Scholar J
Computational Science

Research Article

Artificial Intelligence-Based Prediction of Diabetes Mellitus Using Health Checkup Data

K. V. Odunuga[1], R. E. Ochogwu[2], C.I. Osuji[3], O. E. Owoicho[4], F. B. Oredipe[5], O. A. Bamgbose[6], S. I. Okogu[7], M. A. Sunmola[8] and A. O. Adebanjo[9]

[1,5]Olabisi Onabanjo University Teaching Hospital, Sagamu, Nigeria.
[2]Department of Computer Science, University of Nigeria, Nsukka, Nigeria.
[3]Department of Computer Science, Federal University, Wukari, Nigeria.
[4,8] Corporation for Africa and Overseas (CFAO) group, Nigeria.
[6]Olabisi Onabanjo University, Sagamu, Nigeria.
[7]University of Uyo Teaching Hospital, Uyo, Nigeria.
[9] Federal University of Technology, Minna, Nigeria.

Received 13 August 2024; Acceptance 3 September 2024; Published 11 September 2024.

Abstract

Diabetes Mellitus (DM) is a chronic condition due to chronic high blood glucose levels caused by relative or absolute insulin deficiency. AI could improve diabetes management by providing real-time health information to patients or providers, facilitating patient self-management, and enhancing intervention targeting high-risk populations. The study builds a machine-learning model to predict patients with diabetes mellitus. The study employed a quantitative analysis to understand the predictive power of machine learning algorithms for the risk of developing diabetes. The data of 768 patients was obtained from Kaggle Health data and was used to train a predictive model in Jupyter Notebook. The variables include socio-demographic data, clinical measurements, medical history, and diabetes outcomes. The study adopted the Support Vector Machines (SVM) as the model of choice used for classification. The model was trained by splitting the data into training (70%) and testing (30%) sets. The model was evaluated by assessing the precision and accuracy scores. Data was simulated; thus, no actual participants were involved. The dataset size was 768 samples, and the outcome distribution included Non-diabetic: 500 cases and diabetic: 268 cases. Glucose had a mean = 69.11 and range = 0 to 199. The Body Mass Index had a mean = 31.99 and a range = 0 to 67.10. The mean age was 33.24, and the range = was 21 to 81. The features were standardized using `StandardScaler` with a mean of 0 and a standard deviation of 1. The data was split into training (70%) and testing (30%) sets. The model used a Support Vector Machine (SVM) with a linear kernel. The model performance was Training Accuracy 78.66% and Test Accuracy 77.27%. **The model had a high accuracy score within the sample size used.** The model could predict patients who had diabetes by the parameters.

***Keywords**:* Data Security, DDoS Attack, cybersecurity and e-government

Correspondence to: Kayode V. Odunuga, e-mail: kayodeodunuga94@gmail.com

How to Cite: Odunuga et al. (2024). Artificial Intelligence – Based Prediction of Diabetes Mellitus Using Health Checkup Data. *Scholar J Computational Science,* **1(9)**. DOI: 10.5281/zenodo.13748331

Introduction

Diabetes mellitus (DM) is a chronic metabolic disorder characterized by hyperglycemia with multiple disturbances of carbohydrate, protein, and fat metabolism due to absolute or relative insulin deficiency with dysfunction of organ systems [1-2]. The increasing prevalence of diabetes has become a global public health issue in this century [3]. In the past, diabetes mostly occurred in developed 'Western' economies; nowadays, diabetes happens all over the world because of the increasing consumption of nutrient-poor and energy-dense foods and an increasingly sedentary lifestyle [4]. The latest published figure by the International Diabetes Federation (IDF) is 425 million persons living with diabetes worldwide, of whom nearly 50% are not diagnosed [5]. The developing economies of Africa and Asia contribute significantly to the above figures. The burden of complications of DM has also risen along with the ever-increasing prevalence of the disease [6]. There is now an increasing rate of amputations, cerebrovascular disease, heart disease, and kidney disease in populations that were not known for these challenging health problems in the past [7].

The rise of digital health technologies (DHTs) - artificial intelligence (AI) in particular - may be able to overcome these hindrances and alleviate the disease burden of diabetes in the future since AI-supported DHTs can help to implement better prevention interventions in high-risk populations manage patients with diabetes who are not able to attend physician appointments in person, provide real-time health and metabolic information, facilitate better patient self-management, and save time and money by reducing the need to travel to an in-person appointment [8-10].

Machine learning could be a game-changer for healthcare, particularly for predicting and managing chronic diseases like diabetes. Machine learning is a form of artificial intelligence (AI) that allows computers to make predictions or decisions without being explicitly programmed [7]. It is capable of discovering intricate patterns in databases and uses sophisticated mathematical rules to make 'inferences' not immediately apparent from standard statistical approaches [11-12].

Diabetes is a chronic disease directly related to the pancreas, and the body cannot produce insulin. Insulin is the main reason for maintaining blood glucose levels [1]. Many factors can cause problems for a person affected by diabetes, such as being overweight, physical inactivity, high blood pressure, and cholesterol levels [7]. It has many complications, but the most common is the increase in urination. It can damage the skin, eyes, and nerves, and if people with diabetes do not treat that in time, it can cause kidney failure, diabetic retinopathy, and ocular disease [13].

Several challenges exist in preventing and controlling diabetes in the traditional face-to-face medical setting [14]. First, the prevention and early diagnosis of diabetes have always been obstacles to control since many cases of diabetes have been diagnosed years after the onset of diabetes [10]. Secondly, the management of patients with diabetes is based on regular follow-up of meticulous examination of blood glucose control and diabetic complications, which require integrated diabetes management by endocrinology, podiatry, nutrition, nephrology, and ophthalmology [15]. Third, medical resources are unevenly distributed, and there is a shortage of high-quality human resources and a low capacity for primary health care [8]. Last but not least, diabetes is perhaps the most typical example of a highly prevalent chronic disease that requires a patient to take an active continuous role in its management because of its dependency on diet and exercise, the wide range of complications across the physiological systems, and the necessity for self-monitoring [6].

AI is a broad branch of computer science that focuses on developing theories, methods, technologies, and application systems to simulate, improve, and expand human intelligence in machines. Machine learning (ML) is a subcategory of AI that uses statistical methods to develop intelligent systems [10,15]. A machine can be trained to automatically learn and improve its performance (e.g., accuracy) without being explicitly

programmed through using a supervised (training algorithm) or unsupervised (unsupervised learning) approach [11].

Deep learning (DL) is a category of AI that uses sophisticated ML techniques to solve computer vision and natural language processing tasks more successfully than traditional ML (including support vector machines, decision trees, and logistic regressions) [4]. Its primary strength is feature extraction and pattern recognition, which uses multiple processing layers (artificial neurons arranged in an input layer, hidden layer, or output layer) to learn representations of data with different levels of abstraction to associate the input with the diagnostic output.

Precision diabetes medicine is a new way of diagnosing, preventing, and treating diabetes that considers individual variation and utilizes information from many different data sources to assess an individual's status, prognosis, and response to therapy [1]. A fundamental component of precision diabetes medicine is precision prognostics, that is, the ability to develop predictive models to estimate the risk of T2DM and its complications for a given individual based on risk profiles1. This facilitates the identification of high-risk individuals, enabling tailored prevention strategies and treatments that target the high-risk individuals most likely to benefit from prevention or treatment to delay or prevent disease onset and its complications [15]. This approach is recommended by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD) Consensus Report [6]. It promotes targeting high-risk individuals with lifestyle interventions and glucose-lowering medications to prevent or delay T2DM. Thus, predictive models are translated into clinical practice through several stages: development, evaluation, and translation into clinical decision support [8]. The development of a model requires longitudinal data on individuals' biological, lifestyle, and environmental interactions. The next step is the evaluation of a model's performance [12]. The hallmarks of an excellent predictive model are its accuracy in estimating an individual's risk: it is well calibrated (its predictions closely match observed outcomes), has good discrimination (it reliably distinguishes individuals at high versus low risk of the outcome), and generalizable (its performance is similar across populations) [11]. Thus, calibration and discrimination can be evaluated internally (using the same dataset on which the model was developed) or externally (using a different dataset). An external validation is typically preferred as it better tests the model's generalisability [15].

Materials and Method

Study Design

This study will employ a quantitative research design to evaluate the effectiveness of machine learning algorithms in predicting the likelihood of diabetes in patients. The approach involves developing and validating predictive models using historical patient data.

Data Source

This study collected data from Kaggle Health data to create a predictive model in Jupyter Notebook to assist in diagnosing diabetic patients. The model aims to support healthcare professionals in making informed decisions using reliable electronic sources. Data from 768 patients were utilized.

Data Processing

Data variables will include socio-demographic characteristics (age), clinical measurements (Body Mass Index, Blood pressure, Skin Thickness, Insulin, glucose levels), medical history (previous pregnancies, diabetes pedigree function), and outcome.

Data Cleaning

Handling Missing Data: If imputation is not viable, use imputation techniques to fill in missing values or exclude incomplete records.

Outlier Detection: Identify and address outliers using statistical methods or domain expertise.

Data Transformation

Normalization/Standardization: Normalize or standardize numerical variables to ensure they are on a similar scale.

Encoding Categorical Variables: Convert categorical variables into numerical formats using techniques such as one-hot encoding.

Machine Learning Models

Model Selection

The study adopted the support vector machines (SVM) for classification.

Model Training

Training Data Split: The data was divided into sets of training (70%) and testing (30%), both of which were trained.

Model Evaluation

The model was evaluated for performance using precision and accuracy scores.

Model Validation

External Validation

The model was validated with a sample of the data.

Ethical Considerations

The data was set as an imaginary data set. Thus, no participants were involved in the study.

Results and Discussion

Table 1. Datasheet

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000
mean	3.845052	69.105469	20.536458	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	19.355807	15.952218	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	240.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

kaydiab.head ()

In a Jupyter Notebook, running kaydiab.head() returned the first few rows of the DataFrame kaydiab. This is a useful way to get a high-level look at the contents of the dataset by viewing the top entries.

kaydiab: This is a one of the most used data manipulation libraries of Python.

.head(): this method returns the first five rows of the DataFrame by default. If you want to display a different number of rows, you should pass an integer to the method (kaydiab.head(10) e.g., to return the first 10 rows).

Machine learning methods enable us to develop AI applications that discover new patterns, previously not specifiable, in the data without specifying decision rules for particular tasks and without considering complex interactions among features (Mackenzie, Sainsbury, & Wake, 2024). Thus, ML has become the formal framework for building AI utilities. Therefore, this model can serve as an AI-assisted digital healthcare

ecosystem that includes diabetes prevention and management as a potential future vision for diabetes care [15].

Prediction of diabetes onset is a part of preventive medicine because it identifies those likely to develop diabetes at the pre-disease stage [14]. Thus, such a technology could reduce the incidence of diabetes by treating these people medically early on before they develop diabetes [12].

The dataset comprehensively overviews various health indicators and their distributions. It includes measurements like glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age, alongside the outcome of whether or not the individual has diabetes. The `Outcome` variable is binary, with more non-diabetic individuals than diabetic ones in the dataset. The data also shows significant variability in the health indicators, reflecting the diversity of the population.

The study revealed “500” instances where the `Outcome` is `0`. Thus, considering the diabetes prediction dataset, an `Outcome` of `0` usually signifies that the row represents a person who is “not diabetic”. Similarly, there are “268” instances where the `Outcome` is `1`. Thus, an `Outcome` of `1` typically means that the person is “diabetic.” For instance, this information could be used to examine how many rows within your dataset are considered to be “diabetic” or “not diabetic”; this can help you better understand the balance of your dataset, which is a helpful guideline when training or evaluating machine learning models.

Standardization, or Z-score normalization, ensures that data has a mean of 0 and a standard deviation of 1. This is helpful for many machine learning algorithms that are sensitive to the scale of the data, such as gradient descent-based models. The standardized values of the dataset were shown in the study. Thus, each value in the array is a standardized score (or Z-score) for the given feature and instance in the dataset. Positive values indicate that the original value is above the mean of that feature. 0.63994726 means that the original value was 0.64 standard deviations above the mean. The negative values indicate that the original value is below the mean. For example, `-0.84488505` means the original value is 0.84 standard deviations below the mean. Thus, values Close to 0 suggest that the original value is close to the mean. For instance, `0.00330087` is almost at the mean value of that feature.

The author [11], revealed that standardization helps ensure that all features contribute equally to the model by bringing them to a standard scale. It improves the performance of many algorithms, such as logistic regression, k-nearest neighbors, and support vector machines [1]. It also helps achieve faster convergence during training, particularly for gradient-based optimization algorithms [9].

The study revealed that the X.shape was `(768, 8)` – The original dataset `X` has 768 samples and eight features. Thus the `X_train.shape`: `(614, 8)` – The training set `X_train` consists of 614 samples (80% of the original data) and retains eight features. The `X_test.shape`: `(154, 8)` – The test set `X_test` consists of 154 samples (20% of the original data) and retains eight features. Thus, the dataset was split into training and testing subsets. The training set contains 80% of the original data, and the test set contains 20%. The `stratify=Y` parameter ensures that the distribution of the target variable `Y` is the same in both training and testing sets. The `random_state=2` parameter ensures that the split is reproducible.

A Support Vector Machine (SVM) with a linear kernel is used. This type of SVM attempts to find a linear decision boundary between classes. The model is trained using `X_train` (features) and `Y_train` (target labels). The SVM model is trained on the training data. The linear kernel is used, meaning the decision boundary the model learns will be linear.

The accuracy_score measures how many predictions made by the model are correct compared to the actual labels `Y_train.` The accuracy score of approximately 0.79 (or 78.66%) means that the model

correctly predicted the outcomes for about 79% of the training data. This indicates the model's performance based on the data on which it was trained.

Limitations

However, applying machine learning to predict diabetes is not straightforward, and there are some problematic statistical issues to consider. These include ensuring that datasets are of sufficient quality and represent the population, choosing the most appropriate algorithms, and interpreting results in a clinically meaningful way.

A significant limitation of the study is that it had a limited subset of risk factors as input features, which cannot reflect the interactions among different biological systems implicated in T2DM. Furthermore, the source of the data was reliant on previous literature for the selection of their predictors, which may restrict the scope of the model and neglect novel or less-studied predictors or aspects of the pathogenesis of T2DM.

Conclusion

Diabetes Mellitus (DM) is a chronic condition due to chronic high blood glucose levels caused by relative or absolute insulin deficiency. The incidence of diabetes has risen significantly worldwide, from mainly developed countries to populations of different groups of continents – like Africa and Asia. This is due to the excessive consumption of unhealthy foods and a sedentary way of living. AI could improve diabetes management by providing real-time health information to patients or providers, facilitating patient self-management, and enhancing intervention targeting high-risk populations. The dataset size was 768 samples, and the features included `Pregnancies, ` `Glucose, ` `Blood Pressure, ` `Skin Thickness, ` `Insulin, ` `BMI, ` `Diabetes Pedigree Function, ` and `Age` The outcome distribution include: non-diabetic: 500 cases and diabetic: 268 cases. Glucose had a mean = 69.11 and range = 0 to 199. The Body Mass Index had a mean = 31.99 and a range = 0 to 67.10. The mean age was 33.24, and the range = was 21 to 81. The features were standardized using `StandardScaler` with a mean of 0 and a standard deviation of 1. The data was split into training (70%) and testing (30%) sets. The model used a Support Vector Machine (SVM) with a linear kernel. The model performance was Training Accuracy 78.66% and Test Accuracy 77.27%.

References

1. Dankwa-Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2019). Transforming diabetes care through artificial intelligence: the future is here. *Population health management*, *22*(3), 229-242.
2. Guan Z, Li H, Liu R, Cai C, Liu Y, Li J, Wang X, Huang S, Wu L, Liu D, Yu S, Wang Z, Shu J, Hou X, Yang X, Jia W, Sheng B. (2023). Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. Cell Rep Med. 17;4(10):101213. doi: 10.1016/j.xcrm.2023.101213.
3. Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., & Yoneda, T. (2021). Artificial intelligence in current diabetes management and prediction. *Current Diabetes Reports*, *21*(12), 61.
4. Mohsen, F., Al-Absi, H. R. H., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digital Medicine, 6*(1), 197. https://doi.org/10.1038/s41746-023-00933-5
5. World Health Organization. (2021, November 10). *Diabetes fact sheet*. Retrieved May 7, 2024,

from https://www.who.int/news-room/fact-sheets/detail/diabetes

6. Mackenzie, S. C., Sainsbury, C. A. R., & Wake, D. J. (2024). Diabetes and artificial intelligence beyond the closed loop: A review of the landscape, promise, and challenges. *Diabetologia, 67*(2), 223–235. https://doi.org/10.1007/s00125-023-06038-8
7. Rigla, M., García-Sáez, G., Pons, B., & Hernando, M. E. (2018). Artificial intelligence methodologies and their application to diabetes. *Journal of diabetes science and technology*, *12*(2), 303-310.
8. Gautier, T., Ziegler, L. B., Gerber, M. S., Campos-Náñez, E., & Patek, S. D. (2021). Artificial intelligence and diabetes technology: A review. *Metabolism, 124*, 154872. https://doi.org/10.1016/j.metabol.2021.154872
9. Kaul, S., & Kumar, Y. (2020). Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. *SN Computer Science*, *1*(6), 322.
10. Yuk, H., Gim, J., Min, J. K., Yun, J., & Heo, T.-Y. (2022). Artificial intelligence-based prediction of diabetes and prediabetes using health checkup data in Korea. *Applied Artificial Intelligence: An International Journal, 36*(1), 2145644. https://doi.org/10.1080/08839514.2022.2145644
11. Buchanan, C., Howitt, M. L., Wilson, R., Booth, R. G., Risling, T., & Bamford, M. (2020). Predicted influences of artificial intelligence on the domains of nursing: Scoping review. *JMIR Nursing, 3*(1), e23939. https://doi.org/10.2196/23939
12. Wadhwa, S., & Babber, K. (2020). Artificial intelligence in health care: predictive analysis on diabetes using machine learning algorithms. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20* (pp. 354-366). Springer International Publishing.
13. Wang, S. C. Y., Nickel, G., Venkatesh, K. P., Raza, M. M., & Kvedar, J. C. (2024). AI-based diabetes care: Risk prediction models and implementation concerns. *npj Digital Medicine, 7*(1), 36. https://doi.org/10.1038/s41746-024-01034-7
14. Ziajor, S., Tomasik, J., Sajdak, P., Turski, M., Bednarski, A., Stodolak, M., Szydłowski, Ł., Żurowska, K., Krużel, A., Kłos, K., & Dębik, M. (2024). The use of artificial intelligence in the diagnosis and detection of complications of diabetes. *Journal of Education, Health and Sport, 65*, 11-27. https://doi.org/10.12775/JEHS.2024.65.001
15. Ellahham, S. (2020). Artificial intelligence: the future for diabetes care. *The American Journal of Medicine*, *133*(8), 895-900.