World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

Comparison of feature reduction and feature selection in early diabetes classification

Mustafa Çalışkan [1, *], Kenan Türkeri [2] and Mehmet Ali Dağdeviren [1]

[1] *Şehit Karayılan Vocational and Technical Anatolian High School, Gaziantep, Turkey.*
[2] *Şehit Şahinbey Special Education Practice School (I.II.III.LEVEL), Gaziantep, Turkey.*

World Journal of Advanced Engineering Technology and Sciences, 2024, 11(01), 209–215

Publication history: Received on 22 December 2023; revised on 04 February 2024; accepted on 07 February 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.11.1.0038

Abstract

Diabetes is one of the common health problems encountered today, and this problem is increasing day by day due to unbalanced and unconscious eating habits. Once a person is diagnosed with diabetes, the likelihood of recovery from this disease is often low. If a person is diagnosed with diabetes, the individual must usually take medication and/or follow a strict diet program for life. While this may be somewhat more manageable for patients living in developed countries, it is often difficult for citizens in developing countries to access these facilities. Because it is generally more difficult and costly to access medicine and healthy nutrition in these countries. Therefore, in this study, ways to diagnose diabetes at an early stage using machine learning techniques are examined. In the study, symptoms such as age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital fungus, blurred vision, itching, irritability, delayed healing, partial paralysis, cramps, hair loss and obesity are examined and which parameters were more effective in diagnosing diabetes.

Keywords: Machine learning; Diabetes; AI; Habits

1. Introduction

According to the World Health Organization reports, there are 422 million diabetic patients in the world [1]. Diabetes affects almost the entire body, causing irreversible losses in the most critical areas of the body such as the heart, kidneys, eyes, brain, nerves and blood vessels. However, every year 1.6 million people lose their lives directly due to diabetes [1]. Although diabetes is seen in all countries of the world, it mostly affects people living in low- and middle-income countries. People living in these countries only consult a physician if they develop a serious complication related to diabetes. This causes the disease to reach untreatable levels [2].

Diabetes is a chronic disease that affects not only physical health but also psychological and emotional health. Living with diabetes requires constantly checking blood sugar, taking insulin injections, and following a special diet. This situation requires constant awareness and effort in the individual's daily life, which can cause psychological problems such as stress, anxiety and depression. Additionally, concern about the long-term health complications of diabetes and the effects of this disease on social life can also affect psychological health. In this context, the process of coping with diabetes should include not only medical treatment but also psychological support and counseling services. Thus, individuals can both manage their physical health and maintain their emotional well-being [2], [3].

This study aims to diagnose early diabetes based only on the person's appearance and the basic questions the person answers. While this test can be performed easily, it can prevent time and economic losses by providing a serious guide to the specialist physician. The main purpose of this study is to diagnose the person quickly and effectively for preventive purposes. It is known that some problems seen as a result of diabetes before the diagnosis of diabetes is made are important in the early diagnosis of diabetes. The most basic parameters examined when diagnosing a person

[*] Corresponding author: Mustafa Çalışkan

with diabetes are age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital fungus, blurred vision, itching, irritability, delayed healing, partial paralysis, cramps, alopecia areata and obesity. is a custom [4], [5].

The data set used in the study includes 16 features [4], [5]. The classification process [6], [7] was carried out by systematically selecting attributes ranging from eight to one among these attributes. In addition, these 16 features were examined in terms of all principal components between 1 and 16 with the help of principal component analysis [8]. Accordingly, it has been revealed which approach is more advantageous for this data set. Throughout the process, classification was carried out using support vector machines (SVM) [9], Decision tree (DT) [10], Nearest Neighborhood (KNN) [11] and Navie Bayes (NB) [12] algorithms.

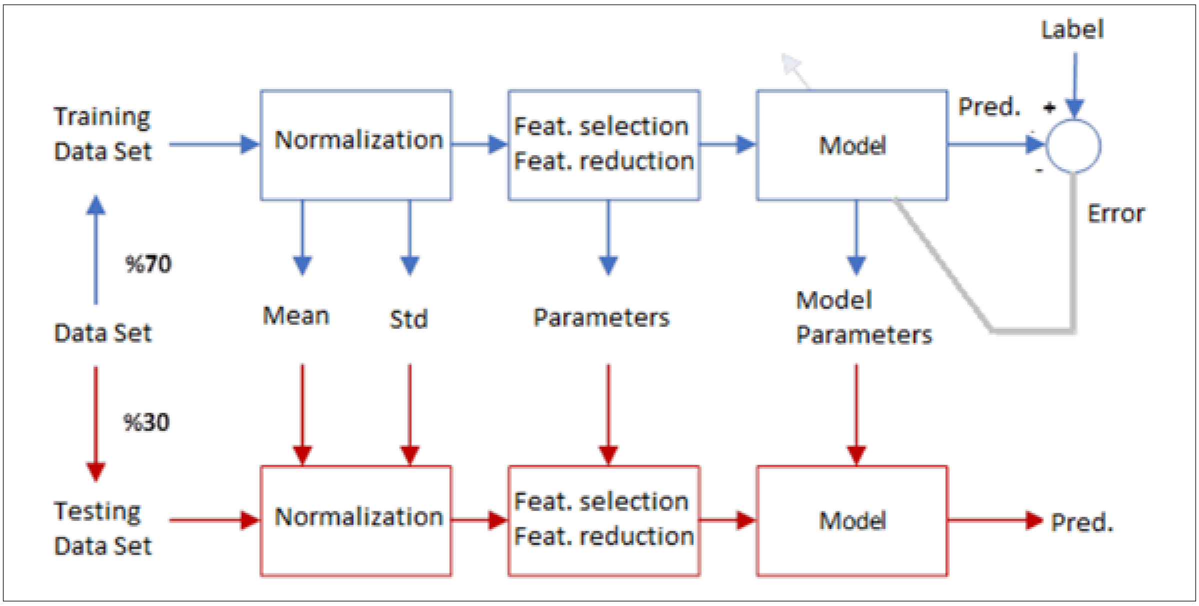

Figure 1 Proposed method

2. Method

Classification is used in many areas of science. In recent years, the increase in data received from patients and the high workload of specialist physicians in the healthcare system have led to the increase of machine learning-based methods in solving medical problems. The input matrix for the classifier can be defined as a combination of these features in a certain arrangement.

Such studies contribute to important issues in the field of medicine, such as early diagnosis of diseases, treatment planning and patient management. The use of classification science in the field of medicine allows patients to be evaluated more quickly and effectively, to reach the correct diagnosis and to optimize treatment processes.

There are 16 features in the data set used in this study. These attributes are age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital fungus, blurred vision, itching, irritability, delayed healing, partial paralysis, cramps, alopecia areata and obesity. The input of the classifier is a matrix as follows. This matrix is a structure in which each attribute is arranged in a certain order and contains the values that each patient has for these attributes. Classification algorithms have the ability to recognize a specific disease or classify a condition by analyzing this input matrix.

$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \dots\ \mathbf{x}_k] \text{ .... } (1)$$

where $\mathbf{x}_1,\ \mathbf{x}_2, \mathbf{x}_3 \dots \mathbf{x}_k$ p denote p-dimensional feature vectors. In this study, $k = 16, p = 520$. The output of the data set is created as shown below:

$$\mathbf{y} = [y_1\ y_2\ y_3\ \dots\ y_p]^{\mathrm{T}} \text{...............} (2)$$

where $y_1\ , y_2, y_3\ and\ y_p$ represent the labels related to the above examples. $y_i = 0\ or\ 1$. Diabetes is expressed as 1 and healthy person as 0. In classification problems, the data set is generally divided into two groups: training

$\{X_{eğitim}, \mathbf{y}_{eğitim}\}$ data and testing $\{X_{test}, \mathbf{y}_{test}\}$ data. **X** and **y** are defined in Equation-1 and Equation-2. In some cases, data can be grouped in different ways using the dataset cross-validation technique. In this study, 30% of the data set is used for testing and the remaining part is used for training.

The proposed classification approach is seen in Figure-1. According to this figure, the data set is first divided into two separate groups. The training set is passed through the stages symbolized by blue boxes. First, pre-processing is applied to the data set. In this pre-processing stage, the mean of each attribute of the input data is brought closer to zero and the standard deviation is brought closer to one. This process is of great importance for many classifiers to work efficiently. The parameters obtained from the normalization process are recorded to normalize the test data. In the second stage, feature selection or reduction is performed. The parameters obtained from this process are also recorded. Finally, the parameters of the classifier are set with the normalized features, processed features and relevant labels. Similar to the previous steps, the parameters are recorded and the classifier is made available for diagnosis of diabetes

In this study, support vector machines (SVM), decision trees (DT), Naive Bayes (NB) and k-nearest neighbor (KNN) algorithms are used to perform classification. The unique features and advantages of each algorithm are evaluated in accordance with the nature of the data set and included in the classification process.

Support vector machines (SVM) is a classification method that is especially effective on high-dimensional data sets. The reasons for using SVM in this study include its ability to adapt to the complexity of the data set and its tendency to perform well overall.

Decision trees (DT) offer a user-friendly approach to explaining and interpreting classification problems. This algorithm is valuable for understanding relationships in the data set and transparently visualizing its decisions.

Naive Bayes (NB) is a probability-based classification method that is especially successful in applications such as text mining. This algorithm works by underestimating the dependencies between attributes in the data set, and with this feature it can be especially effective in large data sets.

K-nearest neighbor (KNN) is an example-based classification method. This algorithm performs classification by bringing similar examples together. The flexibility of KNN can be especially advantageous in identifying structural features in the data set and classifying on an example basis.

The classification process using these four different algorithms was carried out in accordance with the characteristics and structures of the data set, and a more effective classifier was designed by comparing the performance of these algorithms. In this way, a significant contribution to future similar studies may be made by better understanding the internal dynamics of the data set.

2.1. Data Source

The main data source in this study is a dataset predicting the risk of early-stage diabetes, and this dataset (*the early-stage diabetes risk prediction dataset)*) was collected at Sylhet Diabetes Hospital in Sylhet, Bangladesh [4], [5]. Data were recorded through surveys taken directly from patients [4], [5]. This survey contains comprehensive information about patients' health status, lifestyle and genetic history.

This dataset may provide important information for improving methodologies used in early-stage diabetes-related risk prediction and monitoring patients more effectively. Analysis of such data sets can support important steps towards improving public health by providing a valuable contribution to applications in the healthcare sector.

Most of the questions in the data set have categorical features. This indicates that information in different areas, such as patients' demographic characteristics, genetic predispositions and lifestyles, is grouped into certain categories. Sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital fungus, blurred vision, itching, irritability, delayed healing, partial paralysis, cramps, alopecia areata and obesity are all binary options. Based on these questions, it is possible to diagnose early diabetes.

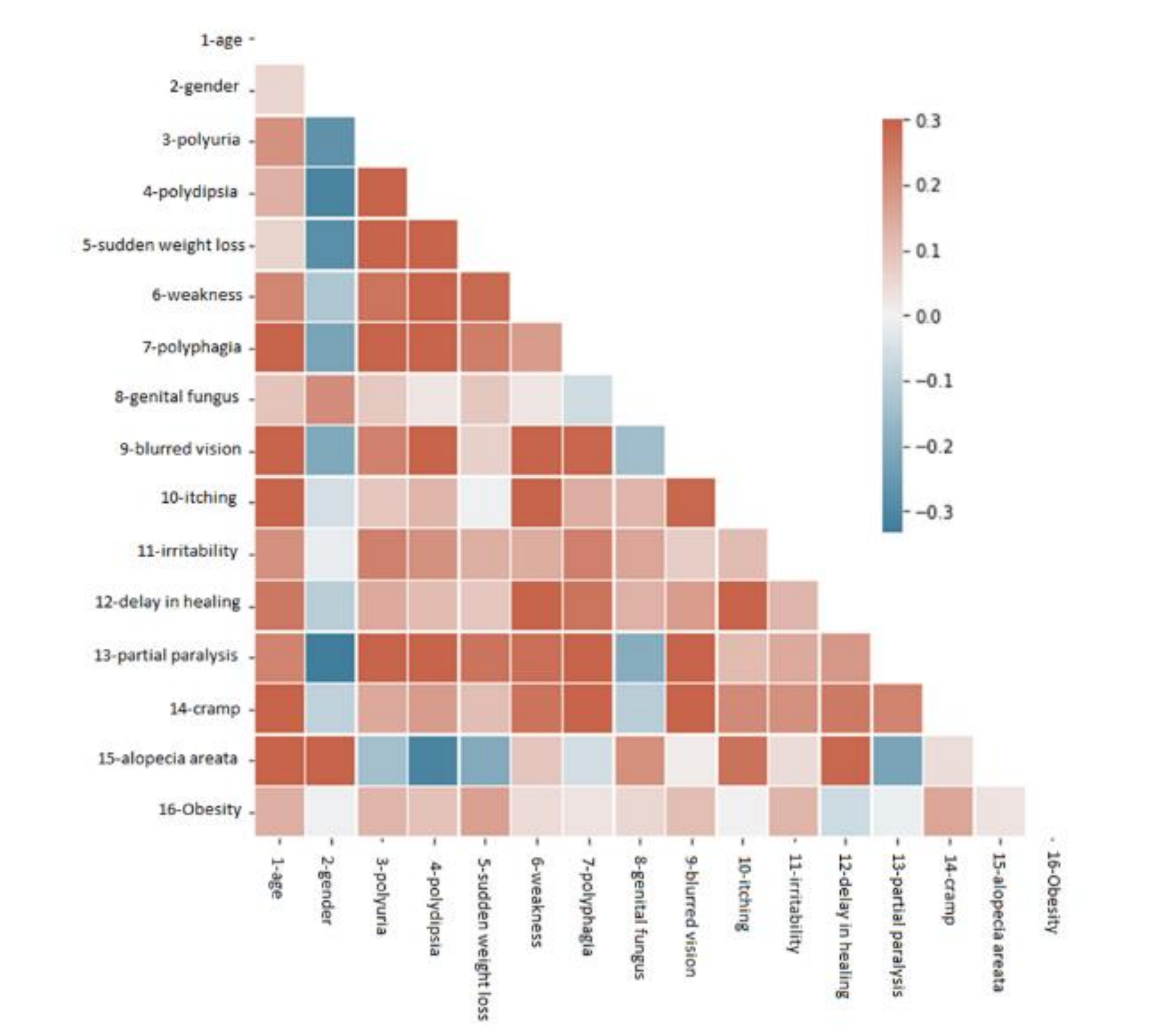

Figure 2 Correlation matrix

3. Experimental results and discussion

The relationships between each attribute in the data are given in Figure 2. In general, it appears that there is no significant correlation between the attributes. This shows that the features identified for early diabetes diagnosis are well selected. However, their effects on classifiers need to be examined in more depth.

Interpretation of experimental results is examined in two main parts. The first of these, the PCA method, briefly referred to as principal component analysis, which is a feature reduction technique, is applied to the data set.

Table 1 Performance of classifiers for feature reduction

	DT	SVM	NB	KNN
1-Feature	85.85±3.03	78.21±2.87	78.08±2.65	90.12±2.38
8- Feature	92.9±2.1	89.86±2.12	87.26±2.51	95.4±2.15
16- Feature	93.74±2.07	91.5±1.82	88.6±3.11	96.62±1.24

For this purpose, the data set was divided into two: 70% training and 30% testing. First of all, the training set is normalized. The parameters used for normalization were recorded. The normalization process is effective on the performance of many classifiers, except some classifiers such as DT. After normalization, the features are reduced with the help of the PCA algorithm. The reduced features and their associated labels are used to train the classifier. The parameters of the trained classifier are also recorded. Then, in the training phase, testing is carried out with the optimum parameters obtained in the normalization, PCA and classification stages. For testing, data that has not been experienced by the model is used. In other words, 30% of the data set other than the training set is used. The effects of feature reduction have been observed with SVM, DT, NB and KNN classifiers. The training and testing process was repeated 50 times after the data was randomly sorted.

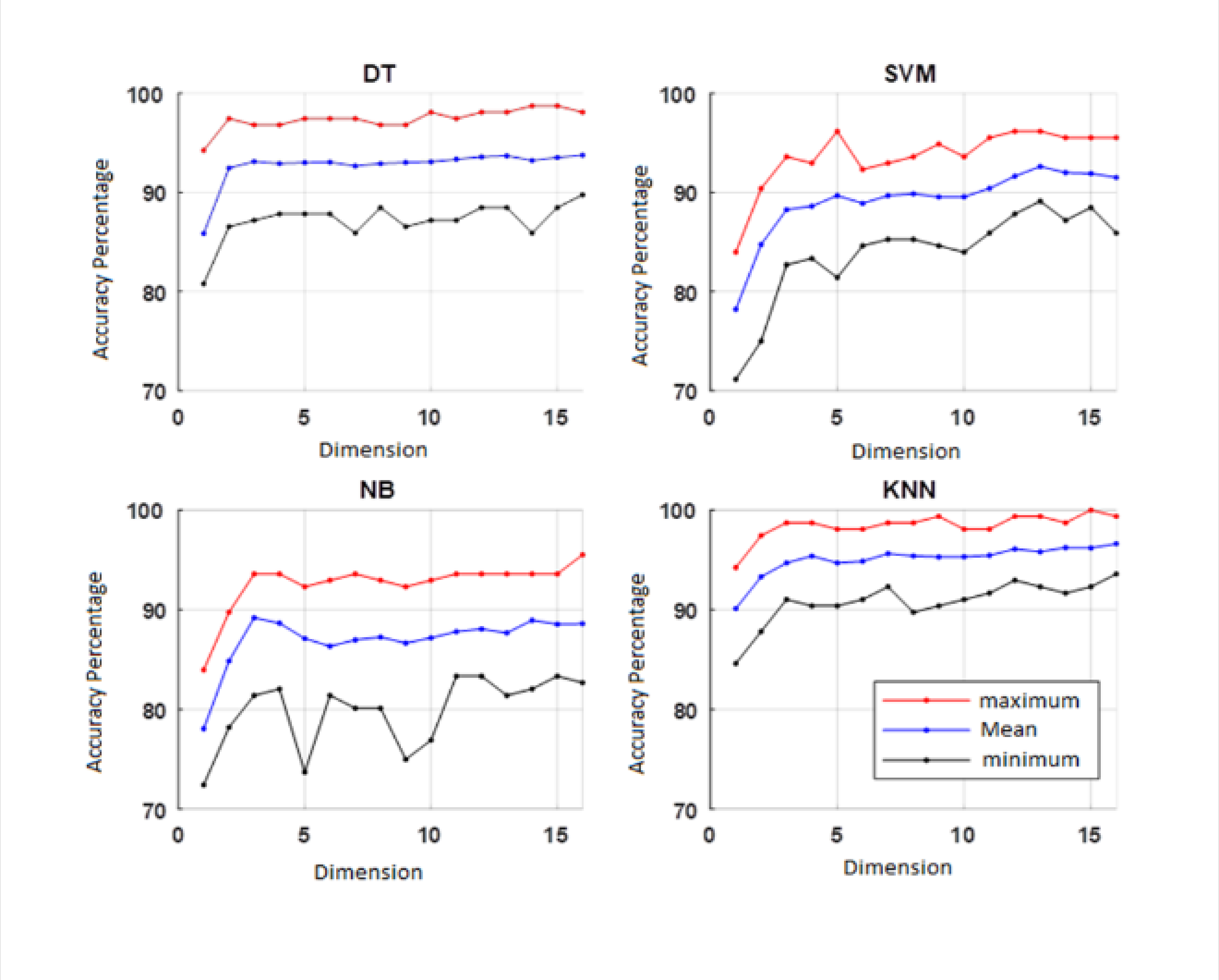

Figure 3 Performance of Classifiers

The average, maximum and minimum values of these results are plotted in Figure 3. According to the figure, it can be seen that good results are obtained as a result of the classification process generally made with 16 features. The results obtained from this figure are summarized in Table 1. Accordingly, it appears that KNN is the most successful classifier.

In the second part of the study, feature selection was emphasized instead of feature reduction technique. This process has tried all attribute combinations between 1 and 8. The normalization and training procedures performed for PCA above are also valid in this section. Therefore, this section specifically focuses on how feature selection is made. For this process, feature groups ranging from 1 to 8 were taken from the 16-element set we had. The performances of each of these groups were measured with the accuracy values obtained with the KNN algorithm.

Table 2 Feature Selection Effects

S*	**Features**													**DY****
	1	**2**	**3**	**4**	**5**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
1				4										75.58
2		2	3											81.99
3	1	2	3											86.67
4	1	2	3						11					91.28
5	1	2	3						11				15	94.29
6	1	2	3	4			9						15	96.41
7	1	2	3	4			9			12		14		97.051
8	1	2	3	4	5					12		14		97.115
9	1	2	3	4		8		10	11				15	97.179
10	1	2	3	4		8		10				14	15	97.244
11	1	2	3	4			9	10	11				15	97.051
12	1	2	3	4			9		11	12		14		97.115
13	1	2	3	4			9		11	12			15	97.372
14	1	2	3	4			9			12	13	14		97.436
15	1	2	3	4			9			12	13		15	97.051
16	1	2	3	4				10		12				97.051
17	1	2	3	4					11	12	13		15	97.179
18	1	2	3	4					11	12		14	15	97.500
S*: shows simulations repeated 10 times.														
DY**: Average accuracy percentage of 10 runs														
1-age, 2-gender, 3-polyuria, 4-polydipsia, 5-sudden weight loss, 6-weakness, 7-polyphagia, 8-genital fungus, 9-blurred vision, 10-itching, 11-irritability, 12-delay in healing, 13-partial paralysis, 14-cramp, 15-alopecia areata														

When Table 2 is examined, it is clearly seen that polydipsia, that is, the need to consume a lot of water, alone has a serious impact on the diagnosis of early diabetes. Just by looking at this attribute, it can be said that there is a 75% probability that the person has diabetes. Apart from this, polyuria also has serious effects with age. In this regard, the attributes of age, gender, polyuria, polydipsia, irritability, delay in recovery, cramps and alopecia areata enable the diagnosis of early diabetes to be predicted with a high accuracy of 97.5%.

4. Conclusion

In this study, feature reduction techniques and feature selection activities, which play an important role in early diabetes diagnosis, are discussed. The analyzes clearly showed that extremely successful classification results were achieved with effective feature selection and appropriate classifier selection. Feature selection is critical to determining what information should be obtained from the patient. This process helps identify the most important markers to diagnose diabetes, which optimizes the diagnostic process and provides healthcare professionals with more focused and effective information. On the other hand, the use of feature reduction techniques can increase the generalization ability of the model by reducing the complexity in the data set. This is an important factor in achieving more reliable and sensitive results in the diagnosis of diabetes. This research emphasizes the importance of early diagnosis and shows that feature selection and reduction techniques can be successfully applied in the early diagnosis of diabetes. These approaches can both improve patients' quality of life and provide cost-effective solutions to healthcare systems.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Diabetes. Accessed: Jan. 10, 2024. [Online]. Available: https://www.who.int/westernpacific/health-topics/diabetes

[2] A. Mujumdar and V. Vaidehi, Diabetes Prediction using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/j.procs.2020.01.047.

[3] "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques | IEEE Journals & Magazine | IEEE Xplore. Accessed: Jan. 24, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8585030

[4] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis, M. Gupta, D. Konar, S. Bhattacharyya, and S. Biswas, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2020, pp. 113–125. doi: 10.1007/978-981-13-8798-2_12.

[5] "UCI Machine Learning Repository." Accessed: Feb. 19, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[6] A. Caliskan, C. O'Brien, K. Panduru, J. Walsh, and D. Riordan, "An Efficient Siamese Network and Transfer Learning-Based Predictive Maintenance System for More Sustainable Manufacturing," Sustainability, vol. 15, no. 12, p. 9272, 2023.

[7] A. Caliskan, D. Riordan, and J. Walsh, "The Biomonitoring of Ireland's River Network Using a 1D Convolution Neural Network," in 2023 IEEE International Symposium on Technology and Society (ISTAS), Sep. 2023, pp. 1–4. doi: 10.1109/ISTAS57930.2023.10305993.

[8] M. Ringnér, "What is principal component analysis?," Nature Biotechnology, vol. 26, no. 3, Art. no. 3, Mar. 2008, doi: 10.1038/nbt0308-303.

[9] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," Neurocomputing, vol. 55, no. 1, pp. 169–186, Sep. 2003, doi: 10.1016/S0925-2312(03)00431-4.

[10] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660–674, May 1991, doi: 10.1109/21.97458.

[11] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 4, pp. 580–585, Jul. 1985, doi: 10.1109/TSMC.1985.6313426.

[12] L. Jiang, H. Zhang, and Z. Cai, "A Novel Bayes Model: Hidden Naive Bayes," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1361–1371, Oct. 2009, doi: 10.1109/TKDE.2008.234.