Frontiers in Health InformaticsISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2025; Vol 14: Issue 1
Open Access

Diabetes Detection Using Machine Learning Algorithm

Hye-Kyeong Ko
Department of Computer Engineering, Sungkyul University, 53 Seonggyeoldaehak-ro, Manan-gu, Anyang-si, Gyeonggi-do, Republic of Korea
ellefgt@sungkyul.ac.kr

Cite this paper as: Hye-Kyeong Ko (2025). Diabetes Detection Using Machine Learning Algorithm. *Frontiers in Health Informatics,* 14(1), 12-23.

ABSTRACT
Diabetes is a chronic condition in which blood glucose levels are elevated and is responsible for several conditions that can cause disabilities resulting in poor quality of life. The prevalence of diabetes has been observed to increase not only in affluent sections of society but also in socio-economically poor sections. This reflects the seriousness of this medical condition and indicates the changing lifestyle of people worldwide. It is suggested that by 2040, there will be 642 million cases of the disease, globally. This research attempts to create a system based on machine learning (ML) to forecast a patient's risk of having diabetes. In the present study two ML algorithms, Logistic regression (LR) and K-Nearest Neighbor (KNN), were used. The LR employs odds ratios (OR) and p-values to determine diabetes risk variables, on the other hand, KNN uses nearest neighbor distance based on Euclidean distance to identify new cases based on its learning. The results of evaluation metrics such as precision or sensitivity, recall and F1 score showed that LR was better in predicting diabetes than KNN. The overall accuracy obtained with LR was 77% as compared to KNN which provided 72% accuracy. The macro-average values, which gives equal weightage to all classes (irrespective of their size) also indicated a better performance of LR. We suggest creating a better dataset which incorporates comprehensive details factors related to the lifestyle of people which may prove helpful in improving the performance of these models.

Keywords: Diabetes, Blood Glucose, Patient's, Machine Learning, Models

INTRODUCTION

Diabetes has been a cause of concern because of its long duration and non-curable characteristics. The fact that it is related to the lifestyle of people is the primary reason for its high prevalence in today's society, which is progressing towards automation, reducing most physical work to sitting jobs. The statistics related to diabetes are also disturbing. It affected nearly 415 million people in 2015 and is projected to affect around 642 million by 2040 (Godlee, 2018; Sonnet, 2019). Higher diabetes prevalence causes higher incidences of diseases such as diabetic vascular complications, such as retinopathy and nephropathy, which significantly impact patients' quality of life and cause an increase in mortality risks (Ametov & Sayamov, 2022). In addition, there are high economic impacts too related to diabetes care, which has been projected to exceed 627 billion US dollars by the year 2035, affecting low-income as well as middle-income countries and overloading healthcare systems worldwide (Sabanayagam & Wong, 2022; Adamiak & Napierała, 2012).

There are several underlying reasons contributing to the high prevalence of diabetes, such as obesity, genetics, ageing, reduced physical activity, and environmental influences (Kaul et al., 2012; Kapur et al., 2015; Mekala & Bertoni, 2020; Bloomgarden & Handelsman, 2023; Bagga et al., 2024). Changes in how people live, such as eating habits, decreased activity, and urban living, are major factors driving the increase in diabetes cases, particularly in countries with lower and medium incomes (Sally & Marshall, 2019). Moreover, the effects of early life development influenced by a mother's health and nutrition have an impact on raising the likelihood of diabetes later in life. Social factors like income levels, education access, and living conditions also play a role in determining health outcomes related to diabetes prevalence. Additionally, environmental aspects such as pollution exposure, lifestyle decisions, and lack of sleep can contribute to insulin resistance (Hill et al.,

2013; Vainshnav & Dave, 2022). Worsening the progression of diabetes. In essence, the intricate interplay between predisposition, environmental influences, and lifestyle choices highlights the growing challenge posed by diabetes (Dong et al., 2019).

Many people experience delays in getting diagnosed with diabetes because they don't have access to healthcare, lack awareness about the disease, and face stigma. Detecting diabetes early is vital for those at risk. Diagnosis involves checking fasting plasma glucose levels, conducting oral glucose tolerance tests, and measuring levels. It's important to focus on improving health programs and increasing healthcare availability to address this issue (Philomena, 2020). There hasn't been an emphasis on using factors and long-term data to predict diabetes despite its complex causes. The unpredictable behaviour of patients (like changes in how they visit or how long they stay) and the diverse nature of conditions make it challenging to gather data over a lengthy period.

Present technological advancements, such as machine learning, are proving helpful in the diagnosis and detection of diabetes. Large amounts of patient data can now be analyzed by machine learning algorithms very rapidly, assisting in identifying patterns and risk factors for diabetes (Bong-Hyun et al., 2024). The present study was conducted to develop a model for the diagnosis of diabetes in female patient.

REVIEW OF LITERATURE

Artificial intelligence (AI) has made a substantial impact on almost every field ranging from management to scientific disciplines (Wasik and Pattinson, 2024; Porwal et al., 2024). Nowadays, AI is used extensively for the detection of diseases in animals, plants and humans (AlZubi, 2023; Cho, 2024; Moses, 2022) and several intelligent machines such as wearable sensors are available in the market which use deep learning algorithms. Many studies have been conducted to test whether these ML algorithms can detect the disease with provided data. These studies have demonstrated that AI can non-invasively detect hypoglycemia (Cisuelo, et al., 2023), predict early stages of diabetes with high accuracy (Ali et al., 2023) and integrate body vitals from smartwatches to identify diabetes using a hybrid AI model (Hariharan et al., 2023; Desai et al., 2024). Furthermore, AI, specifically deep learning, has been instrumental in diagnosing and screening diabetic retinopathy (DR) with high specificity and sensitivity Early models for diabetes detection were based on an ensemble approach such as logistic regression (LR). Joshi & Dhakal (2021) employed LR for evaluating the risk factors associated with type 2 diabetes and found the 5 most important predictors to be age, glucose, body mass index, diabetes pedigree function, and pregnancy. The accuracy obtained was 78.26% with an error rate of 21.74%. Data science techniques, such as support vector machines and deep neural networks, have also been applied to improve the accuracy of diabetes prediction.

Kumar and Venu (2023) created a model and employed SVM, DT, and KNN classifiers, with variable accuracies for identifying diabetes and found that SVM performed the best with the highest accuracy of 99.65%.

Maniruzzaman et al. (2020) tested four classifiers (Random Forrest, Decision Tree, naïve Bayes, Adaboost) on a new dataset using 7 factors to predict diabetes. These classifiers were then compared for 3 partition protocols (K2, K5, and K10) and also studied the effect of data size on the performance of these classifiers. The accuracy was found to increase from K2 to K5 to K10. Also, the RF-based classifier performed better (94.25% accuracy) than other models at all levels.

Ahmed et al. (2022) proposed a model named the “Fused Model for Diabetes Prediction (FMDP)” which consisted of 2 phases, the Training and the testing. They utilized the UCI Machine Learning Repository dataset. Then SVMs and Artificial Neural Networks (ANNs) were trained. The output of SVM and ANN were fed as input to the fusion step, where fuzzy rules were employed to for the final prediction. This newly developed model was then evaluated on a new dataset acquired from another medical database as a testing dataset. The accuracy of the suggested method was 94.87%.

Massari et al. (2022) made use of ‘ontology’, that is, all those factors from which this disease diabetes takes

its form were used to develop a new model where they combined the semantic web with machine learning. Features such as its origin, symptoms, and effects, along with any additional problems were included in this model. The objective of using ontology was to provide a structured vocabulary for the description of diabetes-related knowledge. It used semantic reasoning to understand existing relationships among the dataset and identify patterns. They were able to obtain an accuracy of 73.8%.

The present study aims to develop a new model for the prediction of diabetes.

MATERIAL AND METHOD

Under this heading, the dataset, proposed machine learning (ML) framework, and performance metrics that were used to assess the framework's ability to predict diabetes are presented.

1. Dataset

The dataset employed in the present study was downloaded from Kaggle, which is an open-access repository for various kinds of datasets (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data). This dataset includes values for different predictor variables that are used to test the sugar levels in patients. The patients included are all females and the sample size is 768 belonging to the Pima Indian community in Phoenix, Arizona (Smith et al., 1988). Out of total instances, 500 patients are non-diabetic and 268 are diabetic. The total number of variables used are 8 whose description is provided in Table 1.

Table 1: Dependent and independent variables and their description

Sl no	Variables	Description
Dependent Variables		
1	Number of Pregnancies	The number of pregnancies in women has been found to have greater fasting glucose levels and HBA1c (Lv et al., 2019). Hence, this variable will be useful in predicting diabetes.
2	Glucose	The glucose level in the blood plasma is 2 hours after drinking a glucose-concentrated juice. The purpose of this test is to assess the ability to handle glucose by the human body. It is the most common test to diagnose diabetes and is known as the Oral Glucose Tolerance Test (OGTT) (Eyth et al., 2023).
3	Blood Pressure (BP)	Diabetes and BP have similar risk factors. A person having one condition has more chance of developing the other (Sharma et al., 2020). Here, The Diastolic Blood Pressure was measured (mm Hg) to see if it can predict the presence of diabetes.
4	Skin_Thickness	Triceps skin overlap thickness (mm). Skin thickness has been observed to be higher in diabetic patients (Alkhatib et al., 2020). Hence this can be an indicator for knowing the presence or absence of diabetes.
5	Insulin_Level	It measures the insulin level in serum after 2 hours of glucose solution administration comparing it with fasting level (Khalili et al., 2023).
6	BMI (Body mass index)	It is defined as the weight ratio in kilograms to the square of height in meters. High BMI is associated with diabetes (Medhi et al., 2021
7	Diabetes Pedigree Function	This function is used to know the probability of developing diabetes based on family history (Akemese, 2022).
8	Age	Age as a predictor of diabetes was also evaluated.
Independent Variable		
1	Presence or absence of Diabetes	Diabetic or not (0 or 1)

Table 2 presents the descriptive statistics which provides an insight into the mean values and standard

deviation for all the dependent variables along with minimum and maximum values**.** It is observed that 17 was the highest number of pregnancies with an average of approximately 4 pregnancies per woman. The age varied from a minimum of 21 years to a maximum of 81 years with a mean age of 33.3 years.

Table 2: Descriptives Statistics

	Descriptive Statistics for N=768			
	Minimum	Maximum	Mean	Std. deviation
NOPeg	0	17	3.8	3.37
Gl_L	0	199	121.0	32.0
DBP	0	122	69.1	19.4
SK_T	0	99	20.5	16.0
In_L	0	846	79.8	115.2
BMI	0	67.1	32.0	7.9
DPF	0.08	2.42	0.57	0.33
Age	21	81	33.2	11.8

NOPeg-Number of pregnancies; Gl_L-Glucose Level; DBP-Diastolic Blood Pressure; SK_T-Skin Thickness; In_L-Insulin Level; BMI-Body Mass Index; DPF-Diabetes pedigree Function

2. **Data Normalization**:

The CSV file downloaded consisted of 769 rows and 9 columns. This data was cleaned by scrutinizing any missing values or outliers. It was then normalized by using the z-score normalization process. This involves transforming data in a way that the value of the mean becomes 0 and the standard deviation is 1. This helps to standardize the range of features facilitating easy comparison between them, especially if they were measured on different scales. It is the most commonly used method in k-nearest Neighbors (kNN) where ensuring equal feature contribution from each feature is required for measuring the distances.

The formula utilized for calculating z-score is given as

$$Z = \frac{X - \mu}{\sigma}$$

Where Z is the z-score, X is the value to be normalized, μ is the mean of the feature, σ is the standard deviation of the feature.

3. Training the Models with the Dataset

KNN- The KNN model was trained by using the training dataset, which was obtained by splitting the original dataset into training and test datasets in an 80:20 ratio. For this, an optimum k-value was selected through the process of repeated testing and validation. In this study, the k=7. It is to be noted that in KNN, training simply involves storing the training data.

KNN

KNN is a simple form of supervised machine learning that is used to solve both classification and regression problems. It is often named as an instance-based learning model as it trains itself by memorizing the instances present in the dataset where predictions are made by comparing new instances to those memorized. The fundamental idea of KNN is straight as it just stores the labelled training data. As a new instance is received, it classifies it according to the similarity of the stored data (Lopez-Bernal et al., 2021). Visualization of KNN works is given in Figure 1. The red triangles and blue squares represent two classes of items and the green dot is an unknown item. Here, the classification was made by using the distance of the green dot present inside the inner circle which is the 'k' representing the number of nearest neighbour items (blue squares or red triangles) that the user set to make the classification. In the given figure, k=3 which means that 3 closest data points (neighbors) were taken into account to decide the class of the green dot. Based on this, the green dot is classified as a red triangle.

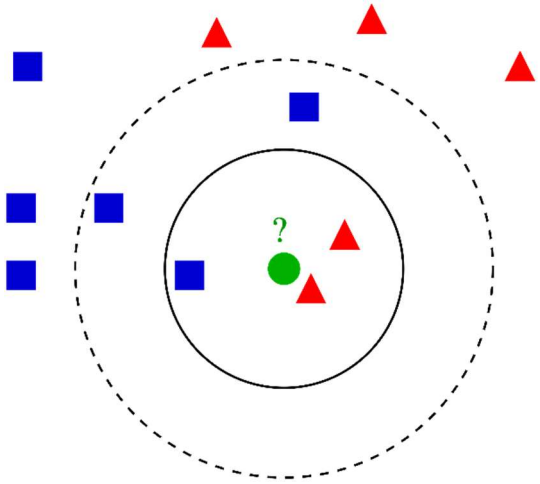

Figure 1: Visulaization of KNN classification

There are several ways to calculate this distance. The Euclidean Distance was used in the present study to measure the distance of the new data point. The Euclidean distance for points in an n-dimension space is given as

$$d = \sqrt{\sum_{i=1}^{n} (x_{i2} - x_{i1})^2}$$

where 'x_{i1}' and 'x_{i2}'are the coordinates of the two points in the ith dimension and 'd' is the distance between two points. To determine the class (diabetes or not) of the new data point, conduct a majority vote among the k nearest neighbors.

Logistic Regression (LR)

Machine learning classifications for two classes (or binary classifications) is usually performed with the help of a statistical technique known as Logistic regression. It can be considered as one of the types of regression analysis where the outcome variable is categorical. LR is based on the logistic (sigmoid) function, which maps any real-valued number into a value between 0 and 1. This function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where z is the linear combination of input features.

Here, these input features have a certain weight which is then added and put through this sigmoid function to convert them into numbers between 0 and 1. For example, here features like age, glucose level, etc. are multiplied by their weight, and then all these values are added up which are then interpreted as the probability of the item belonging to one or the other class. This is done by computing the odds ratio (ratio of the probability of the event taking place with the probability of that event not taking place).

$$odds = \frac{P}{1 - P}$$

In logistic regression, the dependent variable is the log of the odds (log-odds, logit) defined as:

$$logit\ (p) = log\left(\frac{P}{1 - P}\right)$$

An item's categorization is determined by a threshold, often set at 0.5. The instance is placed in one class if the estimated probability exceeds the threshold; if not, it is placed in the other class.

4. Evaluation metrics

When evaluating any model's performance, the statistical metrics that is frequently utilized is sensitivity. Sensitivity, also known as true positive rate (TPR) or Recall, measures how well true positives were identified (in this case, how well the model can classify patients who actually have diabetes). It is be defined as

$$Sensitivity \text{ or } Recall = \frac{True\ positive}{True\ Positive + Fals\ \ Negative}$$

where, True positives refer to the predictions made by the model that were correct (Diabetes present and correctly predicted as diabetes), and False negatives are the predictions that were incorrect (Diabetes present but predicted as not present).

Another measure is the Precision which measures how accurately true predictions were made i. e. out of all predictions that were true how many were correct. It is defined as

$$Precision, P = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ Positive (FP)},$$

The overall performance of a model is denoted by accuracy which is defined as the total true predictions made by the model and is defined as:

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

In addition, F1 Score is also used which combines precision and recall, by taking their harmonic mean for its calculation and is represented as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Macro averaged Precision, recall and F1: It presents the average values for total instances by considering equal contribution by all the classes.

E.g. Macro average precision $= \frac{1}{k}\sum_{k=1}^{k} Precision\ score\ for\ all\ classes\ k$

Here, every class is treated similarly. Macroaveraging is helpful if the dataset has unbalanced classes and has better predictive accuracy.

Weighted average Precision, recall and F1: The weighted average for each class is computed by considering the number of samples per class.

$$M_{Weighted} = \frac{\sum_{i=1}^{N}(M \times Support)}{\sum_{i=1}^{N} \quad Support_i}$$

N = Number of Classes

M = Performance metric (Precision, Recall or F1)

support $_{i=}$ is the number of instances in class *i*

Confusion Matrix: It is a two-dimensional matrix that shows how accurate a model is at **categorization**. An n x n matrix, where n is the number of labels in a given dataset, is called a confusion matrix. Predicted labels are represented by each column, while actual labels are represented by each row. The model's performance is displayed in the confusion matrix, as seen in Figure 1.

	Predicted: No (Class 0)	Predicted: Yes (Class 1)
Actual: No (Class 0)	TN	FP
Actual: Yes (Class 1)	FN	TP

Figure 2: The Confusion Matrix

True positive (TP) represents the number of positively labelled data that are correctly classified in the confusion matrix; true negative (TN) represents the number of negatively labelled data that are correctly classified; false positive (FP) represents the number of negatively labelled data that are mistakenly classified as positive; and false negative (FN) represents the number of positively labelled data that are mistakenly classified as positive.

RESULT AND DISCUSSION

Table 2 presents the performance metrics values obtained in the present study when KNN and LR methods were used for the identification of patients having diabetes.

It was observed that for class 0 (i. e. No Diabetes), the precision was 0.79, indicating that 79 % of the cases predicted by Logistic regression as not having diabetes were correctly classified. On the other hand, in class 1 (Diabetes present group), the precision was 0.6, i.e. 60% of the cases predicted as having diabetes were correctly classified. The Recall value which measures the ability of the classifier to find all the positive cases was 0.77 for class 0 and for class 1 it was 0.64. Likewise, the values for the F1 score were higher for class 0 (0.78) than for class 1 (0.62). Overall, the accuracy of the KNN model was 72%.

Similar to KNN, the LR model also showed higher values for precision, recall and F1 score for class 0 (0.81, 0.83, and 0.82, respectively) and for class 1, the values were 0.68, 0.65 and 0.67 for precision, recall and F1 score, respectively.

Table 3: The evaluation metrics for KNN and LR

	KNN			LR				
Metric	**Class 0 (No Diabetes)**	**Class 1 (Diabetes)**	**Macro Avg**	**Weighted Avg**	**Class 0 (No Diabetes)**	**Class 1 (Diabetes)**	**Macro Avg**	**Weighted Avg**
Precision	0.79	0.6	0.7	0.72	0.81	0.68	0.75	0.76
Recall	0.77	0.64	0,7	0.72	0.83	0.65	0.74	0.77
F1-Score	0.78	0.62	0.7	0.72	0.82	0.67	0.74	0.77
Support (Count)	99	55	154	154	99	55	154	154
Accuracy	0.72							0.77

When the two models were compared, it was observed that LR outperformed KNN in terms of precision and recall for both classes suggesting LR's superiority over KNN in correctly classifying instances of both diabetes and no diabetes. Similarly, the F1 values which consider both precision and recall were higher for LR as compared to KNN, indicating a better balance between precision and recall.

Macro-average has been considered important in cases where outcomes are dominated by excessively frequent kinds (Gowda, 2021). For example, in this study, the number of patients with no diabetes was much greater than the number of patients with diabetes. The justification for giving this metric is to evaluate 'classifier performance equally w.r.t. all classes', or 'does not neglect rare classes' (Opitz, n.d.). Figure 3 presents the macro averages of precision, recall and F1 score. The macro recall is a better indicator of a model's performance due to its interpretability and closer relationship with accuracy (Opitz, n.d.). It is observed that macro average recall is higher for LR than KNN.

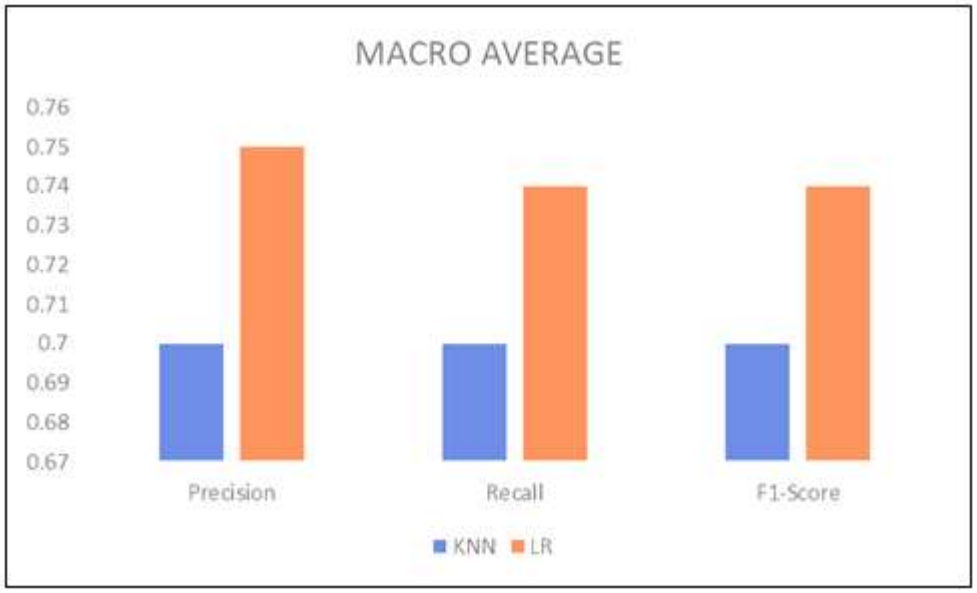

Figure 3: The Macro-Average Precision, Recall and F1 value

Table 4 presents the comparison of confusion matrices for KNN and LR. Logistic Regression model appears to perform better overall compared to the k-Nearest Neighbors model, especially in identifying patients who were not diabetic whereas KNN misidentified more non-diabetic patients as diabetic. The results of this study indicate that both LR and KNN can be helpful in the prediction of diabetes using variables such as glucose level, insulin level, number of pregnancies, blood pressure, skin thickness, body mass index, diabetes pedigree function and age. However, LR performed better than KNN.

LR		**Predicted**	
		0	**1**
Actual	**0**	82	17
	1	19	36
KNN		**Predicted**	
Actual	**0**	76	23
	1	20	35

When looking at the studies conducted by other researchers it is observed that the accuracies obtained by the models used in the present study were comparable. Also, the finding that LR was better for the prediction of diabetes than KNN was corroborated by these studies (Table 4). While the Pima Indian Diabetes Dataset produced almost similar results in different studies along with the present study, the evaluation metrics obtained by Ganie et al. (2022) who utilized a different dataset, showed better performance. This can be attributed to their dataset's extensive and pertinent characteristics which were able to capture a wider range of variables that influence the lifestyle of individuals subsequently affecting the outcome i.e. presence or absence of diabetes, hence, enabling improved ML model performance.

	LR			KNN			Dataset used
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Pima Indian Dataset
Panda et al. (2022)	0.77	0.58	0.66	0.73	0.52	0.6	Pima Indian Dataset
Kangra & Singh (2023).	0.74	0.74	-	0.67	0.66	-	Pima Indian Dataset

Ganie et al. (2022).	0.876	0.931	0.903	0.793	0.77	0.783	Indian Population from Jammu and Kashmir (India)
Khanam & Foo(2021).	0.788	0.789	0.788	0.804	0.794	0.798	Pima Indian Diabetes dataset

CONCLUSION

Diabetes has been observed to be an emerging epidemic worldwide and its prevalence is increasing day by day. It not only compromises the health of patients reducing their quality of life but exerts undue stress on their relatives, both financially and emotionally, worth mentioning the economic burden on the healthcare system. Hence, it becomes imperative for its early detection. Predictive algorithms offer a superior option for the automated diagnosis of diabetes. This paper presented the ML predictive algorithms KNN and LR to help forecast diabetes. The Pima Indian Diabetes Dataset was used for the experiment's trial run. The results suggest that LR perform better for terms of all the evaluation metrics. Comparing across different studies it is revealed that if dataset features can be improved, the models can perform better. So we intend to collect a local dataset collecting comprehensive features representing various lifestyle factors in addition to medical test results and evaluate the performance of these models in our future works.

ACKNOWLEDGEMENT

'This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT). (NO.NRF-2021R1A2C1012827) in (2023)'

Declaration of Conflicts of Interests

The author declares that there is no conflict of interest.

Availability of data and materials

The datasets used in the current study available from the corresponding author on reasonable request.

Use of Artificial Intelligence

Not applicable

Declarations

The author declare that all works are original and this manuscript has not been published in any other journal.

REFERENCES

Adamiak, M. & Napierała, T. (2012). Financing the development of tourism in the Łódź Voivodeship and Oppland county (Norway). Acta Innovations, 5, 57–75. https://www.actainnovations.com/index.php/pub/article/view/5_2 ,

Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529-8538.

Ali, Z., Ismail, N., & Ahmad, K. (2023). A Study of Imam Al-Ghazali's Approach in Strengthening Spirituality, Psychology and Mental Health of Muslims. Journal for ReAttach Therapy and Developmental Diversities, 6(10s (2)), 409-421.

Alkhatib, A., Sindiani, A., Alshdaifat, E., Funjan, K., Khader, Y. (2020). Skin Thickness can Predict the Progress of Diabetes Type 2: A New Medical Hypothesis. *EC Diabetes and Metabolic Research* 4.8 (2020): 08-12.

AlZubi, A.A. (2023). Artificial Intelligence and its Application in the Prediction and Diagnosis of Animal

Diseases: A Review. Indian Journal of Animal Research. 57(10): 1265-1271. https://doi.org/10.18805/IJAR.BF-1684

Ametov, A.S., & Sayamov, Y.N. (2022). Bioethics of Diabetes Mellitus as a Global Problem of the Modern World. Doctor.Ru. https://doi.org/10.31550/1727-2378-2022-21-2-56-58

Bagga, T., Ansari, A. H., Akhter, S., Mittal, A. & Mittal, A. (2024). Understanding Indian Consumers' Propensity to Purchase Electric Vehicles: An Analysis of Determining Factors in Environmentally Sustainable Transportation. International Journal of Environmental Sciences, 10(1), 1-13.

Bloomgarden, Z., Handelsman, Y. (2023). Diabetes Epidemiology and Its Implications. In: Jenkins, A.J., Toth, P.P. (eds) Lipoproteins in Diabetes Mellitus. Contemporary Diabetes. Humana, Cham. https://doi.org/10.1007/978-3-031-26681-2_31

Bong-Hyun, K., Alamri, A. M. and AlQahtani, S. A. (2024). Leveraging Machine Learning for Early Detection of Soybean Crop Pests. Legume Research. 47(6): 1023-1031. https://doi.org/10.18805/LRF-794.

Cho, O.H. (2024). An Evaluation of Various Machine Learning Approaches for Detecting Leaf Diseases in Agriculture. Legume Research. https://doi.org/10.18805/LRF-787

Cisuelo, O., Stokes, K., Oronti, I. B., Haleem, M. S., Barber, T. M., Weickert, M. O., ... & Hattersley, J. (2023). Development of an artificial intelligence system to identify hypoglycaemia via ECG in adults with type 1 diabetes: protocol for data collection under controlled and free-living conditions. BMJ open, 13(4), e067899.. https://doi.org/10.1136/bmjopen-2022-067899

Desai, G.N. Patil, J. H., Deshannavar, U. B. & Hegde, P. G. (2024). Production of Fuel Oil from Waste Low Density Polyethylene and its Blends on Engine Performance Characteristics . Metallurgical and Materials Engineering, 30(2), 57–70. https://doi.org/10.56801/MME1067,

Dong, G., Qu, L., Gong, X., Pang, B., Yan, W., & Wei, J. (2019). Effect of social factors and the natural environment on the etiology and pathogenesis of diabetes mellitus. International Journal of Endocrinology, 2019.. https://doi.org/10.1155/2019/8749291

Eyth E, Basit H, Swift CJ. Glucose Tolerance Test. [Updated 2023 Apr 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532915/

Ganie, S. M., Malik, M. B., & Arif, T. (2022). Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches. *Journal of Diabetes & Metabolic Disorders*, *21*(1), 339-352. doi: 10.1007/s40200-022-00981-w

Godlee, F.. (2018). The growing problem of diabetes. BMJ, 363. https://doi.org/10.1136/BMJ.K4921

Gowda, T., You, W., Lignos, C., & May, J. (2021). Macro-average: rare types are important too. *arXiv preprint arXiv:2104.05700*.

Hariharan, S., Sridharan, D., R, K., T A, M., Tamilselvi, C., & Sam, D. (2023). Real-time Monitoring and Early Detection of Diabetes with Bioactive and Biological Impedance Sensors using Hybrid Machine Learning Algorithm. 2023 4th International Conference for Emerging Technology (INCET), 1-8. 1-8. https://doi.org/10.1109/INCET57972.2023.10170610

Hill, J., Nielsen, M., & Fox, M. H. (2013). Understanding the social factors that contribute to diabetes: a means to informing health care and social policies for the chronically ill. The Permanente Journal, 17(2), 67. https://doi.org/10.7812/TPP/12-099

Huang X, Wang H, She C, Feng J, Liu X, Hu X, Chen L and Tao Y (2022) Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. Front. Endocrinol. 13:946915.

https://doi.org/10.3389/fendo.2022.946915

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health, 18(14), 7346.

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1728-1737. DOI: 10.11591/eei.v12i3.4412

Kapur, A., Schmidt, M. I., & Barceló, A. (2015). Diabetes in socioeconomically vulnerable populations. International Journal of Endocrinology, 2015, 247636-247636. https://doi.org/10.1155/2015/247636

Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., & Chibber, R. (2012). Introduction to diabetes mellitus. Advances in experimental medicine and biology, 771, 1–11. https://doi.org/10.1007/978-1-4614-5441-0_1

Khalili, D., Khayamzadeh, M., Kohansal, K., Ahanchi, N. S., Hasheminia, M., Hadaegh, F., Tohidi, M., Azizi, F., & Habibi-Moeini, A. S. (2023). Are HOMA-IR and HOMA-B good predictors for diabetes and pre-diabetes subtypes?. BMC endocrine disorders, 23(1), 39. https://doi.org/10.1186/s12902-023-01291-9

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432-439. https://doi.org/10.1016/j.icte.2021.02.004

Kumar, A. A., & Venu, N. (2023). Machine Learning for Diabetes Prediction Based on Clinical Data and Risk Factor Measures. High Technology Letters. 29(11), 235-250.

Lopez-Bernal, D.; Balderas, D.; Ponce, P.; Molina, A. (2021). Education 4.0: Teaching the Basics of KNN, LDA and Simple Perceptron Algorithms for Binary Classification Problems. Future Internet, 13, 193. https://doi.org/10.3390/fi13080193

Lv, C., Chen, C., Chen, Q., Zhai, H., Zhao, L., Guo, Y., & Wang, N. (2019). Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. *Menopause (New York, N.Y.)*, *26*(9), 1010–1015. https://doi.org/10.1097/GME.0000000000001349

Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8, 1-14.

Massari, H. E., Mhammedi, S., Sabouri, Z., & Gherabi, N. (2022). Ontology-based machine learning to predict diabetes patients. In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21 (pp. 437-445). Springer International Publishing.

Medhi, G. K., Dutta, G., Borah, P., Lyngdoh, M., & Sarma, A. (2021). Prevalence of Diabetes and Its Relationship With Body Mass Index Among Elderly People in a Rural Area of Northeastern State of India. *Cureus*, 13(1), e12747. https://doi.org/10.7759/cureus.12747

Mekala, K. C., & Bertoni, A. G. (2020). Epidemiology of diabetes mellitus. In Transplantation, bioengineering, and regeneration of the endocrine pancreas (pp. 49-58). Academic Press. https://doi.org/10.1016/B978-0-12-814833-4.00004-6

Moses, M. B., Nithya, S. E. & Parameswari, M. (2022). Internet of Things and Geographical Information System based Monitoring and Mapping of Real Time Water Quality System. International Journal of Environmental Sciences, 8(1), 27-36.

Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. *IAES International Journal of Artificial Intelligence*, *11*(1), 284. Doi:10.11591/ijai.v11.i1.pp284-290

Philomena, V. S. (2020). A Statistical Evaluation of the Data of Ovarian Cancer and Uterine Cancer in Women. Bio-Science Research Bulletin, 36(1), 40-49. ,

Porwal, S., Majid, M., Desai, S. C., Vaishnav, J. & Alam, S. (2024). Recent Advances, Challenges in Applying Artificial Intelligence and Deep Learning in the Manufacturing Industry. Pacific Business Review (International), 16(7), 143-152

Sabanayagam, C. & Wong, T. N. (2022). In Ed(s) Cheng C.-Y., & Wong, T. Y. Ophthalmic Epidemiology-Current Concepts to Digital Strategies. CRC Press. Boca Raton. 185-194.

Sally, M., Marshall. (2019). A life course perspective on diabetes: developmental origins and beyond.. Diabetologia, 62(10):1737-1739. https://doi.org/10.1007/S00125-019-4954-6

Sharma, A., Mittal, S., Aggarwal, R., & Chauhan, M. K. (2020). Diabetes and cardiovascular disease: inter-relation of risk factors and treatment. *Future Journal of Pharmaceutical Sciences*, *6*, 1-19. https://doi.org/10.1186/s43094-020-00151-w

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

Sonnet, E. (2019). Software in Diabetes. Handbook of Diabetes Technology, 83-93.

Vainshnav, S. P., & Dave, K. K. (2022). A study on Artificial Intelligence and Machine Learning in Banking Sector with special reference to term loan. Pacific Business Review (International), 15(3), 34-53.

Wasik, S. and Pattinson, R. (2024). Artificial Intelligence Applications in Fish Classification and Taxonomy: Advancing Our Understanding of Aquatic Biodiversity. FishTaxa, 31: 11-21.