Journal of Associated Medical Sciences 2024; 57(3): 155-165

JAMS
Journal of Associated Medical Sciences

Scopus Indexed & Thai-Journal Citation Index Centre (TCI)

Journal of Associated Medical Sciences

Journal homepage: https://www.tci-thaijo.org/index.php/bulletinAMS/index

Exploring machine learning approaches for early diabetes risk prediction: A comprehensive examination of health indicators and models

Nihar Ranjan Panda[1*] Jatindra Nath Mohanty[2] Ruchi Bhuyan[1] Prasanta Kumar Raut[3] Manulata[4]

[1]Department of Medical Research, IMS & SUM Hospital, SOA Deemed to be university, Bhubaneswar, Odisha, India.
[2]School of Basic and Applied Science, Centurian University of Technology & Management, Bhubaneswar, Odisha, India.
[3]Trident academy of technology, Bhubaneswar, Odisha, India.
[4]All India Institute of Medical Science, Bibinagar, India.

ARTICLE INFO

Article history:
Received 31 May 2024
Accepted as revised 21 July 2024
Available online 23 July 2024

Keywords:
Health indicators, diabetes, classification, decision tree, random forest.

ABSTRACT

Background: The increased prevalence of morbidity and mortality associated with Type 2 diabetes is due to changing lifestyles, demanding improved disease management measures. To tackle this, scientists are increasingly looking to technological advances, notably machine learning, for illness prevention and management, particularly in non-communicable diseases. The emphasis is on establishing an early detection system to identify Type 2 diabetes risk factors, enabling prompt treatments and preventative steps to reduce the disease's rising prevalence.

Materials and methods: The research aimed to assess the association of diabetes class with health indicators. Five machine learning models were employed with cross-validation techniques to predict early diabetes risk. The performance matrices of the models were evaluated and compared with the existing work.

Results: In multivariate analysis, we found polyuria (β=3.492; Aor=32.872; 95% CI=11.09,97.35; p<0.001), polydipsia (β=-4.100; Aor=60.378; 95%CI=18.28,199.37; p<0.001), polyphagia (β=1.181; Aor=3.25; 95%CI=1.23,8.57; p=0.017), genital thrush (β=1.08; Aor=2.96; 95%CI=1.26,7.53; p=0.023), irritability (β=2.28; Aor=9.82; 95%CI=3.41,28.26; p<0.001), and partial paresis (β=1.2406; Aor=3.45; 95% CI=1.35,8.79; p=0.009) are the potential health risk indicators for positive diabetes class.

Conclusion: Using an interpretable feature learning approach for early diabetes prediction improves the use of global health data. This method forecasts hazards correctly and gives insights into influential aspects. As a result, a more proactive healthcare strategy is implemented, allowing for more prompt treatments and encouraging a more hopeful future by improving patient outcomes and lowering the total burden of diabetes on individuals and healthcare systems.

Introduction

Diabetes mellitus is a chronic condition that affects people of all ages. It can cause blindness, renal failure, amputation, heart failure, and stroke, among other serious effects.[1] After we eat, our bodies convert calories into glucose, and the insulin produced by the pancreas nettles the cells to absorb glucose. The pancreas would not produce enough insulin, or the body would not respond efficiently to the insulin produced, leading to increased glucose levels in the patient. The World Health Organization says that diabetic disease was the ninth most significant cause of death on the globe in 2019 and was number six in upper-middle-income countries. Diabetes is

∗ **Corresponding contributor.**
Author's Address: *Department of Medical Research, IMS & SUM Hospital, SOA Deemed to be university, Bhubaneswar, Odisha, India.*
E-mail address: niharpanda1994@gmail.com
doi: 10.12982/JAMS.2024.057
E-ISSN: 2539-6056

classified into three types: T1, T2, and GDM. Presence of Type 1 diabetes arises when the pancreas does not do its job correctly in producing enough insulin, and it frequently strikes kids, and teenagers.[2] Common symptoms include Excessive thirst, dry mouth, weight loss, impaired eyesight, and frequent urination. Individuals with Type 1 diabetes are more likely to develop heart disease. Type 2 diabetes is distinguished by cells that fail to react effectively to insulin, resulting in insulin resistance.[3-5] Type 2 diabetes accounts for approximately 90% of all diabetes cases worldwide. While less severe than Type 1, it can cause health problems, particularly in the kidneys, nerves, and eyes. Type 2 diabetes, which was previously only observed in adults, is increasingly impacting youngsters. Gestational diabetes, the third primary form, occurs in pregnant women who have no history of diabetes and causes blood sugar levels to rise.[6,7] It usually goes away after delivery, but up to 10% of cases might evolve into Type 2 diabetes later in life. Babies born to moms who have gestational diabetes face problems. Diabetes prevalence is highest in China and India.

Related work

Isfafuzzaman *et al.* used a semi-supervised model using extreme gradient boosting (XGBoost) to predict the insulin characteristics of the private dataset.[8] SMOTE and ADASYN techniques were used to address the class imbalance issue. Machine learning classification methods such as decision tree, SVM, Random Forest, Logistic Regression, KNN, and different ensemble approaches were employed to establish the best prediction outcomes. With 81% accuracy, 0.81 F1 coefficients, and an AUC of 0.84, the XGBoost classifier with the ADASYN technique produced the best results.B. Shamreen*et et al*. tested various machine learning classifiers for predicting Type 2 diabetes mellitus, including logistic regression, XGBoost, gradient boosting, decision trees, ExtraTrees, random forest, and Light Gradient Boosting Machine (LGBM). LGBM had the best accuracy among these classifiers, reaching 95.20% and outperforming the other methods.[9] Aishwarya *et al.* For improved diabetes categorization, research presents a diabetes prediction model that adds external parameters and regular criteria such as glucose, BMI, age, insulin, etc. When compared to the previous dataset, the new dataset improves classification accuracy. A pipeline model is also enforced to boost classification accuracy for diabetes prediction. The research describes using clustering, especially K-means clustering, on the dataset to categorize each patient as diabetes or non-diabetic. Clustering was performed using highly associated characteristics, such as glucose and age. Sandip *et al*. provided a diabetes prediction model that uses a variety of machine learning algorithms, including Logistic Regression, SVM, Nave Bayes, Random Forest, XGBoost, LightGBM, CatBoost, Adaboost, and Bagging.[10] CatBoost is the most successful ensemble approach tested, attaining a high accuracy rate of 95.4%. CatBoost also surpasses XGBoost, with a better AUC-ROC score of 0.99 vs. XGBoost's accuracy rate of 94.3% and AUC-ROC score of 0.98. Quan *et al.* used decision trees, random forests, and neural network approaches to predict diabetes mellitus, with random forests attaining the most remarkable accuracy (ACC=0.8084) when all characteristics were applied.[11] The research also utilized mRMR to pick characteristics and discovered that the first five variables (height, HDL, fasting glucose, breath, and LDL) predicted diabetes well using the Luzhou dataset. The first three characteristics (glucose, 2-hour serum insulin, and age) were chosen for the Pima Indians dataset.

The research aimed to evaluate and compare the effectiveness of five machine-learning models in predicting diabetes using lifestyle data from the NHANES database. The following are the study's primary conclusions and information:

1. **Comparison of model**
   - The researchers examined the predictive abilities of CATBoost, XGBoost, RF, LR, and SVM.
   - In terms of predictive performance, CATBoost beat the other models.
   - The models were most likely assessed using sensitivity, accuracy, precision, AUC, and ROC measures.
2. **Performance of CATboost**
   - CATBoost produced an AUC (Area Under the Curve) of 0.83 and an accuracy of 82.1%.
   - According to these criteria, CATBoost displayed a high degree of accuracy and discrimination in discriminating between those with and without diabetes.
3. **Contributing factors**
   - The study found that calorie, carbohydrate, and fat intake levels were the most important predictors of diabetes patients.
   - This means that the machine-learning algorithms, notably CATBoost, used nutritional data to produce accurate predictions.

Materials and methods

This work is based on the ''Early-Stage Diabetes Risk Prediction'' dataset from the University of California, Irvine (UCI) machine learning repository.[29] This dataset contains details of 520 people who report diabetes-related symptoms. It gives us information about people, including features that can cause diabetes to develop. This section describes the method of identifying diabetes using a machine learning technique. The prediction algorithm is initially built using a publicly accessible diabetes dataset. EDA is used to find significant patterns among parameters in discriminating between diabetes and non-diabetic patients. Basic preprocessing activities, such as addressing missing values, numerical representation of categorical data, and normalization, were carried out. After that, the dataset is partitioned into training and test sets, with the training set further subdivided into training and validation subsets. Feature selection minimizes dimensionality, using insights from EDA and an additional trees classifier to find key features. Several machine learning techniques are used to build the prediction model, which is then validated on the validation set. Various measures are used to evaluate performance, and model parameters are fine-tuned to

improve efficacy. The resulting model is then applied to the test set for additional assessment, concluding the development and evaluation of a machine learning-based diabetes detection system. The detailed roadmap of the work is shown in Figure 1. A logistic regression model was used to find potential risk factors for diabetes.[31]

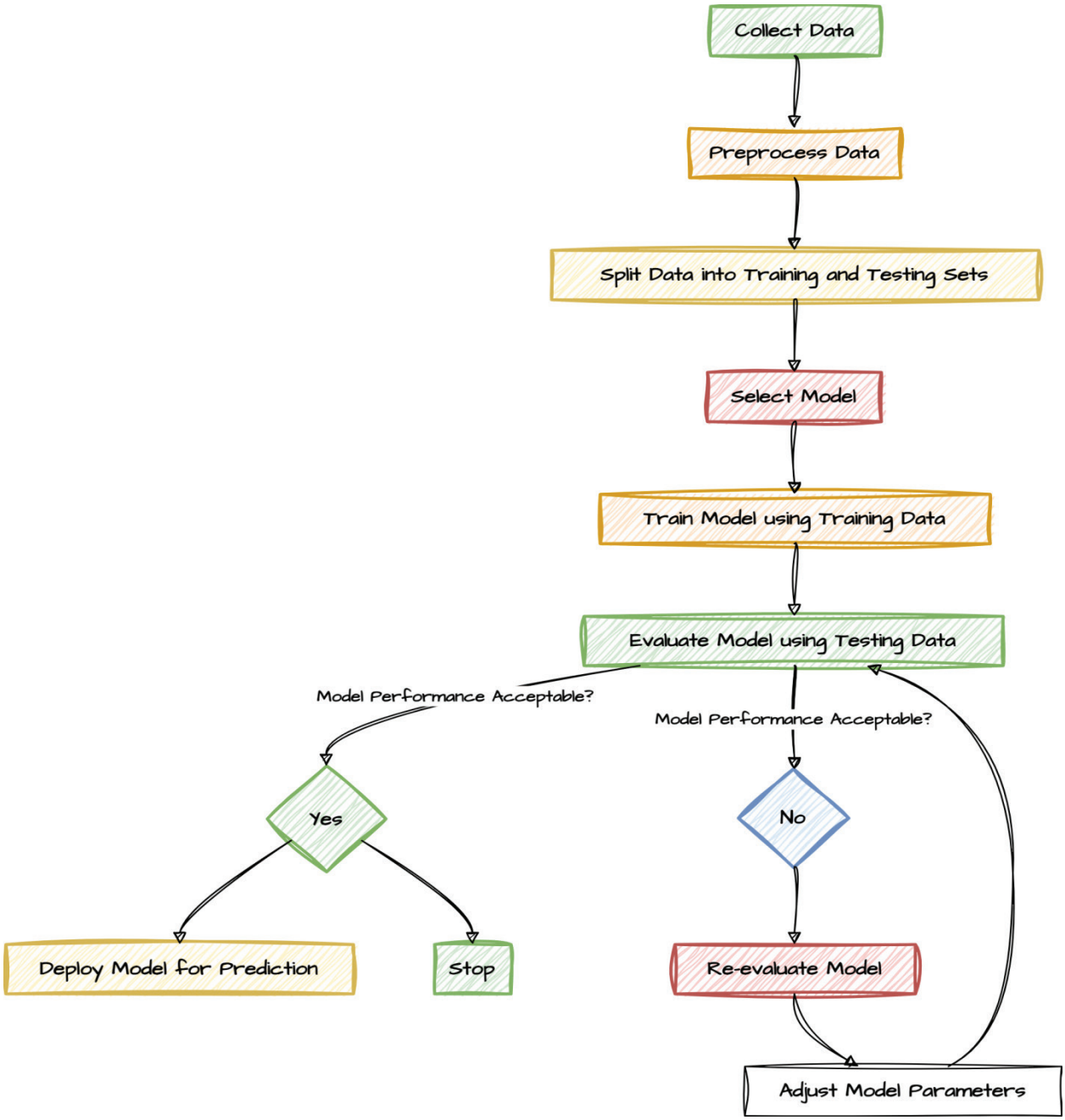

***Figure 1**. A roadmap for building a classification model for predicting type 2 diabetes.*

Statistical analysis and machine learning models

The data were expressed in frequency (%) for categorical variables and mean±SD for continuous variables. A chi-square test was applied to find the association between diabetes class and other parameters. For the mean difference between the two classes, we performed the test. The significance level was 0.05. After conducting an association between diabetes class and other parameters, we used the multivariate logistic regression method to find the most significant risk factors. The factors found to be substantial in the Univariate analysis were run in the final multivariate analysis. To predict early diabetes risk, we used five machine learning models. All the models were tested by their performance matrices.

Data processing and Eda

The effectiveness of a machine learning model is dependent on precise data preparation. It is critical for best model performance to analyze the dataset for redundancy, missing, or irrelevant data. The primary goal is to fine-tune the dataset for training and testing. Using

visualizations, such as charts, may help you understand how attributes affect the target variable. This astute data-driven strategy improves the model's capacity to generalize and uncover important patterns, resulting in robust prediction outcomes in cases such as diabetes diagnosis. The association between diabetes class and some health indicators is shown in Figure 2.

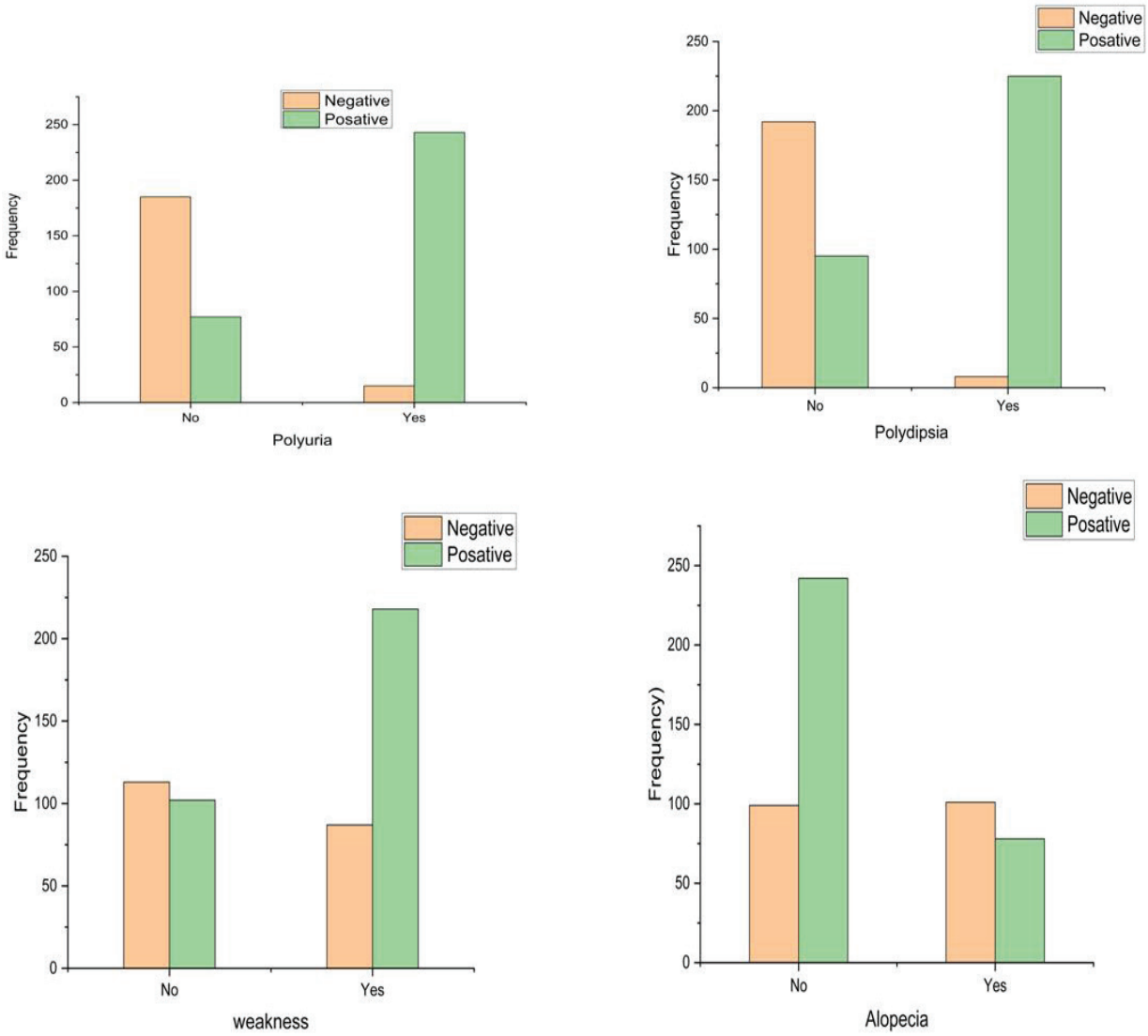

***Figure 2**. Association between diabetes classes with some health indicators.*

Boosting classification

Combining numerous weak learners into a robust ensemble model improves prediction performance.[12] Five boosting methods, including the Gradient Boosting Machine (GBM), were used in this study. GBM combines predictions from numerous decision trees to create powerful learners. For optimum splits, each tree's nodes employ unique feature subsets. Sequential trees rectify prior faults, constantly enhancing the model. GBM performance is further enhanced by hyperparameter optimization, resulting in a considerable improvement. GBM can generate accurate predictions thanks to this iterative learning process that leverages the collective wisdom of numerous trees, making it a valuable tool in machine learning applications.

Decision tree

A Decision Tree is a supervised learning technique with a tree-like structure that starts at the root node and branches out.[13] It makes judgments based on input features and is mainly used for categorization. The tree's leaf nodes indicate outcomes, while inside nodes include dataset attributes and branching decision criteria. The information gain, calculated for each characteristic at each node, directs feature selection, optimizing predicted accuracy. Grid Search CV and Randomized Search CV are used in hyperparameter optimization. The minimum leaf sample size (10), maximum depth (6), and the criteria ('gin) are all fine-tuned. This careful optimization procedure refines the structure of the Decision Tree, ensuring that it efficiently captures patterns and correlations within the data, boosting its prediction skills in a supervised learning setting.[14]

Random forest

Random Forest is an ensemble learning approach that builds a 'forest' of decision trees, each of which has been trained using the Bootstrap Aggregating (bagging) methodology.[15] The primary objective is to improve individual Decision Tree performance by minimizing variation and enhancing model unpredictability. Random Forest reduces overfitting and provides a more robust, generalizable model by developing many trees and using only random subsets of features at each split. Instead of focusing on the most crucial feature during node division, the method chooses the best feature from a random subset, boosting tree variety. Configuring parameters such as min_samples_split, min_samples_leaf, max_depth, max_features, n_estimators, and Bootstrap is part of the offered hyperparameter tuning. It is critical to fine-tune these hyperparameters to optimize the model's performance. However, the cost of computing is evident since the Randomized Search CV technique, used for hyperparameter tuning, takes a long time to execute due to its broad parameter space search. Regardless of the computational requirement, Random Forest is a robust and frequently used machine learning technique that boosts forecast accuracy and manages complicated datasets via its ensemble of varied decision trees.[16]

KNN classification

K-Nearest Neighbours (KNN) is a lazy learning algorithm that differs from eager learners, such as Random Forest, because it does not have a dedicated training phase.[17] KNN is built on the notion of 'feature similarity,' it works by categorizing incoming data points based on their closeness to existing training data. A Minkowski metric was used to measure distances during the model-building process, and a leaf size of 20 was chosen. The best accuracy was obtained after extensive parameter optimization using the Euclidean metric and setting the number of neighbors to four. This means that during the classification step, the algorithm identifies the class of a new data point in the feature space by considering the classes of its four nearest neighbors. KNN's adaptability makes it useful for various applications.[30]

Support vector classification

The Support Vector Machine (SVM) is a popular supervised machine learning technique representing each data point in n dimensions.[18] The data characteristics function as coordinates in this space in this method. SVM performs classification by determining which hyperplane best divides the two classes. A margin, defined as the distance between the decision border and the nearest points of each class, is determined to identify the best hyperplane. SVM chooses the hyperplane with the most significant margin. In other circumstances, however, precise class prediction precedes maximizing the margin. The selection of hyperparameters is critical for the overall accuracy of SVM since they have a major influence on its performance. Randomized Search CV, a method that effectively searches the parameter space, was used to optimize the hyperparameter.[19]

Results

Table 1 describes the association between diabetes class and health indicators. The table shows that the positive diabetes class in females is higher (54.1%) than in males (45.9%). Which is statistically significant (p<0.001). The number of polyuria cases is relatively higher, 243 (75.9%), in the positive diabetes class, compared to the negative diabetes class, 15 (7.5%). We found that polydipsia is strongly associated with diabetes class (p<0.001). It can be observed that the polydipsia cases are much higher, 225 (70.3%) as compared to negative diabetes class (4%). Similarly, sudden weight loss, weakness, polyphagia, visual blurring, irritability, partial paresis, and alopecia were found to be highly significant (p<0.001) with diabetes class. The proportion of all these symptoms is found to be higher in positive diabetes classes. However, delayed healing(p=0.284), obesity, and itching (p=0.760) were not associated with the diabetes class. The mean age of the positive diabetes class was 49.1±12.1, and the negative diabetes class was 46.4±12.1 (p=0.013).

Table 1. Features description of the data set.

Features	Levels	Negative (N=200)	Positive (N=320)	*p*
Sex	Female	19 (9.5)	173 (54.1)	<0.001
	Male	181 (90.5)	147 (45.9)	
Polyuria	No	185 (92.5)	77 (24.1)	<0.001
	Yes	15 (7.5)	243 (75.9)	
Polydipsia	No	192 (96)	95 (29.7)	<0.001
	Yes	8 (4)	225 (70.3)	
Sudden weight loss	No	171 (85.5)	132 (41.3)	<0.001
	Yes	29 (14.5)	188 (58.8)	
Weakness	No	113 (56.5)	102 (31.9)	<0.001
	Yes	87 (43.5)	218 (68.1)	
Polyphagia	No	152 (76.0)	131 (40.9)	<0.001
	Yes	48 (24.0)	189 (59.1)	
Genital thrush	No	167 (83.5)	237 (74.1)	0.012
	Yes	33 (16.5)	83 (25.9)	
Visual blurring	No	142 (71)	145 (45.3)	<0.001
	Yes	58 (29)	175 (54.7)	
Itching	No	101 (50.5)	166 (51.9)	0.760
	Yes	99 (49.5)	154 (48.9)	
Irritability	No	184 (92)	210 (65.6)	<0.001
	Yes	16 (8)	110 (34.4)	
Delayed healing	No	114 (57)	167 (52.2)	0.284
	Yes	86 (43)	153 (47.8)	
Partial paresis	No	168 (84)	128 (40)	<0.001
	Yes	32 (16)	192 (60)	
Muscle stiffness	No	140 (70)	185 (57.83)	0.005
	Yes	60 (30)	135 (42.2)	
Alopecia	No	99 (49.5)	242 (75.6)	<0.001
	Yes	101 (51.5)	78 (24.4)	
Obesity	No	173 (86.5)	259 (80.9)	1.000
	Yes	27 (13.5)	61 (19.1)	
Age		46.4±12.1	49.1±12.1	0.013

Table 2 describes the multivariate analysis among diabetes class and health indicators. We removed itching, delayed healing, and obesity from the analysis since they were found to be nonsignificant in the Univariate analysis. In multivariate analysis, we found polyuria (β=3.492; Aor=32.872; 95%CI=11.09,97.35; p<0.001), polydipsia (β=-4.100; Aor=60.378; 95%CI=18.28,199.37; p<0.001), polyphagia (β=1.181; Aor=3.25; 95%CI=1.23,8.57; p=0.017), genital thrush (β=1.08; Aor=2.96; 95%CI=1.26,7.53; p=0.023), irritability (β=2.28; Aor=9.82; 95%CI=3.41,28.26; p<0.001), and partial paresis (β=1.2406; Aor=3.45; 95%CI=1.35,8.79; p=0.009) are the potential health risk indicators for positive diabetes class. Apart from these, we found gender and age are also associated (p<0.05) in positive diabetes class in multivariate analysis.

Table 2 Multivariate analysis for potential risk health indicators.

	Model coefficients - class					95% Confidence interval	
Predictor	**Estimate**	**SE**	**Z**	***p***	**Odds ratio**	**Lower**	**Upper**
Intercept	2.4927	0.8838	2.8205	0.005	12.0939	2.13936	68.3679
Gender							
Male- Female	-3.7962	0.5170	-7.3433	< 0.001	0.0225	0.00815	0.0619
Polyuria							
Yes-No	3.4926	0.5540	6.3048	**<0 .001**	**32.8720**	**11.09928**	97.3547
Polydipsia							
Yes-No	4.1006	0.6095	6.7281	**<0 .001**	**60.3782**	**18.28469**	199.3756
Sudden weight loss							
Yes-No	0.5206	0.5023	1.0365	0.300	1.6831	0.62885	4.5048
Weakness							
Yes-No	0.0438	0.4723	0.0928	0.926	1.0448	0.41400	2.6367
Polyphagia							
Yes-No	1.1814	0.4939	2.3920	**0.017**	**3.2588**	**1.23783**	8.5794
Genital thrush							
Yes-No	1.0855	0.4766	2.2779	**0.023**	**2.9610**	**1.16357**	7.5350
Visual blurring							
Yes-No	-0.1766	0.5283	0.3344	0.738	0.8381	0.29760	2.3602
Irritability							
Yes-No	2.2851	0.5392	4.2383	**<0.001**	**9.8265**	**3.41564**	28.2699
Partial paresis							
Yes-No	1.2406	0.4763	2.6046	**0.009**	**3.4578**	**1.35943**	8.7953
Muscle stiffness							
Yes-No	-0.6180	0.4992	1.2378	0.216	0.5390	0.20261	1.4341
Alopecia							
Yes-No	-0.5610	0.5044	1.1123	0.266	0.5706	0.21234	1.5335
Age	-0.0552	0.0207	2.6597	0.008	0.9463	0.90863	0.9856

Table 3 describes the classification performances of various machine learning techniques using the data set. The k-fold cross-validation technique was used for all the models. In machine learning, greater k values in cross-validation frequently result in higher accuracy but can also lead to overfitting. Leave-one-out cross-validation is appropriate for small datasets (often less than 100 occurrences) to maximize data utilization, but it can be computationally costly. Holdout validation is commonly recommended for large datasets since it reduces training time. However, this essay seeks to refute that assertion. Despite the increased time investment, it contends that adopting k-fold cross-validation for massive datasets yields significant accuracy benefits. The purpose is to show through results that the time cost of employing k-fold validation over holdout validation is justified, especially when the value of k is kept low enough to provide adequate classification quality. Among all the models, KNN performed the best accuracy (0.962), with an AUC(0.969) and F1 score (0.962). The Roc curves of all the machine learning models are shown in Figure 3.

Table 3. Performance matrices for classification model to predict diabetes class.

Evaluation Metrics	Boosting	DT	KNN	RF	SVM
Accuracy	0.923	0.920	0.962	0.923	0.913
Precision (Positive predictive value)	0.932	0.923	0.965	0.927	0.913
Recall (True positive rate)	0.923	0.923	0.962	0.923	0.913
False positive rate	0.065	0.085	0.031	0.072	0.099
False discovery rate	0.091	0.085	0.047	0.086	0.093
F1 score	0.924	0.923	0.962	0.924	0.913
Matthews Correlation Coefficient	0.843	0.83	0.923	0.842	0.808
Area Under Curve (AUC)	0.973	0.915	0.969	0.977	0.901
Negative predictive value	0.909	0.915	0.953	0.914	0.907
True negative rate	0.935	0.915	0.969	0.928	0.901
False negative rate	0.065	0.085	0.031	0.072	0.099
False omission rate	0.091	0.085	0.047	0.086	0.093
Threat score	4.556	4	10.063	4.271	3.478
Statistical parity	1	1	1	1	1

Note: *DT: decision tree, KNN: K nearest neighbor, RF: random forest, SVM: support vector machine.*

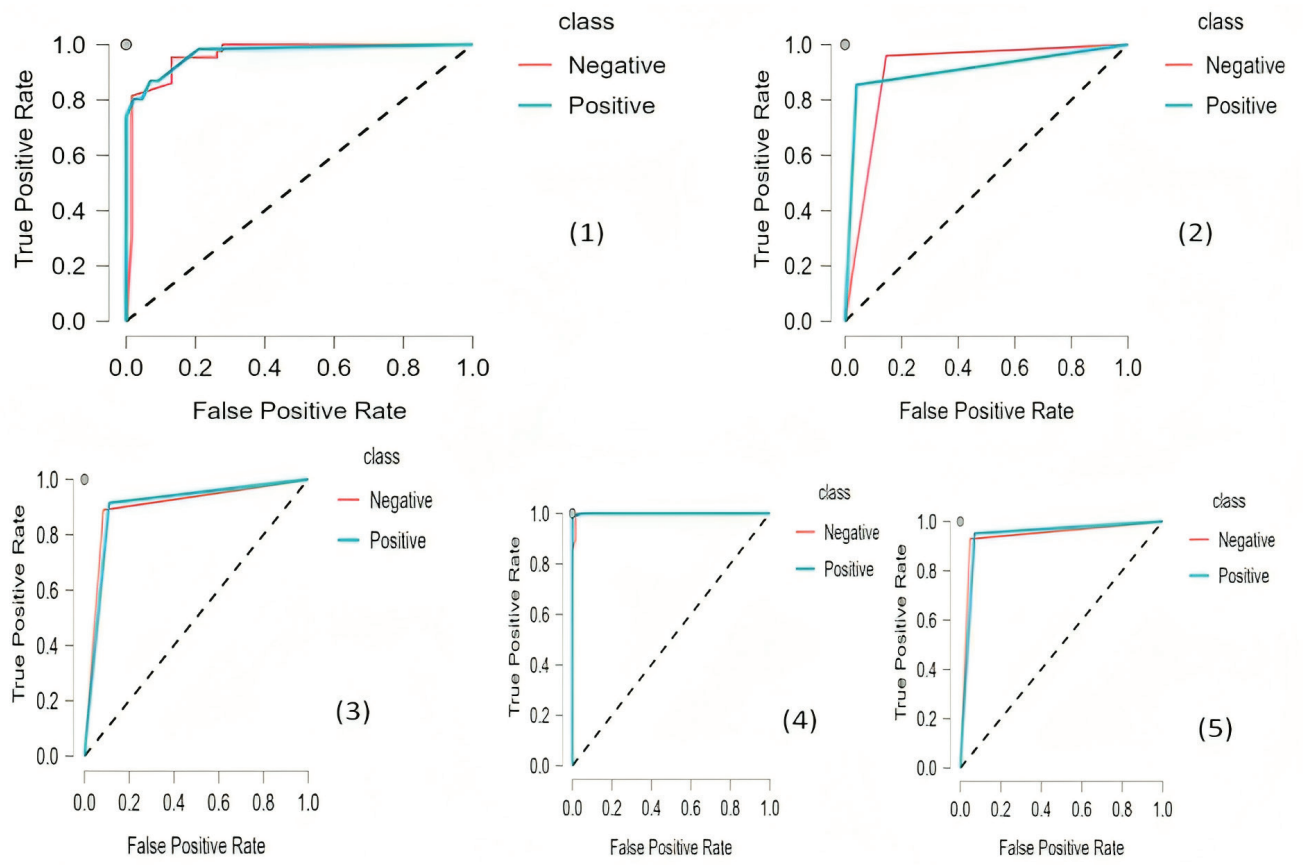

Figure 3. *ROC curves for all the machine learning models.*

Similarly, random forest and boosting algorithms were performed with the same accuracy (0.923). The KNN model beats the others in terms of accuracy, with a score of 0.962, followed by Boosting and Random Forest, with scores of 0.923. Decision Tree and Support Vector Machine (SVM) have lower accuracies of 0.920 and 0.913, respectively. KNN has the best precision and recall and the lowest false positive rate. However, Random Forest has the most significant AUC (0.977), suggesting higher overall performance in classification tasks. Despite its excellent accuracy, KNN has a significantly higher threat score than the other models, indicating that it may be more prone to mistakes. In our research, Figure 4 shows classification accuracy and other related curves of the model.

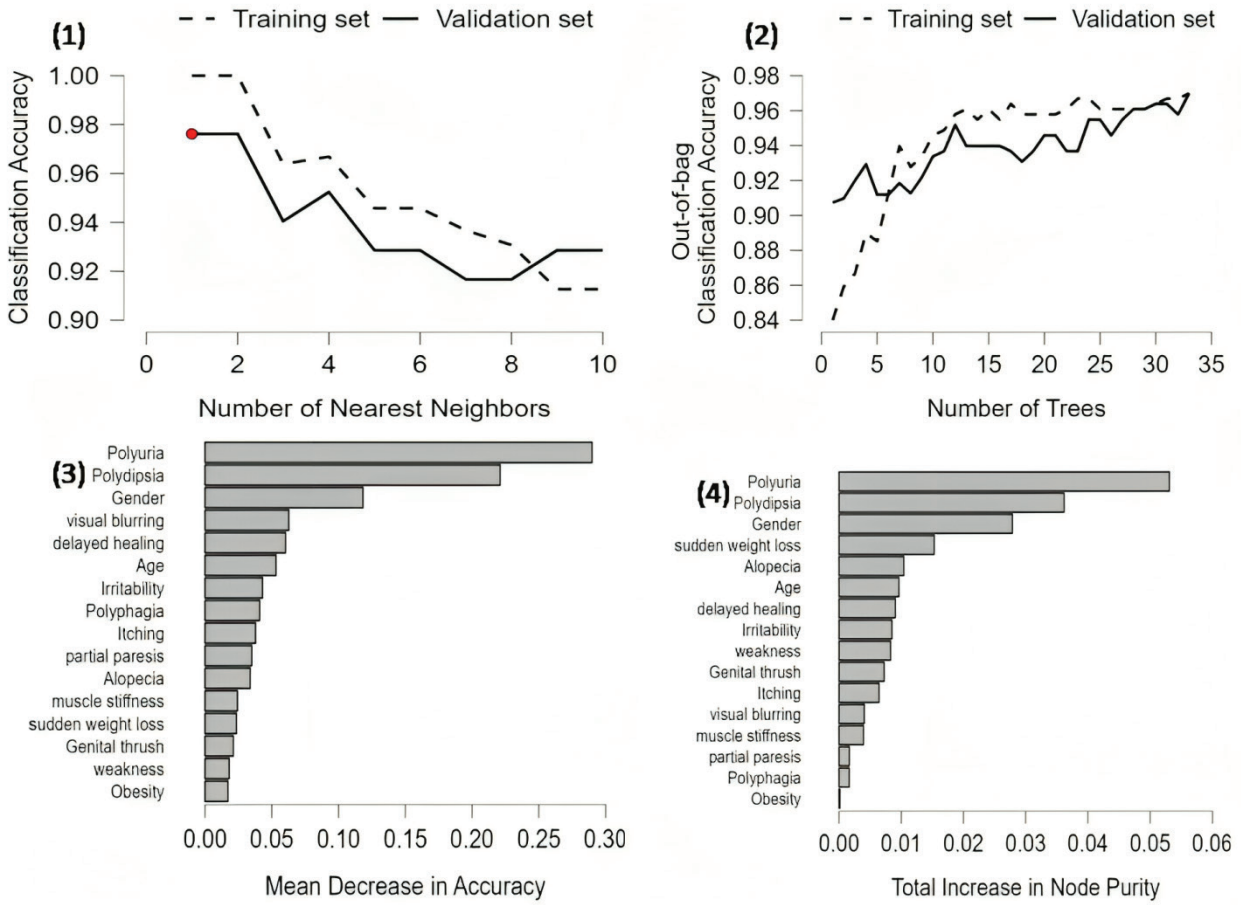

Figure 4*. Classification accuracy and other related curves of the model.*

Discussion

Early detection of any disease facilitates prompt decision-making regarding the condition, mitigates the exacerbation of complications, and conserves both time and money. Machine learning plays a significant role in diabetes prediction by leveraging various data sources to develop predictive models. The study thoroughly assesses the efficacy of different machine-learning techniques in classifying patients as either diabetic or non-diabetic in the early stages. Various metrics such as accuracy, sensitivity, specificity, the ROC curve (AUC-ROC), and precision-recall curve are used to evaluate here the performance of diabetes prediction models. We found that the positive diabetes class in females is higher (54.1%) as compared to males (45.9%), which aligns with a previous study[20,21] but contrasts with the Olufemi *et al*., where gender predictor variable indicates males are more likely to have diabetes than female in their logistic regression model in prediction of diabetes across the US.[23]

The incidence of polyuria cases in our study is found to be significantly elevated, comprising 243 (75.9%) instances within the positive diabetes class, in contrast to the negative diabetes class. In a recent study, Olufemi *et al*. utilized a logistic regression model to predict early diabetes prevalence across the United States. Their findings revealed that polyuria and polydipsia contributed most significantly to predicting the "Positive" class, as evidenced by their parameter values and odds ratios.[23] Various other data mining approaches have been employed to predict diabetes. Using the decision tree algorithm (C4.5), reports indicate that polydipsia is the most influential parameter for diabetes prediction. The performance results demonstrate a notable accuracy rate of 90.38%, suggesting the effectiveness of this algorithmic model.[22] Our comparative machine learning model also found a significant association between polydipsia and the diabetes class (p<0.001). Furthermore, we observed a substantial prevalence of polydipsia cases, accounting for

225 (70.3%) instances, compared to only 8 (4%) cases in the negative diabetes class.

Our study revealed that sudden weight loss, weakness, polyphagia, visual blurring, irritability, partial paresis, and alopecia were significantly associated with the diabetes class (p<0.001). These findings align with the research conducted by Dritsas and Trigka, who also identified polyphagia, irritability, alopecia, visual blurring, and weakness as prominent features correlated with diabetes. In contrast, other features showed negligible correlation (rank <0.2).[24] Utilizing various machine-learning models, they aimed to pinpoint individuals at risk of diabetes based on specific risk factors such as weight loss, weakness, polyphagia, visual blurring, and irritability.[24] Delayed healing (p=0.284), obesity, and itching (p=0.760) were determined to have no significant association with the diabetes class in our study. However, a previous study by Dritsas and Trigka found that diabetes prevalence was notably linked to delayed healing and visual blurring features, with 50% of diagnosed individuals exhibiting these symptoms.[24] Similarly, in our study, the mean age of the positive diabetes class was 49.1±12.1, while for the negative diabetes class, it was 46.4±12.1 (p=0.013). These findings closely resemble those of an interventional study conducted by Wicaksana *et al*. where the average age of diabetic patient participants was reported to be 55.13 years.[25]

In multivariate analysis among diabetes class and health indicators, we also analyzed the indicators other than polydipsia, polyuria, polyphagia, visual blurring, irritability, and alopecia like genital thrush, irritability, and Partial paresis are the potential health risk indicators for a positive diabetes class. Previous studies also obtained similar findings.[26,27] Our comparative analysis found that k-fold validation offers justified time costs compared to holdout validation, mainly when k is maintained at a low level to ensure high-quality classification. Among the various models examined, KNN demonstrated the highest accuracy (0.962), along with notable AUC (0.969) and F1 score (0.962). These findings are consistent with the findings reported in the previous study by Ghosh et al., where they performed another comparative analysis of different machine learning tools in detecting diabetes in their experiments.[28]

Limitations

Although the study identified several possible risk factors and symptoms for early-stage diabetes, certain limitations remain. Since the data set is publicly available, we have only considered a few variables. Further elements or variables might be included to strengthen the study. The collection of data was of a limited size. Large data sets may be used for additional studies to create more precise machine-learning models. Another similar flaw in this research is that internal validation models were used to validate machine learning models by dividing the data into an 80:20 ratio. For generalization to apply to other populations, external validation is necessary. Our future work will focus on external validation of the machine learning models to predict early diabetes risk.

Conclusion

In conclusion, we identified several potential health risk indicators for the positive diabetes class, including polyuria, polydipsia, polyphagia, genital thrush, irritability, and partial paresis. The advancement of diagnostic techniques for diabetes, emphasizing a recent trend towards technology-driven approaches, notably those employing Artificial Intelligence (AI), greatly empowers researchers and healthcare professionals in managing the disease. This shift presents opportunities for early identification of individuals at elevated risk of developing diabetes. Our research indicates that machine learning methods utilizing AI hold promise for accurately predicting diabetes. These models in healthcare effectively mitigate human-induced observation errors, resulting in enhanced outcomes. Timely detection facilitates swift and efficient treatment, potentially lowering the burden of morbidity and mortality linked to the disease.

Funding

No funding was provided for this research.

Conflict of interest

There is no potential conflict of interest.

Ethical Approval

Not applicable.

Acknowledgements

We wish to acknowledge the founder of Siksha O Anusandhan University and the chairman, Prof Dr Manoj Ranjan Nayak, for providing all the research facilities.

References

[1] Caughey GE, Roughead EE, Vitry AI, McDermott RA, Shakib S, Gilbert AL. Comorbidity in the elderly with diabetes: Identification of areas of potential treatment conflicts. Diabetes Res Clin Pract. 2010; 87(3): 385-93. doi: 10.1016/j.diabres.2009.10.019.

[2] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014; 37(Supp1): S81-90. doi10.2337/dc14-S081

[3] DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI, Simonson DC. Type 2 diabetes mellitus. Nat Rev Dis Primers. 2015; 1(1): 1-22. doi: 10.1038/nrdp.2015.19.

[4] Olokoba AB, Obateru OA, Olokoba LB. Type 2 diabetes mellitus: a review of current trends. Oman Med J. 2012; 27(4): 269. doi: 10.5001/omj.2012.68.

[5] Ginter E, Simko V. Type 2 diabetes mellitus, pandemic in 21st century. Diabetes: an old disease, a new insight. 2013: 42-50. doi: 10.1007/978-1-4614-5441-0_6.

[6] Buchanan TA, Xiang AH. Gestational diabetes mellitus. J Clin Invest. 2005; 115(3): 485-91. doi: 10.1172/JCI24531.

[7] Plows JF, Stanley JL, Baker PN, Reynolds CM, Vickers MH. The pathophysiology of gestational diabetes

mellitus. Int J Mol Sci. 2018; 19(11): 3342. doi: 10.3390/ijms19113342

[8] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2023; 10(1-2): 1-10. doi: 10.1049/htl2.12039

[9] Ahamed BS, Arya MS, Nancy V AO. Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques. Front Comput Sci. 2022; 4: 835242. doi.org/10.3389/fcomp.2022.835242

[10] Modak, S.K.S., Jha, V.K. Diabetes prediction model using machine learning techniques. Multimed Tools Appl. 2023; 83: 38523-49. doi: 10.1007/s11042-023-16745-4

[11] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018; 9: 515. doi.org/10.3389/fgene.2018.00515

[12] Sutton CD. Classification and regression trees, bagging, and boosting. Handbook of statistics. 2005; 24: 303-29. doi:10.1016/S0169-7161(04)24011-1

[13] Song YY, Ying LU. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27(2): 130. doi: 10.11919/j.issn.1002-0829.215044

[14] Panda NR, Pati JK, Pati T, Satpathy S, Bhuyan R. Comparison of artificial neural network and decision tree methods for predicting the maternal outcome in a tertiary care hospital in Odisha, India. Nat J Community Med. 2022; 13(11): 821-7. doi.org/10.55489/njcm.131120222262

[15] Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019; 134: 93-101. doi.org/10.1016/j.eswa.2019.05.028

[16] Panda NR, Mahanta KL, Pati JK, Varanasi PR, Bhuyan R. Comparison of Some Prediction Models and their Relevance in the Clinical Research. Int J Stats Med Res. 2023; 12: 12-9. doi.org/10.6000/1929-6029.2023.12.02

[17] Mucherino A, Papajorgji PJ, Pardalos PM, Mucherino A, Papajorgji PJ, Pardalos PM. K-nearest neighbor classification. Data Mining in Agr. 2009: 83-106. doi: 10.1007/978-0-387-88615-2_4

[18] Yu H, Kim S. SVM Tutorial-Classification, Regression and Ranking. Handbook of Nat comp. 2012; 1: 479-506. doi.org/10.1007/978-3-540-92910-9_15

[19] Patle A, Chouhan DS. SVM kernel functions for classification. In: 2013 International conference on advances in technology and engineering (ICATE). 2013 Jan 23 (pp.1-9), IEEE. doi. 10.1109/ICAdTE.2013.6524743

[20] Charoensakulchai S, Usawachoke S, Kongbangpor W, Thanavirun P, Mitsiriswat A, Pinijnai O, Kaensingh S, Chaiyakham N, Chamnanmont C, Ninnakala N, Hiri-o-Tappa P. Prevalence and associated factors influencing depression in older adults living in rural Thailand: A cross-sectional study. Geriatr Gerontol Int. 2019; 19(12): 1248-53. doi: 10.1111/ggi.13804.

[21] Gao M, Jebb SA, Aveyard P, Ambrosini GL, Perez-Cornago A, Papier K, Carter J, Piernas C. Associations between dietary patterns and incident type 2 diabetes: prospective cohort study of 120,343 UK biobank participants. Diabetes Care. 2022; 45(6): 1315-25. doi: 10.2337/dc21-2258.

[22] Permana BA, Ahmad R, Bahtiar H, Sudianto A, Gunawan I. Classification of diabetes disease using decision tree algorithm (C4. 5). In: Journal of Physics: Conference Series 2021; 1869 (1): 012082, IOP Publishing. doi: 10.1088/1742-6596/1869/1/012082

[23] Olufemi I, Obunadike C, Adefabi A, Abimbola D. Application of Logistic Regression Model in Prediction of Early Diabetes Across United States. Int J Sci Manag Res. 2023; 6(05): 34-48. doi: 10.0130/2023230563

[24] Dritsas E, Trigka M. Data-driven machine-learning methods for diabetes risk prediction. Sensors. 2022; 22(14): 5304. doi.org/10.3390/s22145304

[25] Wicaksana AL, Apriliyasari RW, Tsai PS. Effect of self-help interventions on psychological, glycemic, and behavioral outcomes in patients with diabetes: A meta-analysis of randomized controlled trials. Int J Nurs Stud. 2024; 149: 104626. doi.org/10.1016/j.ijnurstu.2023.104626

[26] AtınçYılmaz. PREDICTION OF TYPE 2 DIABETES MELLITUS USING FEATURE SELECTION-BASED MACHINE LEARNING ALGORITHMS. Health Prob Civil. 2022; 16(2): 128-39. doi.org/10.5114/hpc.2022.114541

[27] Sabejon JA, Rejas JB, Lumacad GS, Zarate RL, Mendez EA, Tinoy FM. XGBoost–Based Analysis of the Early-Stage Diabetes Risk Dataset. In: 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT) 2023 Jun 9 (pp. 19-24). IEEE. doi: 10.1109/APSIT58554.2023.10201658.

[28] Ghosh P, Azam S, Karim A, Hassan M, Roy K, Jonkman M. A comparative study of different machine learning tools in detecting diabetes. Procedia Comput Sci. 2021; 192: 467-77. doi.org/10.1016/j.procs.2021.08.048

[29] Early-stage diabetes risk prediction dataset, 2020, doi.org/10.24432/C5VG8H, UCI Machine Learning Repository.

[30] Panda NR, Mahanta KL, Pati JK, Pati T. Development and Validation of Prediction Model for Neonatal Intensive Care Unit (NICU) Admission Using Machine Learning and Multivariate Statistical Approach. J Obstet Gynaecol India. 2024; 74(3): 1-9. doi: 10.1007/s13224-024-02009-0

[31] Panda NR. A review on logistic regression in medical research. Nat J Community Med. 2022; 13(04): 265-70. doi.org/10.55489/njcm.134202222