ORIGINAL PAPER
ORYGINALNY ARTYKUŁ NAUKOWY

Health Prob Civil. 2022, Volume 16, Issue 2
ISSN 2353-6942

PREDICTION OF TYPE 2 DIABETES MELLITUS USING FEATURE SELECTION-BASED MACHINE LEARNING ALGORITHMS

PRZEWIDYWANIE CUKRZYCY TYPU 2 Z WYKORZYSTANIEM ALGORYTMÓW UCZENIA MASZYNOWEGO OPARTYCH NA SELEKCJI CECH

Atınç Yılmaz[1(A,B,C,D,E,F)]

[1] Department of Computer Engineering, Beykent University, Istanbul, Turkey

Authors' contribution
Wkład autorów:
A. Study design/planning
zaplanowanie badań
B. Data collection/entry
zebranie danych
C. Data analysis/statistics
dane – analiza i statystyki
D. Data interpretation
interpretacja danych
E. Preparation of manuscript
przygotowanie artykułu
F. Literature analysis/search
wyszukiwanie i analiza literatury
G. Funds collection
zebranie funduszy

Summary

Background. The aim of this study is to develop and evaluate a machine learning model for the early diagnosis of type 2 diabetes to allow for treatments to be applied in the early stages of the disease.

Material and methods. A proposed hybrid machine learning model was developed and applied to the *Early-stage diabetes risk prediction dataset* from the UCI database. The prediction success of the proposed model was compared with other machine learning models. Pearson's correlation and SelectKBest feature selection methods were employed to examine the relationships between the dataset input parameters and the results.

Results. Of the 520 patients included in the dataset, 320 were diagnosed with diabetes and 328 (63.08%) were males. The most commonly observed diabetes diagnosis criterion was obesity (n=482, 83.08%). While the strongest feature detected with Pearson's correlation was polyuria, the strongest feature detected with SelectKBest was polydipsia. With Pearson's feature extraction, the most successful machine learning method was the proposed hybrid method, with an accuracy of 97.28%. Using SelectKBest feature selection, the same model was able to predict type 2 diabetes with accuracy of 95.16%.

Conclusions. Early detection of type 2 diabetes will allow for a prompter and more effective treatment of the patient. Thus, use of the proposed model may help to improve the quality of patient care and lower the number of deaths caused by this disease.

Keywords: feature selection, health information system, type 2 diabetes, machine learning, nursing care

Streszczenie

Wprowadzenie. Celem niniejszego badania jest opracowanie i ewaluacja modelu uczenia maszynowego umożliwiającego wczesną diagnozę cukrzycy typu 2, która pozwala na podjęcie leczenia na początkowym etapie choroby.

Materiał i metody. Zaproponowany hybrydowy model uczenia maszynowego został przygotowany i zastosowany z wykorzystaniem zbioru danych *Early-stage diabetes risk prediction dataset* pochodzącego z bazy UCI. Proponowany model porównano z innymi modelami uczenia maszynowego pod względem skuteczności przewidywania. Aby zbadać związek pomiędzy parametrami wejściowymi zbioru danych a wynikami, zastosowano metodę korelacji Pearsona oraz metodę selekcji cech SelectKBest.

Wyniki. Spośród 520 przypadków uwzględnionych w zbiorze danych, 320 miało rozpoznaną cukrzycę, a 328 z nich (63,08%) to mężczyźni. Najczęstszym kryterium rozpoznania cukrzycy była otyłość (n=482, 83,08%). Podczas gdy najsilniejszą cechą wykrytą metodą Pearsona była poliuria, najsilniejszą cechą wykrytą metodą SelectKBest okazała się polidypsja. W przypadku ekstrakcji cech Pearsona najskuteczniejszą metodą uczenia maszynowego była zaproponowana metoda hybrydowa, której dokładność wynosi 97,28%. Ten sam model był w stanie przewidzieć zachorowanie na cukrzycę typu 2 z dokładnością 95,16% za pomocą selekcji cech SelectKBest.

Wnioski. Wczesne wykrycie cukrzycy typu 2 pozwoli na szybsze i skuteczniejsze leczenie pacjenta. Dlatego też zaproponowany model może pomóc w podniesieniu jakości opieki nad pacjentami, a także w obniżeniu liczby zgonów spowodowanych tą chorobą.

Słowa kluczowe: selekcja cech, system informacji zdrowotnej, cukrzyca typu 2, uczenie maszynowe, opieka pielęgniarska

Tables: 6
Figures: 5
References: 37
Submitted: 2021 Dec 9
Accepted: 2022 March 14

Yılmaz A. Prediction of type 2 diabetes mellitus using feature selection-based machine learning algorithms. Health Prob Civil. 2022; 16(2): 128-139. https://doi.org/10.5114/hpc.2022.114541

Address for correspondence / Adres korespondencyjny: Atınç Yılmaz, Department of Computer Engineering, Beykent University, Hadım Koruyolu Cd. No. 19, Sarıyer, 34398 Istanbul, Turkey, e-mail: atincyilmaz@beykent.edu.tr, phone: +904441997, ORCID: https://orcid.org/0000-0003-0038-7519

Copyright: © John Paul II University of Applied Sciences in Biala Podlaska, Atınç Yılmaz. This is an Open Access journal, all articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License (http://creativecommons.org/licenses/by-nc-sa/4.0/), allowing third parties to copy and redistribute the material in any medium or format and to remix, transform, and build upon the material, provided the original work is properly cited and states its license.

Introduction

Diabetes is a chronic, broad-spectrum metabolic disorder in which proteins, fats and carbohydrates cannot be used sufficiently by the organism due to insulin deficiency or defect in its effects [1]. Due to modern living conditions, a sedentary lifestyle and obesity have caused this disease to increase gradually worldwide. According to the International Diabetes Federation 2019 report, 1 out of 11 adults has diabetes (463 million), and 1 out of 2 adults (212 million) with diabetes has not yet been diagnosed [2]. It is predicted that there will be 579 million diabetics in 2030 and 700 million in 2045. According to the report, in 2045, Turkey is expected to be among the top 10 countries in the world for the number of diabetic patients. The annual number of deaths due to diabetes, which causes serious morbidity as well as mortality, is 4.2 million [3].

The predictions outlined above for type 1 and type 2 diabetes include individuals aged 20 to 79 years. Approximately 90-95% of diabetes patients consist of type-2 diabetes patients. Type 2 diabetes usually occurs after the age of 40 years, and its incidence increases with advancing age. Recently, obesity and sedentary lifestyle changes have been reported to increase the incidence of type 2 diabetes in children and adolescents. The probability of the occurrence of diabetes can be calculated according to the type 2 diabetes risk factors present (Table 1). However, it cannot be determined when individuals with a high risk score will develop the disease. Type 2 diabetes can be asymptomatic for a long time. Therefore, macrovascular and microvascular complications are typically seen at the time of diagnosis. Diabetes can cause life-threatening complications and seriously affect the heart, blood vessels, eyes, kidneys, nerves and teeth. Diabetes also increases susceptibility to infection [4]. It is very important to detect diabetes at an early stage in order to reduce the risk of acute complications and to avoid long-term chronic complications (retinal, renal, neural, cardiac and vascular), which can lead to extremely serious consequences that will be with the person for life.

Table 1. Risk factors for type 2 diabetes [5]

Question	Yes	No
The body weight is equal or more than the table given on the side	5	0
The age is under 65 and do little or no exercise throughout the day	5	0
The age is between 46 and 64 years old	5	0
The age is over 65	9	0
The sex is female (if she has given birth to a child over 4 kg)	1	0
Relative has diabetes	1	0
Brother/sister has diabetes	1	0
Total points (3-9: low risk, 10+: high risk)		

It is generally accepted that real-time and continuous follow-up of type 2 diabetic patients by doctors provides significant gains in effectively dealing with the disease, raising awareness, and making the patient the protector of their own health by developing positive behaviors. At present, a new generation technologies has been developed that allow for the continuous follow-up and control of these patients by a doctor [6]. Type 2 diabetic patients are diagnosed by classic symptoms (frequent urination, excessive thirst, dry mouth), a fasting blood glucose ≥126 mg/dl, a blood glucose ≥200 mg/dl at the 2nd hour of oral glucose tolerance test, or an HbA1c≥6.5%. It is believed that these technologies, which can monitor and instantly analyze blood glucose, and be used to plan appropriate insulin therapy, diet and individual exercise programs, can significantly contribute to the treatment of type 2 diabetes [7]. In this study, the monitoring of patients at risk of type 2 diabetes with machine learning methods is performed, and the importance of the issue is highlighted.

The artificial intelligence (AI) was originated with the ideas of Turing's "Can Machines Think" and McCarthy's "Intelligent Machine Engineering" [8]. It is a discipline that aims to make machines that have the ability to reason, to benefit from past information, plan, learn, communicate, perceive, and move objects. In today's technology, AI has an important role in the creation of devices that can reason. The idea of thinking machines is aimed to make machines behave rationally like humans. For this reason, it has been proposed to model many rational behaviors in nature, especially the human brain. Machine learning has ensured that the human brain's way of thinking, decision-making, and learning are modeled and imitated by machines [9].

With machine learning methods, systems that imitate certain human behaviors and simulate the human reasoning process related to a specific area of expertise can be created. Medical experts are among the most frequent users of AI, and algorithms have been developed to solve structural questions and provide answers

in this field. Medical expert systems are developed in line with the recommendations of one or more medical experts. Thus, correct results are obtained by considering the most appropriate questions. The purpose of health decision support systems is to make recommendations and suggestions to the doctor based on past experience, rather than making decisions without a doctor. The main areas of interest are to create AI programs that can perform clinical diagnoses and provide recommendations for treatment.

Various studies have used AI approaches to predict or diagnose type 2 diabetes with varying degrees of success. For example, Tigga et al. [10] used different machine learning methods to predict this disease. Logistic regression, k-nearest neighbors (KNN), support vector machine (SVM), Naïve Bayes, decision tree and random forest methods were modeled over this dataset, and the random forest method had the highest accuracy. Choi et al. [11] used diabetes data obtained from the Korea University Guro Hospital and different machine learning methods for diabetes prediction, and the logistic regression method provided higher accuracy than the other methods. Perveen et al. [12] modeled metabolic syndrome and diabetes risk using the Naïve Bayes, C4.5 decision tree, and J48 machine learning methods. In this study, the Naïve Bayes method produced the highest accuracy rate. Mujumdar et al. [13] also applied machine learning methods for diabetes prediction. In this study, decision tree, KNN, Naïve Bayes, SVM, AdaBoost, random forest, perceptron sensor, linear discriminant analysis (LDA), bagging, gradient boost, and logistic regression methods were applied. The highest accuracy (94%) was obtained with the LDA method and the lowest accuracy (60%) was obtained with the SVM method. Sisodia et al. [14] developed a diabetes prediction model over Pima Indians Diabetes Database (PIDD) using Naïve Bayes, decision tree and SVM methods, and the Naïve Bayes method provided the highest accuracy with 76%. Zou et al. [15] also created a diabetes prediction model using random forest, J48 and artificial neural network (ANN) methods. Swapna et al. [16] detected diabetes by using the deep learning method over electrocardiogram (ECG) signals. In this study, a convolutional neural network (CNN) and long-short-term memory deep learning methods were used, and feature extraction was achieved with SVM. The proposed model had an accuracy rate of 95%. Wu et al. [17] also applied data mining methods to predict type 2 diabetes. KNN and logistic regression methods were used in this study, and the highest accuracy was 95%. Naiarun et al. [18] compared Naïve Bayes, ANN, decision tree and logistic regression methods for the diagnosis of diabetes, and the highest accuracy rate obtained was 86%. Rahman et al. [19] also diagnosed diabetes by using multilayer perceptron, Bayes classifier, J48, and JRIP methods. The highest accuracy rate in the study was obtained with the J48 method. Meng et al. [20] determined the risk of diabetes by using logistic regression, ANN, and decision tree methods. Twelve input parameters were used for the model, and the highest accuracy was obtained by using the decision tree method. Huang et al. [21] performed diabetes classification by using C4.5 decision tree and Naïve Bayes methods, and it was revealed that the patient's age, duration of diagnosis, insulin need, and diet control parameters were the most important parameters for blood glucose control. Iyer et al. [22] analyzed diabetes using Naïve Bayes and decision tree methods and demonstrated the applicability of both methods. Patil et al. [23] put forward a hybrid prediction model consisting of the k-means and C4.5 methods, and the results obtained from the clustering algorithm were used in the classification algorithm.

This study aims to improve the prediction success and time process of the previous models implemented with the proposed model. Today, with the increasing use of technology, more data is available to solve complex problems. As outlined above, machine learning methods can be used to predict patients at high risk of developing type 2 diabetes. Early prediction of this disease can allow medical practitioners to intervene much sooner and potentially avoid serious, long-term consequences for patients. The current research is designed as a methodological comparison between various machine learning methods used to predict type 2 diabetes. The hypothesis of the study is that the proposed hybrid machine learning model is more successful than other machine learning methods and previous diagnostic techniques in predicting type 2 diabetes.

Based on this information, making risk assessments for the prevention of type 2 diabetes, planning preventive care by determining the individuals at risk, and systematic management of the disease in the early period will be possible with the machine learning model developed here. With this machine learning model, it is expected that the prevalence of type 2 diabetes will decrease and contribute to the national economy by reducing health care expenditures.

Material and methods

Dataset

The dataset used in this study is the *Early-stage diabetes risk prediction dataset* from the University of California Irvine (UCI) database [24]. There are 520 patients in this dataset, of which 328 (63.08%) are male and 192 (36.92%) are female. The class label of the dataset indicates whether it belongs to the diabetes class. There are 16 input parameters that qualify the class tag. The input parameters are gender, age, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, irritability, itching, delayed healing, partial paresis, muscle stiffness, alopecia and obesity. Of the patients in the dataset, 320 (61.54%) were diagnosed with diabetes and 200 (38.46%) were healthy.

Machine learning algorithms

The Naïve Bayes method is a probability-based machine learning method used in classification problems, which is based on the Bayes theorem. This method, as an algorithm, tries to determine the relationship between the desired goal and the independent variables. As a result of learning, the model enables a calculation of how many times each output occurs in the learning set and identifies these calculated values as priority probabilities [25].

The logistic regression method is similar to the linear regression method in many ways. The most substantial difference between them is what the methods are used for. While linear regression is used for prediction problems, logistic regression is used for classification problems. The aim of logistic regression is to find the most appropriate model to define the relationship between independent variables [26].

Decision tree is a method based on the classification of a large amount of data into subgroups. It renders the dataset of the classifier more understandable, and shows the class options and probability-dependent states in the form of a tree by sorting them in order. Decision tree methods based on entropy and using regression are available [27].

The random forest method is based on the principle of classifying the sample (in the dataset) using more than one classifier. A random forest consists of multiple decision trees, combining them to get a more accurate and consistent prediction. One of the strongest advantages of the random forest is that it can be applied to both classification and regression problems [28].

SVMs are classification methods based on statistical learning. These methods are based on separating the data from two classes in the most appropriate way. There are two types of SVMs: linear and non-linear [29].

The KNN algorithm makes use of data from dataset classes which are known. The distance of the new data without class label information to all labeled data is calculated. Then, the nearest neighbor is checked for the value K [30].

XGBoost is a decision tree-based method that is an optimized version of the gradient boosting method that provides better performance. The important features of the algorithm are that it can obtain a high prediction accuracy rate, prevent overfitting, handle null data, and it has a high throughput speed [31].

Proposed hybrid model

The proposed hybrid model has been used for the Ozone Level Prediction problem [32]. In order to demonstrate the applicability of the hybrid model in different areas and types of problems, the proposed model has also been applied for type-2 diabetes prediction. A two-layer classification is built into the proposed hybrid model. In the first classification layer, the relationship between the input parameters is put forth by the feature selection method and fragmentation is performed on the dataset (Figure 1). First, the 16 input parameters used for type 2 diabetes diagnosis are divided into 4 different groups by the feature selection method. Then, in the first classification layer, each piece separately produces results in the genetic algorithm model. The results obtained from the four different genetic algorithm models are used as input to the XGBoost model, which is the second classification layer. The value calculated using the XGBoost method is the result of the system. The proposed model aims to use the learning advantages of the genetic algorithm method and the strengths of the XGBoost method together.

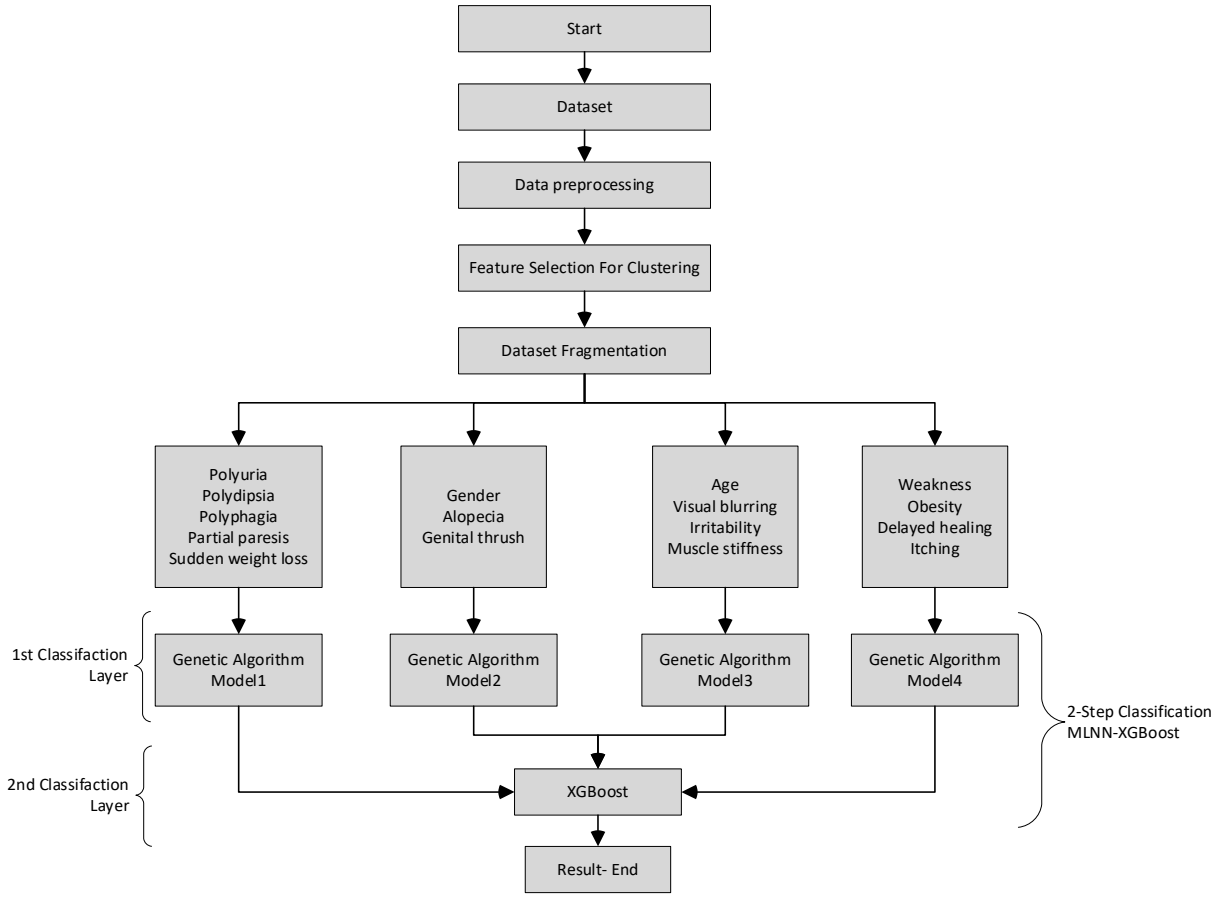

Figure 1. The working principles of the proposed model

Experimental studies

In the first stage of the study, type 2 diabetes is predicted using the proposed hybrid machine learning model (Figure 2). Next, in order to measure the success of the proposed hybrid method, the results are compared with other machine learning methods. For this reason, after pre-processing of the data is completed, the 16 input parameters (age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity) are correlated with the diagnosis result. After calculating the effects of all parameters on the diagnosis result, the 10 parameters that had the greatest impact on the result were selected and classification was made with other machine learning methods (especially the XGBoost and SVM methods).

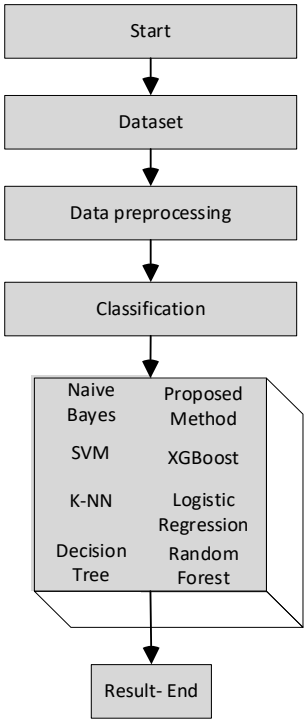

Figure 2. Predicting type 2 diabetes with machine learning

Notes: KNN = k-nearest neighbors; SVM = support vector machine.

The data analysis was performed using the open-source Python programming language and the Jupyter Notebook application. With this programming language, the features of all data in the dataset are extracted and an AI model is created. Pearson's correlation and the SelectKBest feature selection methods were used to determine the effects of dataset input parameters on the result. Pearson's correlation method was used to determine the relationships between continuous and class features. A Pearson's correlation is a number between -1 and 1 that indicates the degree to which the two variables are linearly related. Pearson's correlations are suitable only for metric variables. A value closer to 0 means a weaker correlation (exact 0 indicates no correlation). A value closer to 1 means a stronger positive correlation and a value closer to -1 means stronger negative correlation. The SelectKBest feature selection method scores features uses only one function, and then uses the parameters with the highest scoring properties as "k" for classification.

Results

For the dataset used in the study, the effect of each input parameter on diabetes was focused on. Of the 520 patients in the dataset, 320 were diagnosed with diabetes and 200 did not have diabetes. While the average age of the patients diagnosed with diabetes was 49.07 years, the average age of the non-diabetic individuals was 46.36 years. In addition, 328 (63.08%) patients were male and 192 (36.92%) were female (Figure 3).

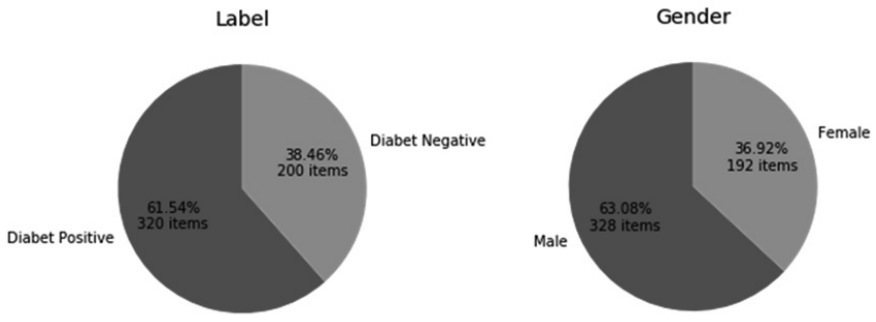

Figure 3. Demographic features of the dataset

The relationships of the 16 input parameters specified as diabetes diagnosis criteria with diabetes status were examined. In the patients included in the dataset, the following characteristic were observed: alopecia (n=341, 68%), genital thrush (n=404, 77.69%), itching (n=267, 52.35%), delayed healing (n=281, 54.04%), muscle stiffness (n=325, 62.50%), irritability (n=394, 75.77%), polyphagia (n=283, 54.42%), weakness (n=305, 58.65%), polydipsia (n=287, 55.19%), obesity (n=482, 83.08%), partial paresis (n=296, 56.92%), sudden weight loss (n=303, 58.27%), visual blurring (n=287, 55.19%) and polyuria (n=262, 50.38%) (Figure 4).

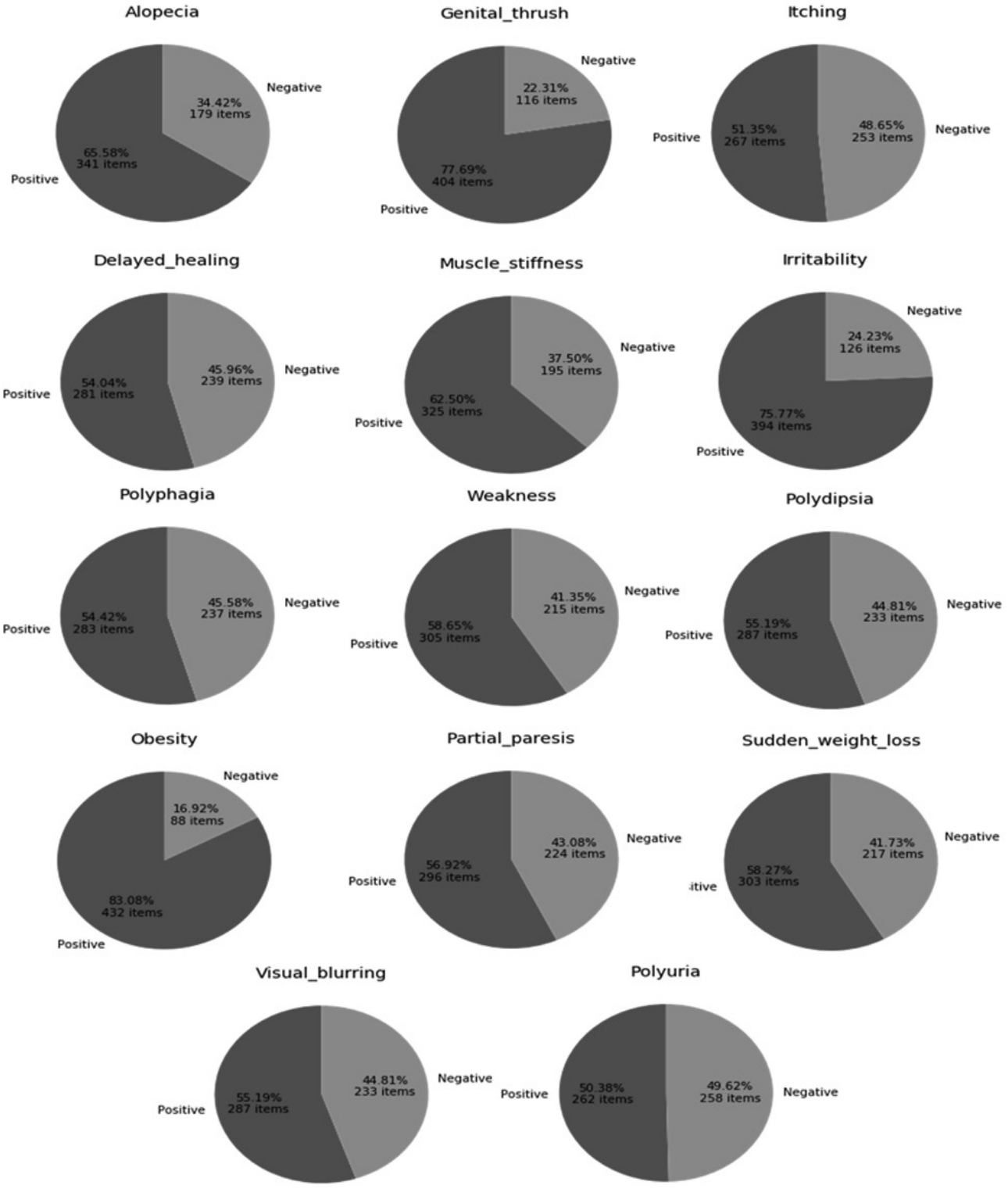

Figure 4. The incidence of diabetes diagnostic criteria in the dataset

Pearson's correlation method was applied to determine the relationships between the dataset input parameters and the result. According to the analysis, the feature most strongly related to the diagnosis of diabetes was polyuria (r=0.66) (Table 2, Figure 5).

Table 2. Type 2 diabetes diagnosis criteria and significance levels with the applied methods

Pearson's Correlation Coefficient		SelectKBest Method	
Parameters	**Correlation Coefficient (r)**	**Parameters**	**SelectKBest Scores**
Polyuria	0.665	Polydipsia	120.785515
Polydipsia	0.648	Polyuria	116.184593
Sudden weight loss	0.436	Sudden weight loss	57.749309
Partial paresis	0.432	Partial paresis	55.314286
Polyphagia	0.342	Gender	38.747637
Irritability	0.299	Irritability	35.334127
Visual blurring	0.251	Polyphagia	33.198418
Weakness	0.243	Alopecia	24.402793
Muscle stiffness	0.122	Age	18.845767
Genital thrush	0.110	Visual blurring	18.124571
Age	0.108	Weakness	12.724262
Obesity	0.072	Genital thrush	4.914009
Delayed healing	0.046	Muscle stiffness	4.875000
Itching	-0.013	Obesity	2.250284
Alopecia	-0.267	Delayed healing	0.620188

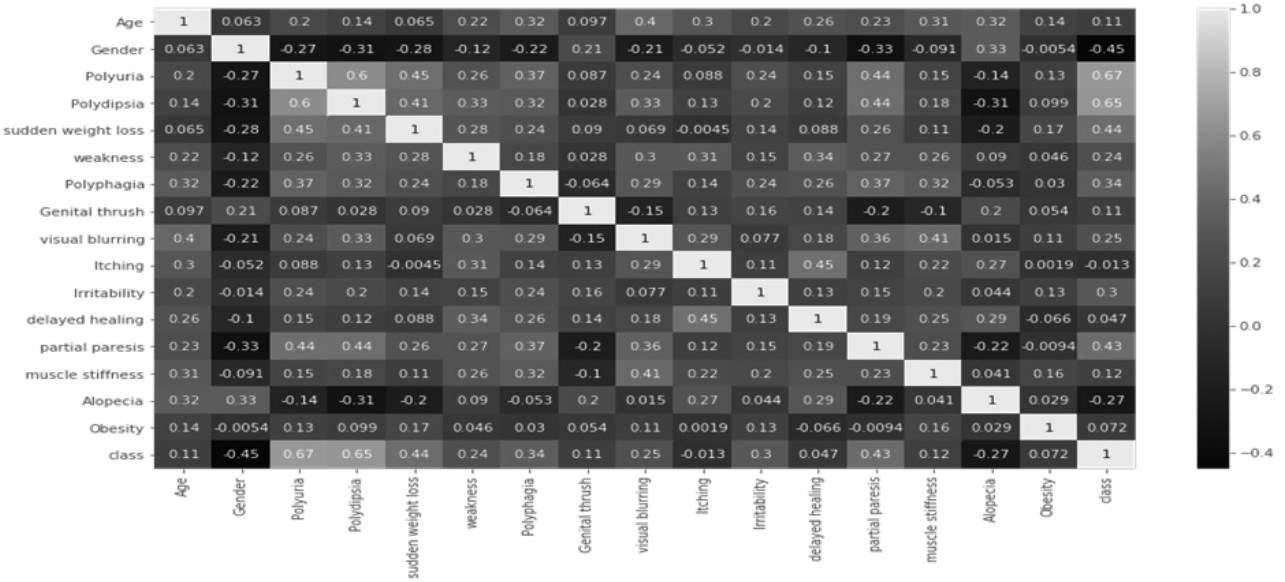

Figure 5. Type 2 diabetes diagnosis criteria and significance levels with Pearson's correlation method

SelectKBest, one of the other feature extraction methods, was applied to select the features that have the strongest relationships with the output. The SelectKBest class scores features use only one function, and then removes all features except the "k" top scoring features. Feature selection with the SelectKBest method is shown in Table 2.

To apply machine learning methods to the diabetes diagnosis problem, 65% of the dataset was used in the training and the remaining 35% was used in the test set. The k-fold cross validation method was used to select the optimum training and test sets. For the k-fold cross validation method, values of 5, 10 and 15 were applied, and

the most effective results were obtained with k=5. Next, using the Pearson's correlation and SelectKBest feature extraction methods, the top 10 features that were related to the diagnosis were selected, and machine learning methods were applied. First, the dataset with the 16 input parameters was used with the machine learning algorithms. However, the statistical results were less successful in terms of accuracy and time. Thus, instead of using all 16 features, temporal performance and accuracy were improved by only choosing the features that had the largest impacts on the result. The reason for choosing 10 features is that the effect of the input parameters on the result decreases after the 10th feature. While the strongest feature identified using the correlation analysis was polyuria, polydipsia was the strongest feature identified using the SelectKBest method. Table 3 shows the feature extraction methods and the criteria used as a basis for the diagnosis of type 2 diabetes.

Table 3. Comparison of the type 2 diabetes diagnosis criteria used with the two feature selection methods

Significance Level	Pearson's Method	SelectKBest Method
1	Polyuria	Polydipsia
2	Polydipsia	Polyuria
3	Sudden weight loss	Sudden weight loss
4	Partial paresis	Partial paresis
5	Polyphagia	Gender
6	Irritability	Irritability
7	Visual blurring	Polyphagia
8	Weakness	Alopecia
9	Muscle stiffness	Age
10	Genital thrush	Visual blurring

The proposed hybrid model was the most successful method when the prediction success of the machine learning methods was measured with Pearson's correlation coefficient. This model was able to predict type 2 diabetes with an accuracy of 97.28% (Table 4).

Table 4. Prediction success of the machine learning methods as assessed using Pearson's correlation coefficient

Method	Accuracy (%)	F_1-Score	R^2	MAE
Naïve Bayes	89.56	0.90	0.89	0.26
Logistic Regression	89.01	0.89	0.88	0.32
Decision Tree	93.12	0.95	0.92	0.18
Random Forest	95.15	0.96	0.94	0.11
SVM	94.05	0.95	0.93	0.15
KNN	91.85	0.93	0.92	0.21
XGBoost	95.604	0.96	0.95	0.08
ANN	94.23	0.95	0.94	0.13
Proposed Method	97.28	0.97	0.96	0.05

Notes: SVM = support vector machine; KNN = k-nearest neighbors; XGBoost = eXtreme gradient boosting; ANN = artificial neural network.

The proposed hybrid method was the most successful method when the prediction success of the machine learning methods was measured with SelectKBest feature extraction. This model was able to predict type 2 diabetes at an accuracy of 95.16% (Table 5).

Table 5. Prediction success of the machine learning methods as assessed using SelectKBest feature extraction

Method	Accuracy (%)	F_1-Score	R^2	MAE
Naïve Bayes	84.61	0.85	0.84	0.52
Logistic Regression	87.58	0.88	0.86	0.46
Decision Tree	90.83	0.90	0.90	0.32
Random Forest	91.12	0.93	0.90	0.24
SVM	88.46	0.92	0.87	0.38
KNN	90.76	0.90	0.90	0.33
XGBoost	93.604	0.95	0.92	0.11
ANN	92.11	0.93	0.91	0.16
Proposed Method	95.16	0.96	0.94	0.09

Notes: SVM = support vector machine; KNN = k-nearest neighbors; XGBoost = eXtreme gradient boosting; ANN = artificial neural network.

A comparison of the current results with the works of other researchers who used the same diabetes dataset is shown in Table 6.

Table 6. Comparison of the current results with similar studies

Study	Accuracy (%)	Method
Ferdousi et al. [33]	94	RF
Wijayaningrum et al. [34]	95.38	MLP
Tan et al. [35]	96.79	CNN
Alehegn et al. [36]	90.36	Proposed Ensemble Method
Fazakis et al. [37]	88.8	Voting (LR-RF)
Proposed Method	97.23	Proposed Method

Notes: RF = random forest; MLP = multi-layer perceptron; CNN = convolutional neural network, LR-RF = logistic regression-random forest.

Discussion

Type 2 diabetes has been diagnosed using different methods from past to present. However, over the last couple of years, there has been a transition from traditional diagnostic methods to technology-friendly diagnostic methods. Diagnosing diabetes with AI, which is the state-of-the-art technology, creates an opportunity to identify individuals at a high risk of developing diabetes at the earliest stage. The results of the current study suggest that machine learning methods using AI are more successful than traditional diagnostic methods in predicting type 2 diabetes. With the proposed hybrid model, it was possible to detect type 2 diabetes with a 97.28% accuracy as assessed by correlation analysis, and at a 95.16% accuracy as assessed by the SelectKBest method.

As machine learning models used in the field of health care do not make human-induced observation errors, they tend to provide better results compared to other methods. The machine learning hybrid model developed here was trained with 520 samples and was able to accurately detect type 2 diabetes with a 97% accuracy. The early detection of type 2 diabetes will allow for earlier and more effective treatments to be applied to patients. Early interventions may lower the morbidity and mortality caused by this disease.

Conclusions

The contributions of the current study are as follows:

1) a novel hybrid model has been proposed based on genetic algorithm and XGBoost;
2) the development of a new model with high performance values and a demonstration of the applicability of the model;
3) machine learning methods to predict type 2 diabetes mellitus.

It is thought that detecting type 2 diabetes before its appearance or in the early period of its occurrence by using AI will minimize the complications of this disease. As a result, health care expenditures will also decrease; thus, having a positive impact the economy.

Disclosures and acknowledgements

The author declares no conflicts of interest with respect to the research, authorship, and/or publication of this article. The research was funded by the author.

References:

1. Baran Ö, Türker PF, Tayfur M. [Assessment of nutritional status, food addiction, and awareness of individuals with type 2 diabetes]. Journal of Başkent University Faculty of Health Sciences-BÜSBID. 2020; 33(3): 226-242 (in Turkish).
2. Atış Ş, Önder A. [Classification and treatment management in new diagnosed diabetes mellitus patients]. Hitit Medical Journal. 2020; 2(3): 262-226 (in Turkish).
3. International Diabetes Federation. IDF Diabetes Atlas 2019 [Internet]. Brussels: IDF [cited 2021 Feb 01]. Available from: https://idf.org
4. Swapna G, Soman K, Vinayakumar R. Diabetes detection using ecg signals: an overview. Deep Learning Techniques for Biomedical and Health Informatics. 2020; 14: 299-327. https://doi.org/10.1007/978-3-030-33966-1_14
5. Nisar N, Khan IA, Qadri MH, Sher SA. Knowledge and risk assessment of diabetes mellitus at primary care level: a preventive approach required combating the disease in a developing country. Pakistan Journal of Medical Sciences. 2008; 24(5): 667-672.
6. Shi Yin B. The importance and strategy of diabetes prevention. Chronic Diseases and Translational Medicine. 2016; 2(4): 204-207. https://doi.org/10.1016/j.cdtm.2016.11.013
7. Eswari T, Sampath P, Lavanya S. Predictive methodology for diabetic data analysis in big data. Procedia Computer Science. 2016; 50: 203-208. https://doi.org/10.1016/j.procs.2015.04.069
8. McCarthy J. What is artificial intelligence? [Internet]. Stanford: Stanford University; 2007 [cited 2021 March 12]. Available: http://www-formal.stanford.edu/jmc/whatisai/
9. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatology. 2018; 154(11): 155-160. https://doi.org/10.1001/jamadermatol.2018.2348
10. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science. 2020; 167: 706-716. https://doi.org/10.1016/j.procs.2020.03.336
11. Choi BG, Rha SW, Kim SW, Kang JH, Park JY, Noh YK. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei Medical Journal. 2019; 60(2): 191. https://doi.org/10.3349/ymj.2019.60.2.191
12. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques. IEEE Access. 2018; 7: 1365-1375. https://doi.org/10.1109/ACCESS.2018.2884249
13. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019; 165: 292-299. https://doi.org/10.1016/j.procs.2020.01.047
14. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Science. 2018; 132: 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122
15. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics. 2015; 9: 515.
16. Swapna G, Kp S, Vinayakumar R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science. 2018; 132: 1253-1262. https://doi.org/10.1016/j.procs.2018.05.041
17. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018; 10: 100-107. https://doi.org/10.1016/j.imu.2017.12.006
18. Naiarun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science. 2015; 69: 132-142. https://doi.org/10.1016/j.procs.2015.10.014
19. Rahman RM, Afroz F. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. Journal of Software Engineering and Applications. 2013; 6(3): 85-97. https://doi.org/10.4236/jsea.2013.63013

20. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences. 2013; 29(2): 93-99. https://doi.org/10.1016/j.kjms.2012.08.016
21. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artificial Intelligence in Medicine. 2007; 41(3): 251-262.
22. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP). 2015; 5(1): 1-14. https://doi.org/10.5121/ijdkp.2015.5101
23. Patil B, Joshi R, Toshniwal D. Association rule for classification of type-2 diabetic patients. Proceedings of the 2nd International Conference on Machine Learning and Computing; 2010 Jan; Bangalore, India. Piscataway Township: IEEE; 2010. p. 330-334.
24. UCI Machine Learning Repository. Early stage diabetes risk prediction dataset [Internet]. Irvine: University of California Irvine [cited 2021 March 1]. Available from: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset
25. Pham BT, Prakash I. A novel hybrid model of Bagging-based Naive Bayes Trees for landslide susceptibility assessment. Bulletin of Engineering Geology and the Environment. 2019; 78: 1991-1925.
26. Kuha J, Mills C. On group comparisons with logistic regression models. Sociological Methods & Research. 2020; 49(2): 498-525. https://doi.org/10.1177/0049124117747306
27. Li MJ, Xu HH, Deng Y. Evidential decision tree based on belief entropy. Entropy. 2019; 21(9): 897. https://doi.org/10.3390/e21090897
28. Athey S, Tibshirani J, Wager S. Generalized random forests. Annals of Statistics. 2019; 47(2): 1148-1178. https://doi.org/10.1214/18-AOS1709
29. Wang MJ, Chen HL. Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. Applied Soft Computing. 2020; 88: 105946. https://doi.org/10.1016/j.asoc.2019.105946
30. Wang CZ, Sh YP, Fan XD, Shao MW. Attribute reduction based on K nearest neighborhood rough sets. International Journal of Approximate Reasoning. 2019; 106: 18-31. https://doi.org/10.1016/j.ijar.2018.12.013
31. Nguyen H, Bu XN, Bui H, Cuong DT. Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study. Acta Geophysica. 2019; 67(2): 477-490. https://doi.org/10.1007/s11600-019-00268-4
32. Yılmaz A. Ozone level prediction with machine learning algorithms. Journal of Aeronautics and Space Technologies. 2021; 14(2): 177-183.
33. Ferdousi R, Hossain MA, Saddik AE. Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access. 2021; 9: 96823-96837. https://doi.org/10.1109/ACCESS.2021.3094063
34. Wijayaningrum VN, Saragih TH, Putriwijaya NN. Optimal multi-layer perceptron parameters for early stage diabetes risk prediction. IOP Conference Series: Materials Science and Engineering. 2021; 1073(1): 012070. https://doi.org/10.1088/1757-899X/1073/1/012070
35. Tan Y, Chen H, Zhang J, Tang, Liu P. Early risk prediction of diabetes based on GA-Stacking. Applied Sciences. 2022; 12: 632. https://doi.org/10.3390/app12020632
36. Alehegn M, Joshi R, Mulay P. Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 2018; 118(9): 871-878.
37. Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, Moustakas K. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access. 2021; 9: 103737-103757. https://doi.org/10.1109/ACCESS.2021.3098691