Comparative Analysis of varied ML algorithms for Diabetes dataset

Sanjana V. Asrani, Karina K., Roshni W., and Simran L.

***Abstract*—Diabetes Mellitus, a chronic metabolic disorder, poses a significant global health challenge impacting millions of people worldwide. In India, diabetes has reached alarming proportions, affecting an estimated 77 million people, as reported by the International Diabetes Federation. Over recent decades, there has been a consistent rise in both the incidence and prevalence of diabetes. To address this critical issue, early detection, prediction and post-diagnosis care and management of diabetes are important. This research focuses on developing a predictive model for diabetes risk identification that fits the PIMA Indian Diabetes Dataset from the National Institute of Diabetes and Digestive and Kidney Diseases accurately by exploring multiple algorithms with varied parameters.**

***Keywords***—Diabetes, PIMA dataset, ML, XAI

1. INTRODUCTION

Diabetes, a multifaceted metabolic disorder, arises from a combination of genetic predisposition, lifestyle choices, and environmental factors. This chronic condition disrupts the body's ability to regulate blood sugar levels effectively. Insufficient insulin production or decreased insulin sensitivity leads to elevated blood sugar levels, a condition known as hyperglycemia. Long-term hyperglycemia can lead to cardiovascular disease, kidney damage, nerve dysfunction, and vision problems. Thus making Early detection, prediction and post-diagnosis care and management of diabetes a very crucial task. Approach:

1. Prediction into categories like Diabetic, Non Diabetic, Pre Diabetic using clinical Datasets.
2. Probabilistic Range of Risk/Health.
3. Type1, Type2 Identification using the Pathological Report Datasets
   1. OVERVIEW
   2. Motivation

In the face of surging diabetes cases, particularly in India, a diabetes prediction and detection project fueled by machine learning emerges as a critical intervention. Rapid urbanization, sedentary lifestyles, and genetic predispositions contribute to escalating risks. Early diagnosis through our ML model becomes a beacon of hope, offering timely interventions to curb the rising

tide of diabetes-related complications. By addressing the unique challenges prevalent in India's healthcare landscape, we strive to empower individuals, alleviate strain on healthcare infrastructure, and foster a paradigm shift towards proactive diabetes management.

* 1. Problem Statement

The aim of this project is to address the challenges associated with diabetes risk prediction and management. Using traditional physical assessments Long queues, limited appointment availability, high costs, and potential geographic barriers hinder timely access to healthcare professionals. This project aims to overcome these challenges by developing an AI-powered application that offers virtual diabetes risk assessment and personalized insights, providing a convenient, cost-effective, and accessible solution for individuals seeking to understand and manage their diabetes risk.

* 1. Objectives
     + Diabetes risk prediction and management
     + Overcome disadvantages of traditional physical assessments: Long queues, limited appointment availability, high costs, and potential geographic barriers hinder timely access to healthcare professionals.
     + virtual diabetes risk assessment and personalized insights.
  2. LITERATURE REVIEW
  3. Survey of Existing Systems

Authors Sivaranjani S et al. presented a study at the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), focusing on predicting diabetes-related diseases. They utilized Support Vector Machine (SVM) and Random Forest (RF) on the Pima Indian diabetes dataset from the UCI repository. Results showed RF outperformed SVM, achieving an 83% prediction accuracy after feature selection and dimensionality reduction, contributing to the literature on diabetes prediction and demonstrating the effectiveness of machine learning techniques [1]. S. Saru et al. used WEKA software and the Pima Indian diabetes dataset. Employing Decision Trees, Naïve Bayes, and k-Nearest Neighbors, the study achieves high accuracies (e.g., 94.4% for Decision Trees). Bootstrapping improves accuracy, particularly for k-NN (k=1), from 69.93% to 93.79%. The research suggests that ensemble methods enhance prediction

performance, achieving an overall accuracy of 90.36%[2]. In their paper presented at the 24th International Conference on Automation and Computing, Muhammad Azeem Sarwar et al. discuss predictive analytics in healthcare. They apply six machine learning algorithms, including KNN and SVM, to the PIMA dataset for diabetes prediction. SVM and KNN stand out with the highest accuracy of 77%, highlighting their effectiveness in healthcare applications [3]. In their paper on predicting Type-2 Diabetes Mellitus, Ahamed BS et al. employ various machine learning classifiers, including logistic regression, XGBoost, and LGBM. Using the PIMA Indian Dataset, LGBM stands out with the highest accuracy of 95.20%, outperforming other algorithms. The study suggests avenues for future research, including exploring different datasets and considering advanced versions of LGBM for enhanced prediction accuracy [4]. In their 2023 paper titled "Diabetes prediction using machine learning and explainable AI techniques," Tasin et al. developed an automatic diabetes prediction system using a private dataset of female patients in Bangladesh and the PIMA dataset. They employed a semi-supervised model with extreme gradient boosting to predict insulin features, addressing class imbalance with SMOTE and ADASYN. Various machine learning classification methods were explored, with XGBoost and ADASYN yielding the best results, achieving 81% accuracy, 0.81 F1 coefficient, and an AUC of 0.84. The authors implemented explainable AI techniques, specifically LIME and SHAP frameworks, to understand the XGBoost model's predictions. Additionally, they created a website and an Android application for instantaneous diabetes prediction [5].In the study by Kirti Kangra et al. on predictive machine learning algorithms for diabetes mellitus, various algorithms, including SVM, Naïve Bayes, KNN, RF, LR, and DT, were evaluated using the Pima Indian diabetic and Germany diabetes datasets. Using WEKA 3.8.6, SVM achieved 74% accuracy for PID, while KNN and RF performed better with 98.7% accuracy for the Germany dataset. The study suggests Logistic Regression (LR) as a preferable choice for both datasets based on accuracy and ROC area. Future work may involve exploring hybrid models and assessing their performance on real-time data [7]. Muhammad Exell Febriana et al. (2020) compared K-Nearest Neighbor (KNN) and Naive Bayes algorithms for diabetes prediction using supervised machine learning. Evaluating the PIMA dataset, they found Naive Bayes outperformed KNN with an average accuracy of 76.07%, precision of 73.37%, and recall of 71.37%,

while KNN achieved 73.33%, 70.25%, and 69.37% respectively. The study suggests future research incorporating neural networks, Particle Swarm Optimization, and practical application development to

enhance accuracy and precision [8]. Md Shahin Ali et. al developed the RFWBP algorithm for early diabetes detection using the PIMA dataset. Achieving a remarkable 95.83% accuracy with cross-validation and 90.68% without, RFWBP outperformed conventional methods (AdaBoost, SVM, logistic regression, naive Bayes, multilayer perceptron, and regular Random Forest). The study utilized data processing and mining techniques to enhance the dataset. Future research aims to broaden the analysis by incorporating more subjects and diverse datasets for precise diabetes identification [9].Umair Muneer Butt et al. present a machine learning approach for diabetes classification and prediction using Random Forest, Multilayer Perceptron, and Logistic Regression. The study, based on the PIMA Indian Diabetes dataset, shows MLP achieving 86.08% accuracy for classification, while LSTM significantly improves prediction with 87.26% accuracy. Kopitar et al. compared machine learning (Glmnet, RF, XGBoost, LightGBM) and traditional regression models for early detection of Type 2 Diabetes Mellitus. Using 6 months of data, a simple regression model had the lowest RMSE at 0.838. Glmnet showed the highest improvement rate (+3.4%) with more data. LightGBM demonstrated stability in variable selection. While machine learning didn't significantly outperform regression models, all methods improved with additional data. The study suggests exploring ensemble methods in future research but warns about challenges in interpreting results for healthcare decisions [12].

* 1. Limitations and Research Gap

The limitations highlight the need for further research in several key areas. One critical area is the diversification of datasets to enhance the generalizability of predictive models across various populations. Additionally, the exploration of emerging algorithms and hybrid models, as well as the systematic evaluation of ensemble methods and alternative classifiers, holds promise for improving predictive accuracy. Furthermore, the integration of explainable AI techniques and the development of real-time monitoring and IoT applications represent promising avenues for enhancing the interpretability and practicality of diabetes prediction models. Finally, cross-disease evaluation, long-term predictive analytics, and considerations for ethnic and gender differences should be prioritized to broaden the applicability and effectiveness of machine learning-based diabetes prediction systems.

* 1. Project Contribution

Our project makes a significant contribution to diabetes risk prediction by incorporating a multifaceted approach

to data collection. Inputs are sourced from users, capturing vital information like pregnancies, glucose levels, blood pressure, and other pertinent features, as well as potential user interface data. The inclusion of model input, derived from a diverse dataset such as PIMA and Sylhet, enhances the generalizability of our predictive models, addressing a key limitation in existing research. We employ a variety of machine learning algorithms, including Decision Tree, Support Vector Machine, Random Forest, Adaboost, and XGBoost, to ensure a comprehensive evaluation of diabetes risk. The transparency of our predictions, coupled with detailed explanations provided to the user, facilitates informed decision-making regarding diabetes risk management. By utilizing the PIMA dataset, and the Sylhet dataset from Bangladesh, our system caters to diverse demographic groups, contributing to the inclusivity of diabetes prediction models. Overall, our project not only advances the accuracy of diabetes risk prediction but also addresses the limitations of existing research by incorporating diverse datasets and fostering user understanding through detailed explanations.

* 1. Dataset Description

Following two datasets have been used:

# PIMA Dataset:

The dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases and serves the purpose of predicting whether a patient has diabetes based on specific diagnostic measurements. Stringent criteria were applied when selecting instances from a broader database, focusing exclusively on female patients aged at least 21 years with Pima Indian ancestry. These datasets encompass various medical predictor variables alongside a single target variable, Outcome. Predictor variables encompass factors such as the patient's number of pregnancies, BMI, insulin levels, age, and others.

**Table 1.** Pima Dataset Features

|  |  |  |  |
| --- | --- | --- | --- |
| Age | 21.0 | 81.0 | 33.24 |

# Sylhet dataset:

Diabetes stands as one of the most rapidly spreading chronic, life-threatening conditions, having impacted an estimated 422 million individuals globally, as reported by the World Health Organization (WHO) in 2018. Given its lengthy asymptomatic period, detecting diabetes at an early stage is crucial for achieving meaningful clinical outcomes. Alarmingly, approximately 50% of diabetes cases remain undiagnosed due to this prolonged asymptomatic phase. The dataset under discussion comprises 520 observations featuring 17 distinct characteristics. These data points were gathered through direct questionnaires and diagnostic findings from patients treated at the Sylhet Diabetes Hospital in Sylhet, Bangladesh.

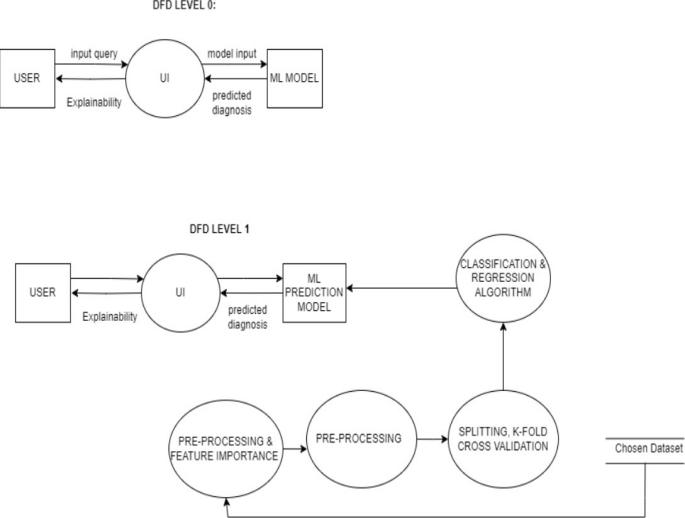
**Table 2.** Sylhet Dataset Features

|  |  |
| --- | --- |
| **Features** | **Meaning** |
| age | Age of the patient |
| gender | Gender of the patient (Male/Female) |
| polyuria | Excessive urination |
| polydipsia | Excessive thirst |
| sudden\_weight\_ loss | Rapid, unintentional weight loss |
| weakness | Generalized weakness |
| polyphagia | Excessive hunger |
| genital\_thrush | Fungal infection in genital area |
| visual\_blurring | Unclear vision |
| itching | Skin itching |
| irritability | Feeling easily annoyed or frustrated |
| delayed\_healing | Slow healing of wounds |
| partial\_paresis | Partial loss of voluntary muscle movement |
| muscle\_stiffness | Stiffness or discomfort in muscles |
| alopecia | Loss of hair |
| obesity | Excessive body weight |

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Min value** | **Max Value** | **Avg Value** |
| Pregnancies | 0.0 | 17.0 | 3.85 |
| Glucose | 0.0 | 199.0 | 120.89 |
| BloodPressure | 0.0 | 122.0 | 69.11 |
| SkinThickness | 0.0 | 99.0 | 20.54 |
| Insulin | 0.0 | 846.0 | 79.8 |
| BMI | 0.0 | 67.1 | 31.99 |
| DiabetesPedigreeF unction | 0.078 | 2.42 | 0.47 |

* 1. IMPLEMENTATION
  2. Detailed Design

User provides information which is input to the model. It requires all the features of pima dataset to be inputted.The Algorithm used here finally is *Random Fores*t for deployment while *Decision Tree*, *SVM* (for non-linear data), *XGBoost* and *ADABoost* were also implemented along with the *XAI* techniques i.e. *LIME* and *SHAP*. It predicts whether the user is diabetic or not using ML techniques. The explanation of the prediction is given in XAI techniques to help user make informed decisions on how to manage their risk of developing diabetes.



between classes by transforming input data using different kernel functions.

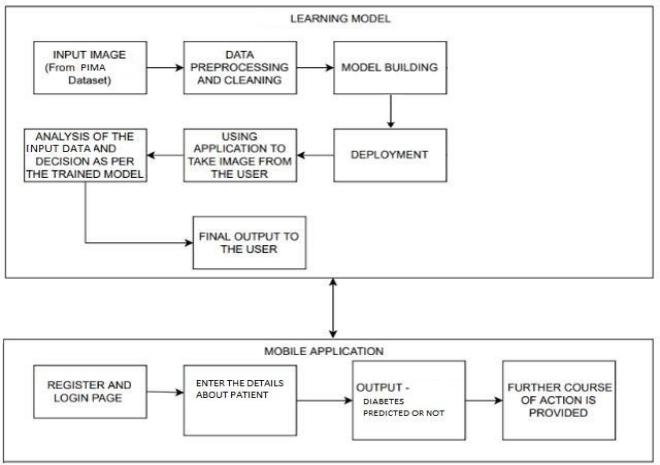
# XGBoost (Extreme Gradient Boosting):

XGBoost, a form of gradient boosting, iteratively builds decision trees to rectify errors made by preceding trees. It uses a gradient descent algorithm to minimize the loss function.

# AdaBoost (Adaptive Boosting):

AdaBoost combines multiple weak learners (simple models that perform slightly better than random guessing). It allocates weights to instances, prioritizing those incorrectly classified by preceding models. The final prediction is a combination of these weighted weak learners.

* 1. Modular Diagram



**Fig 2.** Modular Diagram

Algorithms executed:

# Decision Tree:

**Fig 1.** DFD

The modular diagram explains the model as two sections. A user-interface will collect all the required inputs from the user that will be given to the model. Based on the input, the model predicts whether the user

Decision trees iteratively partition the data based on provided features to form a hierarchical structure comprising nodes and leaves. At each node, the algorithm selects the feature that best separates the data into pure classes (maximizes information gain or minimizes impurity) depending upon this. The process persists until a stopping criterion is satisfied, such as reaching a predefined tree depth or attaining a minimum number of samples in a leaf node, as predetermined.

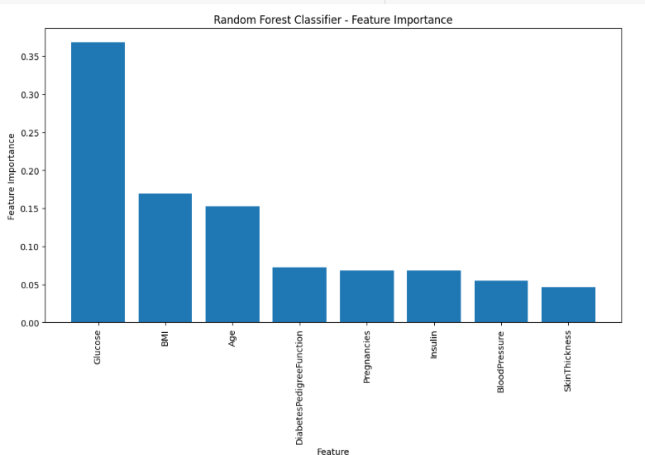
# Random Forest:

Random Forest builds multiple decision trees by using random subsets of both the data and the features given. The final prediction is an average (classification) or a mean (regression) of the predictions of individual trees depending upon the data.Random Forest introduces diversity and reduces overfitting by combining the predictions of multiple trees.

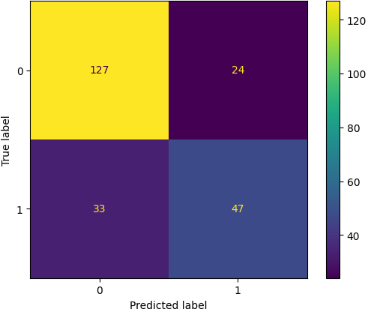
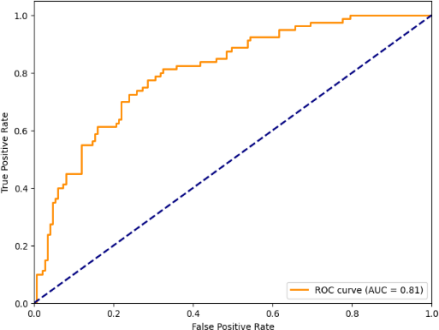
# Support Vector Machine (SVM):

SVM finds the hyperplane that best separates classes in a high-dimensional space.It maximizes the margin

is diabetic or not.Both the sections are combined together for proper functioning of diabetes prediction and detection.



**Fig** 3. Feature Importance Plot for PIMA Dataset

**Fig 4.** Confusion Matrix for RF, PIMA Dataset **Fig.** 5. ROC AUC Curve for RF, PIMA Dataset

There are a number of classification algorithms that help determine the appropriate class. While developing the model, we implemented various algorithms to determine the accuracy of each algorithm.The Random Forest algorithm gave the best accuracy for the dataset chosen.

* 1. RESULTS AND EVALUATION

There are various evaluation measures that are used for determining the performance of a model. Some of them are: Accuracy, precision, recall, F1-score, etc.

**Accuracy:** It is a metric representing the proportion of correct predictions made by a model in relation to the total number of input samples.

Accuracy=(TP+TN)/(TP+FP+TN+FN)

**Confusion matrix:** A confusion matrix offers a summary of a machine learning model's performance on a test dataset, particularly for classification tasks. It displays key metrics like true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

**F1 score:** The F1 score is a balanced metric for assessing the accuracy of a test, calculated as the harmonic mean of precision and recall in classification tasks.

F1 score= 2\*precision\*recall/precision+recall

**Precision:** This quantity represents the ratio of correct positive predictions made by the classifier to the total number of positive instances predicted by the classifier. Precision=TP/(TP+FP)

**Recall:** It is the ratio of correct positive predictions to all relevant samples, which include all instances that should have been identified as positive.

Recall=TP/(TP+FN)

Accuracy Comparison for all 5 algorithms:

PIMA:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SVM(  non-lin ear: rbf kernel ) | Deci sion tree | Rand om Fores t | Adabo ost | XGBo  ost |
| 80:20 | 75% | 75% | 76.6 | 73.37 | 73.37 |
|  |  |  | % | % | % |
| 70:30 | 74% | 70% | 75.3 | 74.45 | 73.59 |
|  |  |  | % | % | % |
| 5-fold | 76.17% | 73% | 85.7 | 80.52 | 76.43 |
|  |  |  | % | % | % |
| 10-fol | 76.43% | 71% | 80.2 | 84.21 | 76.17 |
| d |  |  | % | % | % |

**Table 3**. All Algorithms accuracy comparison wrt PIMA dataset

The sylhet dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SVM | Deci sion tree | Rand om Forest | Adabo ost | XGBoo st |
| 80:20 | 89% | 96% | 100% | 93.26% | 99.03% |
| 70:30 | 94% | 97% | 96.7  % | 92.94% | 98.71% |
| 5-fold | 91.15% | 96% | 100% | 90.38% | 96.35% |
| 10-fol d | 92.31% | 97% | 100% | 94.23% | 96.92% |

**Table 4**. All Algorithms accuracy comparison wrt the sylhet dataset

As per our evaluation, Random Forest Algorithm works best on both the datasets. However we observe the accuracy in the second dataset is 100% which shows that the second dataset is overfitting.

* 1. CONCLUSION

This study compares machine learning algorithms (Decision Tree, Random Forest, SVM, XGBoost, AdaBoost) using PIMA and Sylhet datasets. It assesses accuracy metrics with varied data splits and cross-validation. It also employs XAI techniques (LIME, SHAP) for model interpretability. However, limitations are encountered, such as the inability to create a robust model from the collected reports of diabetic patients due to variations in patient reports. Moreover, the Sylhet dataset demonstrates 100% accuracy, indicating potential overfitting issues. Future work includes integrating hardware for voltage measurement to estimate glucose levels accurately and enhancing voltage accuracy by adjusting parameters.

* 1. REFERENCES

1. Book Referred

[1].Soumya K N, Vigneshwaran P, “Prediction on Type-2 Diabetes Mellitus Using Machine Learning Methods”, *Volume 41: Advances in Parallel Computing Algorithms, Tools and Paradigms, 2022.*

Available:<https://ebooks.iospress.nl/volumearticle/61506>

1. Research Papers referred/ Journals/ Articles referred

[1]Sivaranjani S, Ananya S, Aravinth J, Karthika R, “Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”, *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS),* March 2021.

Available:<https://ieeexplore.ieee.org/document/9441935>

[2]S.Saru, S.Subashree, “ Analysis and Prediction of Diabetes Using Machine Learning”, *International Journal of Emerging Technology and Innovative Engineering Volume 5, Issue 4,* April 2019.

Available: [https://ssrn.com/abstract=3368308](https://ssrn.com/abstract%3D3368308)

1. Muhammad Azeem Sarwar, Nasir Kamal,Wajeeha Hamid; “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”; *24th International Conference on Automation and Computing (ICAC)*: 06-07 September 2018.

Available:[https://ieeexplore.ieee.org/abstract/document/8748992/auth](https://ieeexplore.ieee.org/abstract/document/8748992/authors#authors) [ors#authors](https://ieeexplore.ieee.org/abstract/document/8748992/authors#authors)

1. Ahamed BS, Arya MS and Nancy V AO , “Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques”, *Front. Computer. Sci. 4:835242. doi: 10.3389/fcomp.2022.835242,* 2022.

Available:[https://www.frontiersin.org/articles/10.3389/fcomp.2022.83](https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242/full) [5242/full](https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242/full)

1. Tasin, I., Nabil, T.U., Islam, S., Khan, R., “Diabetes prediction using machine learning and explainable AI techniques”, *Healthc. Technol. Lett*. 10, 1–10 , 2023.

Available:[https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/](https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/htl2.12039) [htl2.12039](https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/htl2.12039)

1. Chang, V., Bailey, J., Xu, Q.A. *et al.* ,“Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms”, *Neural Comput & Applic* 35, 16157–16173 (2023).

Available:[https://link.springer.com/article/10.1007/s00521-022-07049-](https://link.springer.com/article/10.1007/s00521-022-07049-z) [z](https://link.springer.com/article/10.1007/s00521-022-07049-z)

1. Kirti Kangra, Jaswinder Singh, “Comparative analysis of predictive machine learning algorithms for diabetes mellitus “, *Bulletin of Electrical Engineering and Informatics Vol. 12, No. 3, June 2023.*

Available:<https://www.beei.org/index.php/EEI/article/view/4412>

1. Muhammad Exell Febriana, Fransiskus Xaverius Ferdinana , Gustian Paul Sendani, Kristien Margi Suryanigrum, Rezki Yunanda “ Diabetes prediction using supervised machine learning ” , *7th International Conference on Computer Science and Computational Intelligence*, *2020*.

Available: [//pdf.sciencedirectassets.com/280203/](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050922X00197/1-s2.0-S1877050922021858/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEK3%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQDiR%2BvVna7sWtMNVyBSY7lHy7e3bwHIPatBjgrpNkhxdAIgQX%2B2F7n81t4rqqYlfU1hUwV06iMZZiPDjXLTPCOk7moqswUIdhAFGgwwNTkwMDM1NDY4NjUiDPD7ulw8cAiBCHrBfCqQBTt1koFd7wvINXRTfB9%2BiOaHHtjaxzG12ljURdYH%2F1r2lD6YyEyI1lsUSOmLDyeAcCN%2FHOwRnoW1dA8P5NZuea39HyPgRIZAPOgJ8NW%2BhK%2FiG%2FGRKkxVnrAgkoJ19QtlGf1xU1AerXCHoce83JMAE5yZCafO7Qb6qNtI0sAB%2BpDCBUZfaVbj2aQaOwbCVEczSBEJvhvZnQYI5c2BfLDeINYPQcJSmMk8SUJaEOfQdWwSKAjQ8wEMXxyexnHH1mCcpdWT2jKnQaMu%2BK91B65JkyO9E0k7%2BowW%2FiN21CypeY7uKwZvcNF0BC5kSDyuOtmMbf92JN3Uaf9Db3tM0RbNKDuKznKdJdiu7DUf2kalWEV0WxZVPPJqGM3U5gwLm7lVXyXGMqEFMXGt7Fi7zqQqi3w3SJLVzOxfjVmKEnTdoiPoWCGFX8pu8yDuB4rV9%2BvRD1y9Lkc6TALMLFZGrOwRPKIypD785MxtpQiXqDEkQ3Upij7EBeHjc1ukhrW8KLhvEeusLf2oXgNGMzWefnc3NrqJNhrPicXueZfwbnlTXRbVLr5rJwMItboclEwt789L6Nmq%2FHHifn7sBl3rY6Yy5B9ZpYJNaCuHQtuvcoT5fauTX%2F%2FZoN1Z2adeOAlOWUC4IlwovsMCx2rsaAzrJqbyyWuU%2Fi5SVqd8mWkOK0cQ%2BRD7P%2BxTFkVpKWKoqYvSo8DsFYkZydg%2FIGh2sKkMWO4zPqSXj1CuAlavoZeNCX%2B2AWaYQWA8SjqNEFom8U6ArusiALNn4SKQD67c%2FqO2zW6QZMoCpTmx5uwBE%2B4fjSQreezXXK6ETtWkfERvFpldsdHGKQMg7mhkSFgYzBTWeutaMXcwwDZzXWXx8SFaP%2FHo%2B9ZcMMmJracGOrEBVQVfZa4rCr5dHLOuopOs5R2M3zKNjDPSU5ENzfunljXiYvzbUGoSWHHSX8l3ho%2BicCRjiSWQsiPyRf46mkmNl1PcTqe6Ckml%2Fvt1gW7omffH327m7yxrr9adqpcpIpeH15FtJo7hlGUG%2B%2Fp0WCvatLTiv%2FXTb86I9WEeUDgjaqnTEfqUqH%2BuE2q8O0t%2FUaFwBf6kxlG00BnPvIUjritp%2FarWFXx0VIYwatKJiOFDKkx4&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230827T125908Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYSJIHJVLY%2F20230827%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=be7762148e054ecd7f68d3aab17301487e30d723dc88caa784b912ad653a108e&hash=557dc406a605886f99ea61e24017c2e4725acd355388fa1615135bc8f498be87&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050922021858&tid=spdf-6bd91c46-17b3-4a94-aaf8-ef8e66e7808d&sid=cd59b35d4cf74348b31b04e739a946edf7a6gxrqb&type=client&tsoh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&ua=13085705005a500a525603&rr=7fd47d71af916ec2&cc=in)

1. Md Shahin Ali , Md Khairul Islam , A. Arjan Das , D. U. S. Duranta , Mst. Farija Haque , and Md Habibur Rahman , “A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights”, *Hindawi BioMed Research International Volume , 2023*.

Available:<https://www.hindawi.com/journals/bmri/2023/8583210/>

1. Alain Hennebelle , Huned Materwala, Leila Ismail, “HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System”, *The 14th International Conference on Ambient Systems, Networks and Technologies (ANT) , March 15-17, 2023*.

Available: [pdf.sciencedirectassets.com](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050923X00040/1-s2.0-S1877050923005781/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEK7%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIGg5u7m6W0zDwU70IW3KWbneSatpIYawP%2BS%2BxU9cLMkLAiALE4%2BWROrBy7lq6TgOqAqlToJBLwwZ%2FNAhy2rWFRI7XCqzBQh3EAUaDDA1OTAwMzU0Njg2NSIM6%2BaRHw7J7kNrs5snKpAFUs1wPq8MWV%2BjYKXP1p0%2FKd7%2F58nZnsYnx0Dj2tgNNbHylcIRu%2Ba4b13KkSzOUFVaJCob%2BaicL4W3Uv%2F1X6nPejnV9sHHhPWa%2BH%2FuxgpLn1g9aNuPMvkNk2rbCUIVGK8m4XLJFCMhMTmUkpXgIRsrKmwrg9aqiYkur5M0GrZfJdj0CRiR7GOHR%2Fx274qGHINkk%2FshHa7GA%2BsWk8hPoCqyszt8PGs9Ubz%2F0tpX7w6T6xfr9mK7LvtB6qbIa1Tjy4SLJdSy3%2FeWf5r6Icm4qxdpYfG0OKZhVvixjL0Qd5mCZWS0C0mDTAG51XX6615jVvuZhwmS0l%2B1o6xSdsXZHVKdDTL5UfUaj6NMaZ6PIcyml8a3fomWfNLQPRWKmbFCJTz9Eszz1l%2BJNyZuV59tNwsPUzfTcCIiumizBGkBtNAQnVwO0zLUTbq%2BC93%2FMNsIQJxN61vMVHbE7dY5a8JoW8BddFaMwV49qHu%2BYsv6Q%2B0oQgvxjDI3NiMTM0czrvVWnI3aSfzHlwTlRWEXKCTcvYJRtUr1LgJ%2FDZxgVPjrOLQIvXRyMgBRc%2FOlnF%2FmRAYgaW5Xxli7hDM5DF0MuTd2nEVa1Yt57tEjgPkBr3XFgYI1Bc%2FWRDJZfKETlg%2BCOpxVoHPKYGyjp9wAs%2BOYgVB3ohi71GIxuF4dEGyHRjaGD125X9fnEE5gbA5Ptyl3YxZ%2BMTvS0rWxk5DM7%2BLQ1M9%2FMKevkj1S3GFoIpsp%2B2VHoYOAnn8%2BYO6nX7Z9UygV4qn48rLIeC1b8EnTM9%2BVb%2Fh2Gl8y3K3rnqujo63BzjgOgi4xuAqpnUOUfXhCJg56Ri2d3gikh%2FtW9uwRLZhUMfXwV8oL2fdneIti71uS24Gv2D%2BAcBAw7aStpwY6sgEYsqS94eGvEVsngZfYdx3%2F%2B4nfRHul0pm79EBcYbtlf%2F6PFegRi78ZEnj%2BVKPQnP%2B7H9OguykbKZlJ6nfrOyS0KVBJ9Rr6KN7RLJIggCw4%2FgTwzjLZgeyGEVxXyi1Wm%2B%2FuEmNs9TeuB8Iug5vPGy3Jww6jlcLi8mpPkIbMFVXNvGqJrPZxtYnEYnTxYaZCBBid4ZnANt3l5aKhk5bryB3bLQo6XER3mzAR7CM8vq3AEkdu&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230827T140226Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY63JA427P%2F20230827%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=ea950e7140bc63802a8ebcd50c75b4659eeb0cad5a852dded3a9e8fd843a29f7&hash=74396542bbb37620a495a5d81c15aa19e5fd1fc7b748155bebfb7bfaa16844c6&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050923005781&tid=spdf-e700d030-f49c-4c17-b14d-02a79c1c3fbc&sid=cd59b35d4cf74348b31b04e739a946edf7a6gxrqb&type=client&tsoh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&ua=13085705005a030f575e04&rr=7fd4da29f9e885f7&cc=in)

1. Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, and Hafiz Husnain Raza Sherazi, “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications”, *Journal of Healthcare Engineering-Hindawi, Volume 2021.*

Available:<https://www.hindawi.com/journals/jhe/2021/9930985/>

1. Kopitar, L., Kocbek, P., Cilar, L. *et al.*, “Early detection of type 2 diabetes mellitus using machine learning-based prediction models”. *Sci Rep* 10, 11981 (2020).

Available:<https://www.nature.com/articles/s41598-020-68771-z>

[13]Ashwini Tuppad , Shantala Devi Patil, “Machine learning for diabetes clinical decision support: a review”, *Advances in Computational Intelligence , 2022.*

Available:[https://link.springer.com/article/10.1007/s43674-022-00034-](https://link.springer.com/article/10.1007/s43674-022-00034-y) [y](https://link.springer.com/article/10.1007/s43674-022-00034-y)