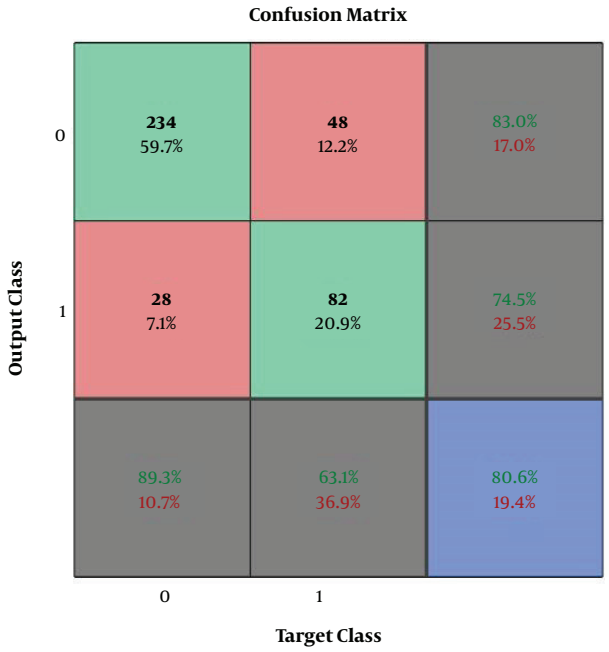

Figure 7. Confusion matrix for proposed ensemble algorithm

In Figure 4, 211 samples were correctly classified as non-diabetic patients, which represent 53.8% of all 392 cases, and similarly, 91 samples were correctly classified as diabetic patients, which represent 23.2% of all cases.

Furthermore, 39 diabetic cases were incorrectly classified as non-diabetic patients, which represent 9.9% of all 392 cases, and similarly, 51 non-diabetic patients were incorrectly classified as diabetic patients, which represent 13.0% of all data.

Additionally, 15.6% of 250 non-diabetic predictions are incorrect and 84.4% are correct. Likewise, 35.9% of 142 diabetic predictions are false, and 64.1% are true. Also, 80.5% of 262 none-diabetic cases correctly predicted as non-diabetic while, 19.5% predicted as diabetic patients. 70.0% of 142 diabetic cases correctly classified as diabetic and 30.0% classified as none-diabetic patients.

Overall, 77.0% of the predictions are correct, and 23.0% are wrong. Also, our final study results presented in Table 3.

Table 3. The Comparative Analysis Between Result of the Proposed Method and Other Research's Results[a]

Authors	Method	Accuracy
De Silva et al. (15)		
	Naïve Bayes classifier	77.86%
	C4.5 tree classifier	78.25%
	SVM classifier	77.47%
	KNN classifier	77.73%
Purushottam et al. (18)		
	Using association rules for C4.5 tree classifier	81.27 %
Proposed method		
	Simple tree classifier	77.0%
	Weighted KNN	77.3%
	Logistic regression	79.3%
	Ensemble method	80.60%

[a]For more research results, comparisons and information check the following link (19): http://fizyka.umk.pl/kis-old/projects/datasets.html#Diabetes

4. Discussion

The idea of utilizing a hybrid approach derived from one of our research background studies and motivated us to check their method with different classifiers (15). In this study, we investigated the efficiency of hybrid classification algorithms in diabetes diagnosis proposing an ensemble method in comparison with single independent classifiers using Pima Indian database. Some research papers applied Pima Indian database of diabetes, and there-

fore have approach attributes and similar conclusions (12, 15). Additionally, they utilized several algorithms independently to compare the efficiency of the algorithms among each other (9, 12). However, some studies suggested the hybrid application of algorithms, a statistical-based method, or the combination of classification, clustering, and even association rule-based systems (5, 7, 15-18). Accordingly, let us compare the results of our proposed technique to other methods' outcomes, which utilized the Pima Indian database. Notice that since the proposed method's accuracy affects by the initial parameters of some classifiers, therefore it can even achieve up to 82% of precision. As shown, it is clear that the result of the proposed ensemble method is always better than each classifier's outcome, and the overall accuracy of the proposed ensemble algorithm is significant comparing to other methods.

4.1. Conclusions

The results of the proposed method show an acceptable improvement of accuracy compared to other methods. Consequently, our results support that the hybrid method is more efficient in comparison with single data-mining methods.

4.2. Suggestions and Future Studies

Here are several suggestions for future studies to improve the performance of the ensemble proposed method, such as using other types of classifiers and then comparing the results. Moreover, applying the certified databases containing moderate positive and negative cases which were acquired with precision. Furthermore, using neuro-fuzzy systems and other data mining methods is also recommended.

Footnotes

Authors' Contribution: Study concept and design: Seyed Ataaldin Mahmoudinejad Dezfuli and Seyed Vafaaldin Mahmoudinejad Dezfuli. Analysis and interpretation of data: Seyedeh Razieh Mahmoudinejad Dezfuli and Younes Kiani. Drafting of the manuscript: Seyed Ataaldin Mahmoudinejad Dezfuli and Seyed Vafaaldin Mahmoudinejad Dezfuli. Critical revision of the manuscript for important intellectual content: Seyed Ataaldin Mahmoudinejad Dezfuli. Statistical analysis: Seyedeh Razieh Mahmoudinejad Dezfuli and Younes Kiani.

Conflict of Interests: We declare that there is no conflct of interests regarding the publication of this paper.

Ethical Approval: The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a prior approval by the Institution's Human Research Committee.

Financial Disclosure: There are no financial interests related to the material in the manuscript.

Funding/Support: Since data has obtained from Pima Indian database free of charge, this study has not funded by any real person or legal institution.

Patient Consent: As we mentioned this article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Zarkogianni K, Litsa E, Mitsis K, Wu PY, Kaddi CD, Cheng CW, et al. A review of emerging technologies for the management of diabetes mellitus. *IEEE Trans Biomed Eng*. 2015;**62**(12):2735–49. doi: 10.1109/TBME.2015.2470521. [PubMed: 26292334]. [PubMed Central: PMC5859570].
2. World Health Organization. *Global report on diabetes*. Geneva: World Health Organization; 2016. Available from: http://www.who.int/diabetes/global-report/en/..
3. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE J Biomed Health Inform*. 2014;**18**(2):555–61. doi: 10.1109/JBHI.2013.2264509. [PubMed: 24608055].
4. Kor LK, Ahmad AR, Idrus Z, Mansor KA. *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. Singapore: Springer; 2019.
5. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW. Extending association rule summarization techniques to assess risk of diabetes mellitus. *IEEE Trans Knowl Data Eng*. 2015;**27**(1):130–41. doi: 10.1109/tkde.2013.76.
6. WHO/IDF. *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: Report of a WHO/IDF consultation*. World Health Organization; 2006. Available from: https://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf.
7. Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed*. 2010;**14**(4):1114–20. doi: 10.1109/TITB.2009.2039485. [PubMed: 20071261].
8. Senthil Kumar B, Rajapandi P, Sridar K, Shanthi D. Mobile based medical diagnosis system using Ann-Arm for the diabetes mellitus. *Int J Innovative Sci Eng Technol*. 2015;**2**(4).
9. Lee CS, Wang MH. A fuzzy expert system for diabetes decision support application. *IEEE Trans Syst Man Cybern B Cybern*. 2011;**41**(1):139–53. doi: 10.1109/TSMCB.2010.2048899. [PubMed: 20501347].
10. International Diabetes Federation. *IDF Diabetes Atlas, 5th ed*. Brussels, Belgium: International Diabetes Federation; 2011. Available from: http://www.idf.org/diabetesatlas.
11. Lee BJ, Kim JY. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE J Biomed Health Inform*. 2016;**20**(1):39–46. doi: 10.1109/JBHI.2015.2396520. [PubMed: 25675467].
12. Thirumal PC, Nagarajan N. Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. *ARPN J Eng Appl Sci*. 2015;**10**(1):8–13.
13. Lorenzo C, Serrano-Rios M, Martinez-Larrad MT, Gonzalez-Villalpando C, Williams K, Gabriel R, et al. Which obesity index best explains prevalence differences in type 2 diabetes mellitus? *Obesity (Silver Spring)*. 2007;**15**(5):1294–301. doi: 10.1038/oby.2007.151. [PubMed: 17495206].
14. Marinov M, Mosa AS, Yoo I, Boren SA. Data-mining technologies for diabetes: A systematic review. *J Diabetes Sci Technol*. 2011;**5**(6):1549–56.

doi: 10.1177/193229681100500631. [PubMed: 22226277]. [PubMed Central: PMC3262726].
15. De Silva LHS, Pathirage N, Jinasena TMKK. Diabetic prediction system using data mining. *Proceedings in Computing, 9th International Research Conference-KDU*. Sri Lanka. 2016.
16. Tafa Z, Pervetica N, Karahoda B. An intelligent system for diabetes prediction. *4thMediterranean Conference on Embedded Computing MECO*. Budva, Montenegro. 2015. p. 378–82.
17. Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes. *IEEE J Biomed Health Inform*. 2015;**19**(2):728–34. doi: 10.1109/JBHI.2014.2325615. [PubMed: 24860043].
18. Saxena K, Sharma R. Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree. Noida, India. IEEE; 2015. p. 1–6.
19. Blake CL, Merz CJ. *UCI repository of machine learning databases*. Irvine: University of California; 1998.
20. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;**1**(1):81–106. doi: 10.1007/bf00116251.
21. Hechenbichler K, Schliep K. Weighted k-nearest-neighbor techniques and ordinal classification. *Col Res Center*. 2004. doi: 10.5282/ubm/epub.1769.
22. Taylor AP, Webb RI, Barry JC, Hosmer H, Gould RJ, Wood BJ. Adhesion of microbes using 3-aminopropyl triethoxy silane and specimen stabilisation techniques for analytical transmission electron microscopy. *J Microsc*. 2000;**199**(Pt 1):56–67. [PubMed: 10886529].
23. Freedman DA. *Statistical models: Theory and practice*. Cambridge University Press; 2009. 128 p. doi: 10.1017/cbo9780511815867.