Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1131

Deep Learning Algorithms for Diabetes Detection: A Comprehensive Exploration and Evaluation

Ayesha Saif[1*], Talia Noreen[2], Samina Bibi[3], Muhammad Azam Hussain[4], Ishrat zubair[5], and Muhammad Umair Bin Yaseen[6]

[1]Department of Informatics and Systems, University of Management and Technology, Lahore, 5700001, Pakistan.
[2]Department of Computer Science, Beaconhouse National University, Lahore, 570002, Pakistan.
[3]Department of Computer Science, University of Management and Technology, Lahore, 570003, Pakistan.
[4]Department of Computer Science & Information Technology, University of Lahore, 570004, Pakistan.
[5]Department of Computer Science, Islamia University of Bahawalpur, Pakistan.
[6]Department of Computer Science, University of South Asia, Lahore, 570006, Pakistan.
[*]Corresponding Author: Ayesha Saif. Email: ayeshawraich51@gmail.com

Received: September 13, 2025 **Accepted:** November 21, 2025

Abstract: Early diagnosis of diabetes can lead to early interventions, lifestyle modifications, and personalized treatment plans that can positively impact patient health outcomes and reduce the burden on healthcare systems. Early detection reduces the risk to the patient's health. Diabetes is spreading rapidly all over the world, and the majority of people over the age of 45 are victims. Therefore, it is important to detect this serious disease as soon as possible. Traditional diabetes screening methods often involve regular blood tests and clinical evaluations, which may not always detect diabetes in its early stages or identify individuals at high risk of developing the condition. A deep learning model is useful to detect this disease, which also reduces the cost of medical care. In this paper, we used different models, including LR, SVM, KNN, and NB, to analyze diabetes and then show their comparative results. Experiments are conducted on two datasets: the Pima Indian diabetes dataset and the Diabetes Health Indicator, both of which are available on Kaggle.

Keywords: Diabetes; SVM; NB; KNN; LR; BMI; EDA

1. Introduction

Diabetes is a condition where the body fails to metabolize sugar (glucose) in the body [1]. This causes a rise in the level of blood glucose to alarming levels. This is referred to as hyperglycemia. In this state of the body, the body cannot generate insulin. The second possibility is that the body is unable to react to the insulin generated. Diabetes cannot be cured; it has to be managed. The heart attack, kidney failure, stroke, and nerve damage are serious complications that can be developed by a diabetic. The statistics of 2017 indicate that there are estimated 8.8% of the total population in the world with diabetes. It is likely to increase to 9.9% by 2045. The prevalence of diabetes portrays a lot of differences in various countries worldwide. Pakistan has one of the highest rates of diabetes having 30.8 prevalence rate and Kuwait comes in with 24.9 prevalence rate. This is the same rate as Nauru, New Caledonia and the Northern Mariana Islands: 23.4%. Conversely, there are some countries in which diabetes is exceptionally low as indicated in Figure 1 In contrast. The rates are lower in Eritrea, Somalia and Guinea-Bissau with a rate of less than 1 which means that the burden of diabetes is much less in these countries. On the same note, Madagascar, Mozambique and Sierra Leone have lower rates of less than 2 and this further supports the trend of reduced prevalence. It is also possible to observe a trend as the nations with the high incidence of diabetes are more likely to be located in such areas like the Middle East, the South Asian region, or the Pacific Islands. In the meantime, other areas such as Africa have diabetes that is scattered at the low end of the list [2].

As of 2021, the global population of individuals living with diabetes reached 537 million. This number is projected to continue rising steadily until at least the year 2045. By that time, the estimated number of people affected by diabetes is expected to reach a staggering 783 million.

The increase in the number of diabetics will have significant implications for public health and healthcare systems worldwide. As the diabetic population expands, the prevalence of diabetes is also anticipated to rise. By the year 2045, it is projected that around 12 percent of the global population will be affected by diabetes.

This projected increase in diabetes prevalence highlights the urgent need for comprehensive strategies to address this growing health concern. Preventive measures, early detection, improved access to healthcare services, and lifestyle interventions will be vital in mitigating the impact of diabetes and reducing its burden on individuals and healthcare systems worldwide [3]. The distribution of diabetes prevalence varies across different regions and income levels. Urban areas have a higher prevalence rate of 10.8% compared to rural areas, where the prevalence is relatively lower at 7.2%. Similarly, high-income countries exhibit a higher prevalence of 10.4.

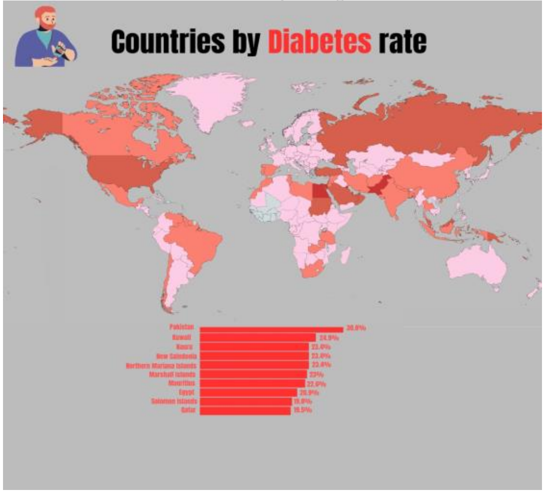


Figure 1. Countries by Diabetes rate

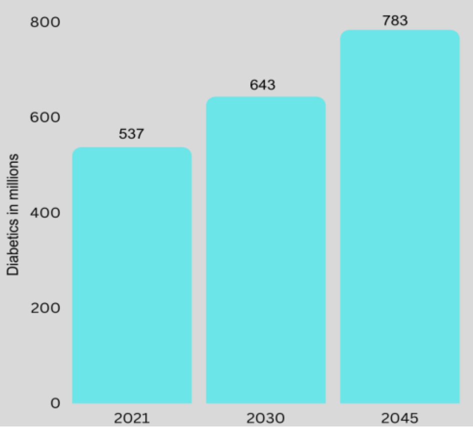


Figure 2. Cases Found in Millions in 2021 Estimated Cases by 2045

A significant portion of the global population remains unaware of their potential diabetes status due to financial constraints or limited access to healthcare services. Many individuals, especially in low-income or undeserved communities, face barriers in seeking regular medical check-ups or diabetes screening tests. Lack of financial resources may deter them from visiting healthcare facilities, obtaining necessary diagnostic tests, or accessing essential medications. Consequently, countless people remain diagnosed and unknowingly live with diabetes, exposing themselves to the risk of developing serious complications associated with the condition [4]. Addressing this issue requires targeted efforts to improve healthcare accessibility and affordability, raising awareness about diabetes, and implement outreach programs to reach vulnerable populations, ensuring timely detection and proper management of the disease.

The American Diabetes Medical Association advises clinicians to conduct fasting plasma glucose tests for patients who exhibit risk factors such as obesity, advanced age, a significant family history of diabetes, or metabolic syndrome-related indicators like elevated triglyceride levels, low high-density cholesterol, impaired plasma glucose tolerance, or high fasting plasma glucose. This screening approach aims to enhance the detection rate of diabetes, especially in individuals with higher susceptibility to the condition. Despite these recommendations, a concerning trend persists where more than half of people diagnosed with diabetes experience delayed treatment initiation because they may not exhibit noticeable symptoms. This delayed treatment can have detrimental effects on the management of the disease, emphasizing the importance of early diagnosis and awareness campaigns to encourage timely medical intervention and support for individuals living with diabetes.

[5] This research aimed to develop a diabetes prediction model using readily available data from Kaggle and utilized advanced deep learning models to assess its effectiveness. By comparing various parameters and analyzing the predictive outcomes, the study successfully identified the occurrence of diabetes. In the future, with the implementation of the appropriate model, healthcare professionals should possess the capability to accurately diagnose diabetes in prospective patients, even in the absence of noticeable symptoms. By analyzing these specific factors, medical practitioners can make informed and precise assessments, leading to early detection and timely intervention in individuals at risk of diabetes. This advancement holds the potential to revolutionize diabetes diagnosis, enabling proactive healthcare measures and improved management of the condition for better patient outcomes [6].

2. Related Work

Diabetes is a long-term type of metabolic disorders that is associated with high levels of blood sugar in the body (hyperglycemia) caused by lack of sufficient insulin secretion or resistance to insulin. Two categories are prevalent: Type 1 that usually occurs in childhood and is caused due to the damage of insulin-producing cells by the immune system and Type 2, common in adults and linked to not only insufficient insulin release but also insulin resistance. The symptoms of the condition include excessive urination, excessive thirst, fatigue, blurred vision, unintended weight loss, and slow wound healing [7]. Attempts to model blood glucose levels in diabetics have resulted in the well-researched methods, primarily through elaborate physiological models. The models subclassify dynamics into meal absorption dynamics, insulin dynamics, and glucose dynamics and end up in analytically solvable equations [8].

Estrada et al. had a research to come up with an ARX (AutoRegressive with eXogenous input) model that is particularly used to forecast future levels of blood glucose in diabetic patients. The model considers two imperative inputs, which are the present blood glucose concentration and the dosage of insulin given. With these included, the ARX model can be successfully used to predict the level of blood glucose within a 45-minute horizon of prediction. The results of the study underline the opportunities of this model to facilitate the optimal management of blood glucose levels in diabetic patients, which may be used in improving the management and treatment approaches to diabetes [9].

Nuryani et al. applied support vectors machine (SVMs) to forecast hypoglycemia with the inclusion of electrocardiograms (ECG) data with blood sugar levels and insulin injections. In the meantime, it was revealed that Marling et al. used SVM models and wearable activity trackers to keep track of heart rate, galvanic skin response, and skin/air temperatures [10].

There was a study by Marling and others, whereby they used SVM models to provide resolution to data collected using wearable activity trackers. These monitors were also fitted to measure several physiological parameters, including heart rate, galvanic skin response, and skin/air temperatures. Through this array of information, the SVM models enabled the researchers to acquire some valuable information about the association between these variables, which could provide new lines of enlightenment on the nature of health and well-being [11].

Zecchin et al. developed a predictor based on the model of a neural network (NN) and a first-order extrapolation algorithm using a first-order polynomial. This is a novel method whereby the past continuous glucose monitoring (CGM) sensor data is combined with carbohydrate intake data. With the combination of inputs, the model can critically predict future glucose levels and this information can be used to manage diabetes and make the best dietary decisions [12].

Hamdi et al. followed the progress of a cohort of 12 patients closely in their study in order to create an elaborate model. It was a new combination of support vector regression and a differential evolution algorithm. This hybrid strategy was supposed to enhance the quality of predictions of the health information of the patient, which, in turn, should allow understanding their diseases and, possibly, help develop more individual and effective health care actions. The research was based on the use of advanced ways of computing into an effort to make significant contributions to the sphere of medical research and patient care that would open the way to better approaches toward disease management [7].

Diabetes type 2 is a complicated disease that has various causes. It is generally linked to a complex of genetic disposition, obesity and lack of physical activities. The etiology of diabetes is however varied and there are other types which are specific. These involve diabetes that is a genetic defect, diabetes caused by the destruction of the pancreatic exocrine activity and diabetes induced by some drugs or chemicals. The complexity of diabetes underlines the necessity of an extensive research and personal methods to control and treat this common metabolic disease successfully [13].

Diabetes is characterized by a rise in blood sugar levels due to insufficient compensatory secretion of insulin. This means that the body’s ability to utilize blood sugar (glucose) effectively is impaired. During digestion, the food we consume is converted into sugar, which then serves as a source of energy for our body. To regulate blood sugar levels, our pancreas secretes insulin, a vital hormone that redirects excess sugar into the body’s cells. When there is an inadequate supply of insulin, the body becomes unable to utilize the surplus sugar, leading to elevated blood sugar levels. This imbalance is a hallmark feature of diabetes, highlighting the critical role of insulin in maintaining proper blood sugar control and overall health [14].

High blood sugar and increased secretion of urine in diabetic pregnancies can result in premature birth as a result of complications such as polyhydramnios. Premature birth may also be increased by infections particularly in the genitourinary tract. Cesarean section is mostly done with overweight babies or when the mother is experiencing complications that might result to premature baby birth or high blood pressure [15].

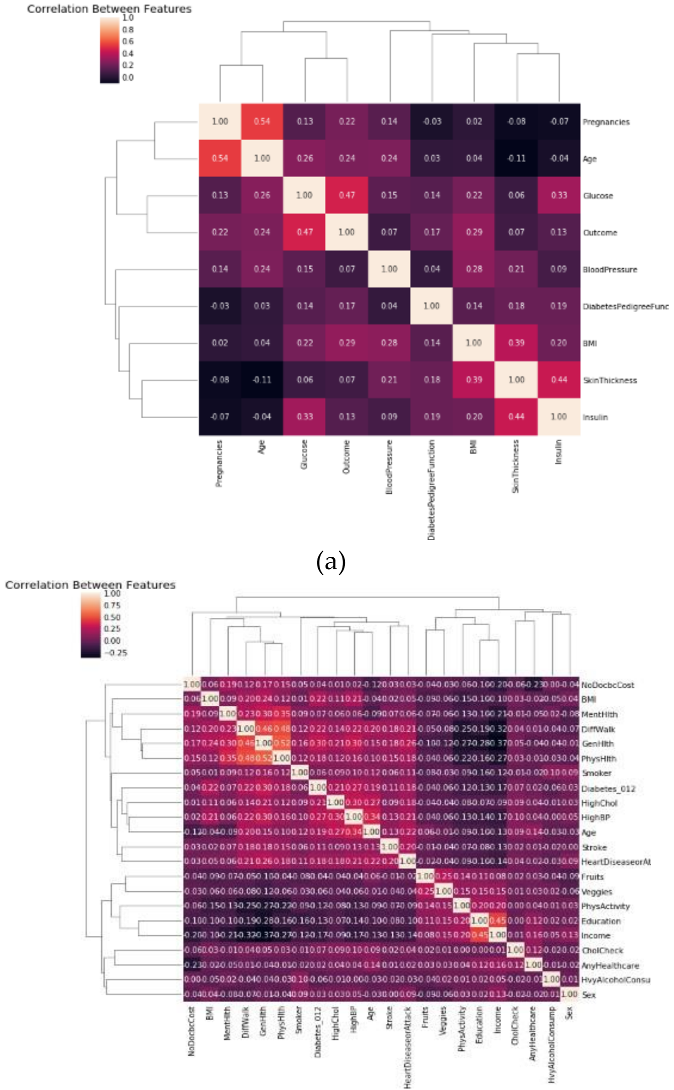
Hassan et al. used eight features and evaluated different machine learning algorithms, namely, decision trees, k-NN (k-Nearest Neighbors), AdaBoost, Random Forest, Naive Bayes, and XGBoost. They have carried out detailed analyses and have discovered that a combination of AdaBoost and XGBoost gave the best and encourage outcomes on their predictive model. The observation highlights the possible advantages of using various machine learning methods together to improve the accuracy and performance of the diabetes prediction models [16].

3. Methodology

In this paper, we explore deep learning methods for diabetes detection using different datasets, including the Pima Indian diabetes dataset and a health indicator dataset. Our main goal is to achieve improved accuracy in diabetes detection with these advanced machine learning techniques. To ensure robust data analysis and modeling, we employ Exploratory Data Analysis (EDA) techniques to gain a deeper understanding of the datasets and preprocess the data effectively.

Evaluating the model on multiple datasets allows us to assess its generalizability and performance across diverse populations. The outcomes of this research could provide valuable insights for early diagnosis and better diabetes management.

Through rigorous experimentation and optimization, we fine-tune the deep learning models to maximize their performance. We carefully validate the accuracy and performance metrics to ensure the reliability of our findings. This study represents a significant step towards developing more accurate and reliable tools for diabetes detection, potentially impacting healthcare outcomes and reducing the global burden of diabetes.



(a)

(b)

Figure 3. Correlation of D1 and D2

3.1. Dataset

In our study, we employed two publicly available datasets, namely the Pima-Indian diabetes dataset and the Diabetes- Health indicator dataset, to predict the occurrence of diabetes. The Pima-Indian diabetes dataset comprises 768 rows and 9 attributes, while the Diabetes-Health indicator dataset consists of 253,681 rows and 22 features are shown in fig 3 and 4. By using these diverse datasets, we aimed to explore the effectiveness of diabetes detection and compare the accuracy of the predictive models derived from each dataset. Our analysis involves assessing the performance of the models to determine which dataset yields more accurate results in predicting diabetes 4.

3.2. Algorithmic Description

In this section, we explored various algorithms to detect diabetes using two distinct datasets. For dataset 1, we employed Logistic Regression (LR), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), and Decision Tree algorithms. Comparing the performance of these models on the Pima-Indian Diabetes dataset revealed that certain models exhibited higher accuracy in predicting diabetes.

Next, we applied Logistic Regression (LR) and k-Nearest Neighbors (KNN) algorithms to dataset 2 and compared the results with those from the Pima-Indian Diabetes dataset. Surprisingly, we found that dataset 2 outperformed dataset 1 in terms of accuracy, suggesting that the combination of LR and KNN was more effective in predicting diabetes on this particular dataset.

3.3. Model Architecture

3.3.1. Visualization and Standardization

The data was imported for feature visualization, and visualization tools were used to observe the data distribution for each field when a patient was diagnosed with diabetes. Additionally, the correlation between diabetes and all variables was examined to understand the relationships between different factors.

3.3.2. Implemented Models

Logistic Regression: Logistic Regression is widely used machine learning algorithm in diabetes research and medical applications. It is a supervised learning method that is particularly suitable for binary classification problems, making it well- suited for predicting the presence or absence of diabetes based on patient data. The evaluation of this classification is measured in terms of several key metrics, including average accuracy, test accuracy, and the confusion matrix. These performance indicators provide valuable insights into the model's effectiveness in correctly classifying diabetic and non-diabetic patients. The average accuracy represents the overall correctness of the model's predictions across all classes, providing an overview of its general performance. Test accuracy specifically assesses the model's accuracy on new, unseen data, offering a more realistic estimation of its real-world performance. 2) We achieved an average accuracy of 75% on Pima-Indian Diabetes. On other dataset we achieve the average accuracy of 84%.

K-Nearest Neighbors: KNN is particularly useful when the data is non-linear and there may not be a clear boundary between diabetic and non-diabetic cases. KNN works by comparing a new patient's data with the data of existing patients in the dataset. The algorithm identifies the k-nearest patients (nearest neighbors) to the new patient based on the similarity of their features, such as blood glucose levels, BMI, age, and other relevant attributes. The class (diabetic or non-diabetic) of these k- nearest neighbors is then used to classify the new patient. We achieve the accuracy of 77% on Pima-Indian Diabetes dataset and 82 SVM: SVM is a supervised learning technique that can be applied to classify and predict diabetes based on data from patients.

SVM has been used for Diabetes Classification. SVM can classify patients into diabetic or non-diabetic groups based on features like blood glucose levels, insulin dosage, body mass index (BMI), and other relevant factors. SVM finds the optimal hyperplane that best separates the two classes, allowing it to accurately classify new patients. We achieve the accuracy of 84%.

NB: The algorithm calculates the probability of a patient belonging to each class (diabetic or non-diabetic) given their feature values. It assumes that the features are conditionally independent. We achieve the accuracy of 84%.

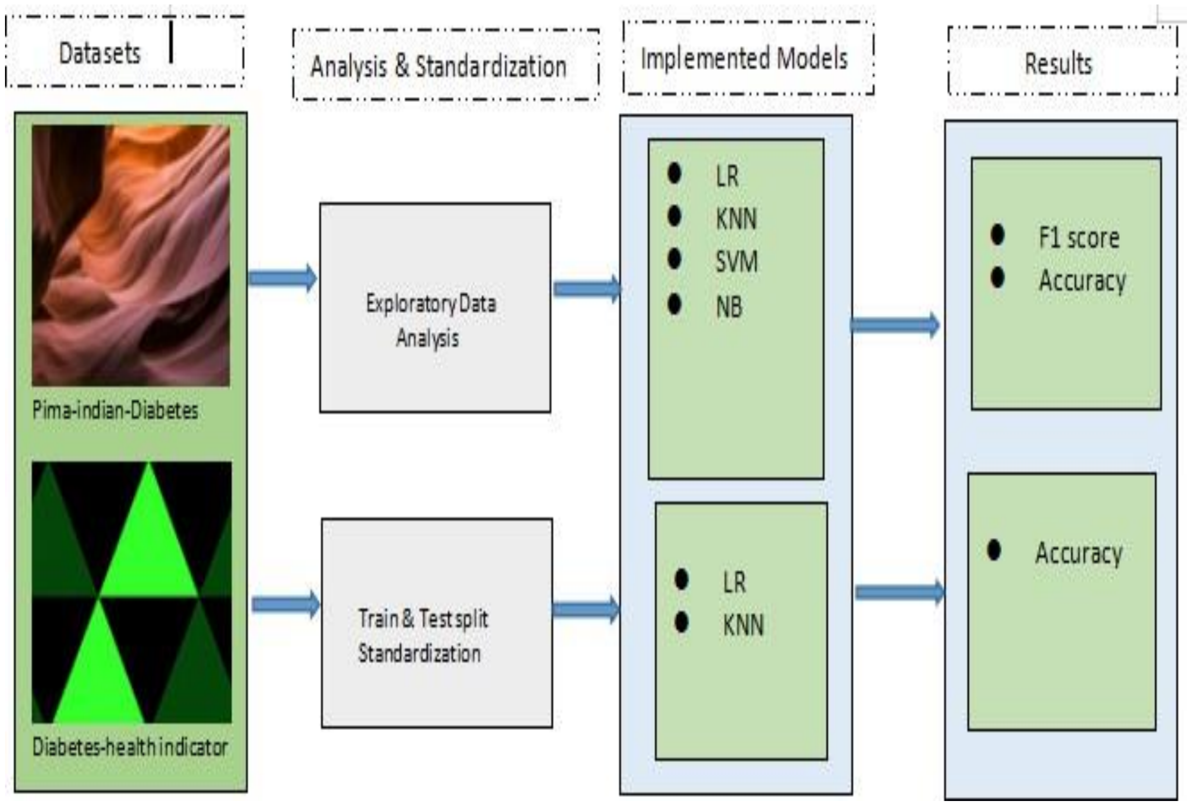

Datasets
Analysis & Standardization
Implemented Models
Results
Pima-indian-Diabetes
Diabetes-health indicator
Exploratory Data Analysis
Train & Test split Standardization
● LR
● KNN
● SVM
● NB
● LR
● KNN
● F1 score
● Accuracy
● Accuracy

Figure 4. Detailed Architecture of the models

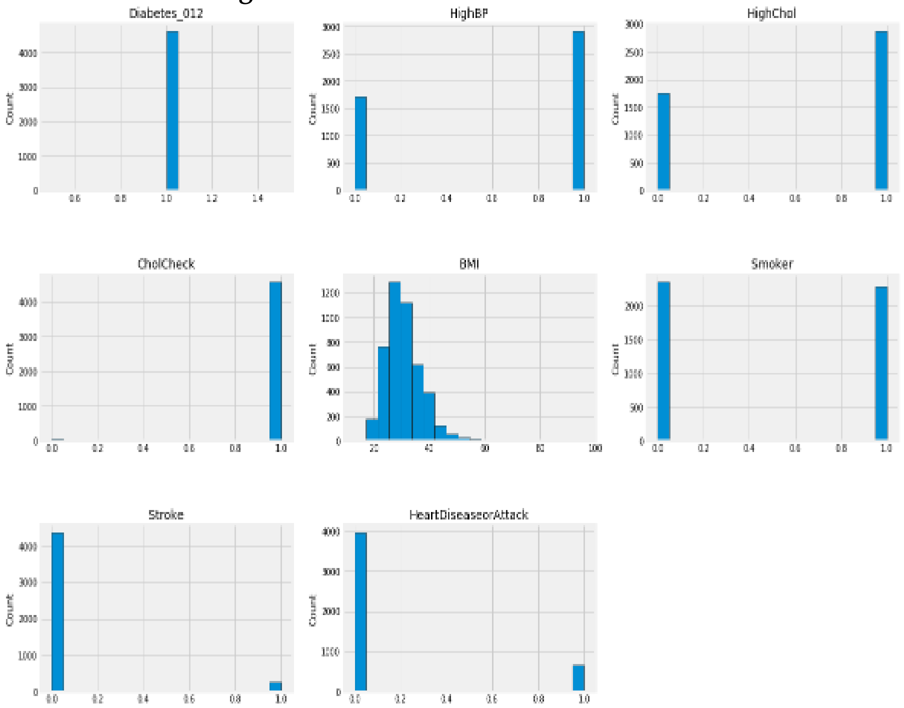


Figure 5. Analysis of Diabetic Cases of Health Indicator

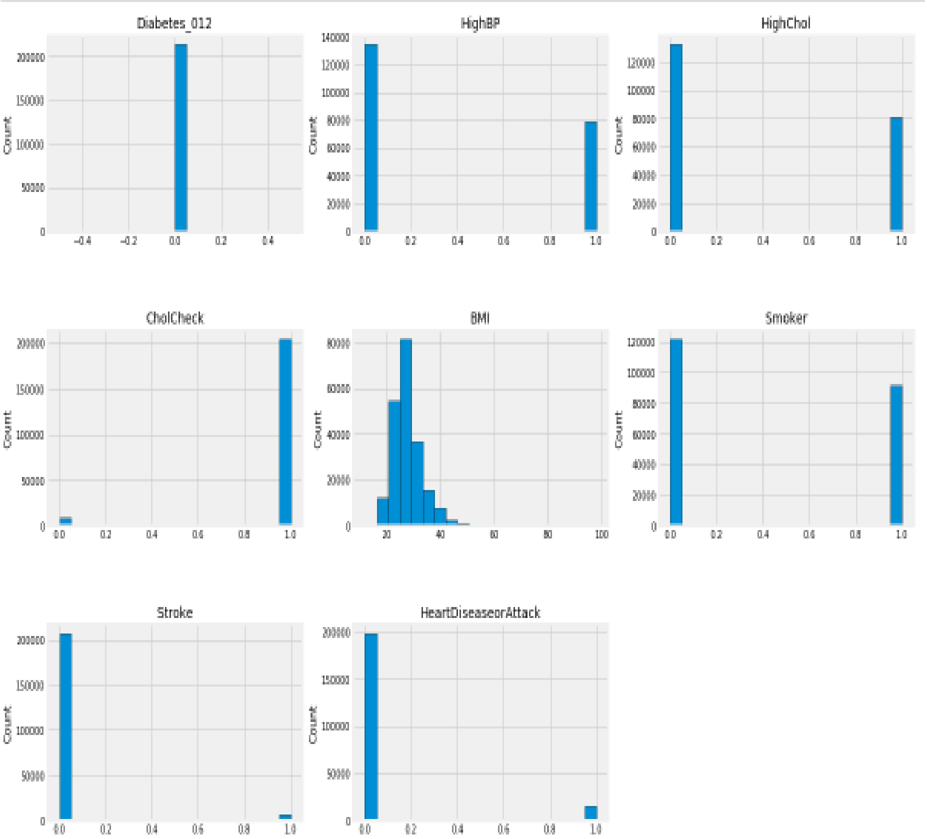


Figure 6. Analysis of Non-Diabetic Cases of Health Indicator

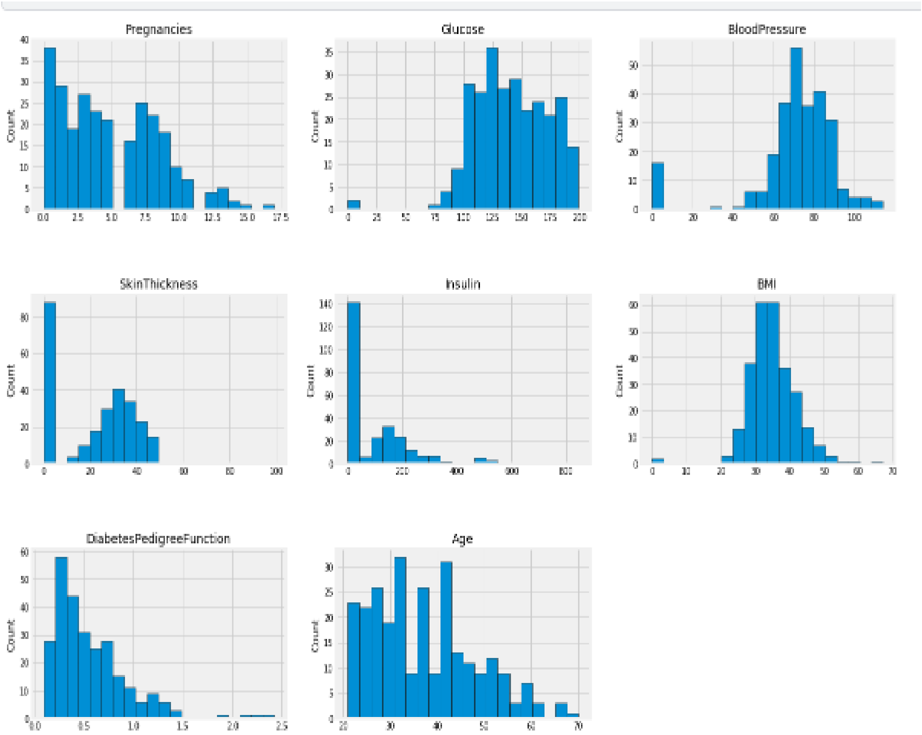


Figure 7. Analysis of Diabetic cases PIMA- Indian

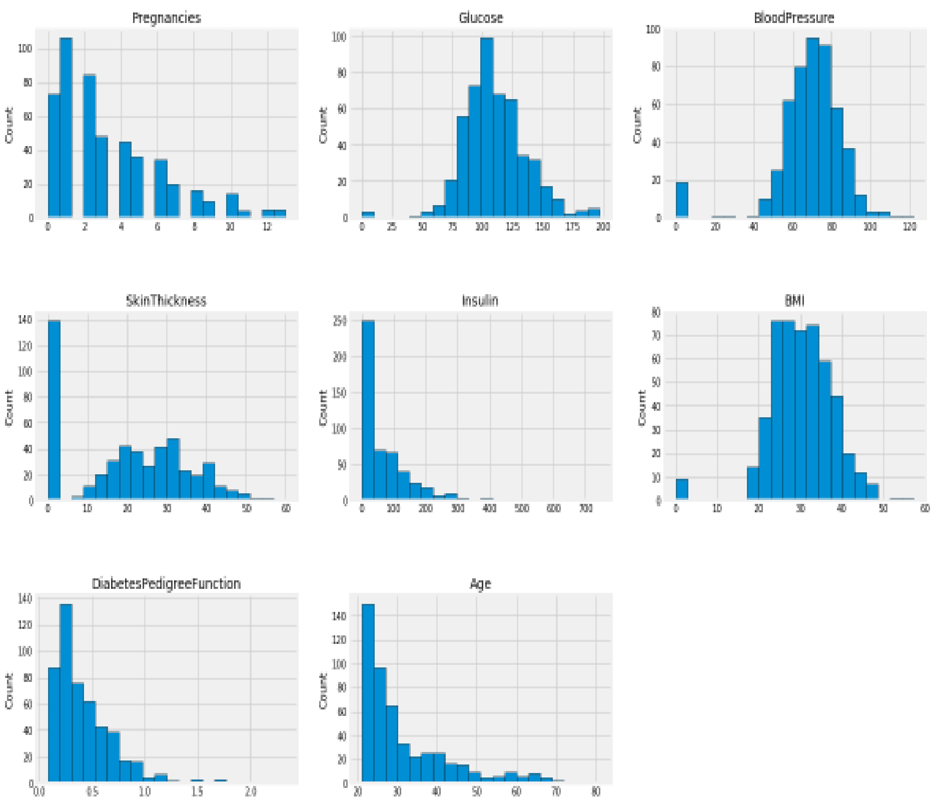


Figure 8. Analysis of non-Diabetic cases of Health Indicator

4. Results

We conducted a comparison of our results with those obtained from Deep Learning models, specifically on the Pima- Indian and Health-Indicator Datasets. The outcomes from our analysis showed variations between the two datasets, with the Health-Indicator dataset yielding more accurate results. These findings highlight the significance of dataset selection in the performance of predictive models for diabetes detection. The superior accuracy achieved on the Health- Indicator dataset underscores the importance of diverse and relevant data sources in enhancing the effectiveness of diabetes prediction models.

Pima-Indian and Health Indicator on different models.

Table 1. PIMA-Indian and Health-Indicator on Same models

Model	Result-1	Result-2
LR	75.00	84.00
KNN	77	82

Table 2. PIMA-Indian on different models

Model	Result-1
SVM	84.00
NB	84.00

Table 3. PIMA-Indian and Health-Indicator on Same models

Model	Result-1	Research-1	Research-2	Research-3
SVM	84.00	65.00		
NB	84	76	78	79

KNN	77	69

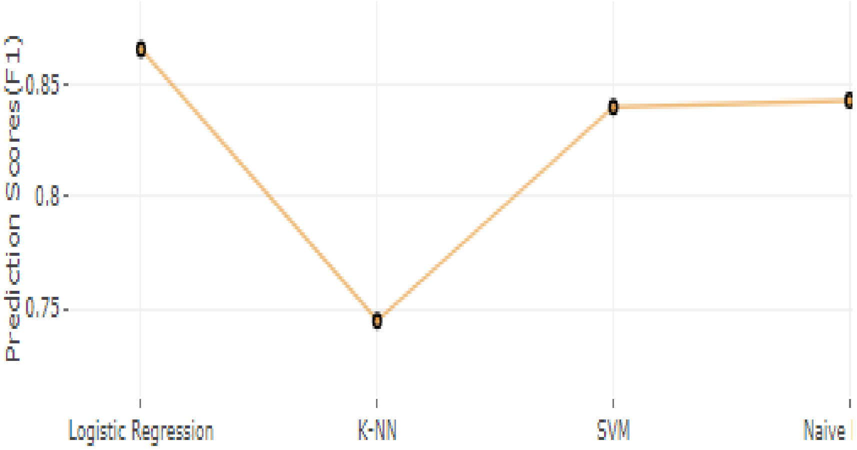


Figure 9. F1-Score of PIMA-Indian dataset

5. Conclusion Future Work

In this research paper, we worked with two datasets and utilized various models for diabetes prediction. From dataset 1, we applied Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). Among these models, SVM and NB showed promising accuracy results. However, when comparing the two datasets, we found that dataset-2 yielded even better accuracy for diabetes prediction. These findings emphasize the importance of dataset selection and model evaluation in achieving more accurate and reliable results in diabetes detection.

As we Move forward, we plan to expand our research by exploring various techniques and incorporating the combination of different models to further improve accuracy. For Dataset 1, we intend to apply additional models to assess their performance and identify potential enhancements. Meanwhile, for Dataset 2, we will focus on leveraging the power of combining multiple models to achieve even better results.

References

1. Centers for Disease Control and Prevention. (Accessed August 2023) Type 2 diabetes. [Online]. Available: https://www.cdc.gov/diabetes/basics/type2.html
2. W. Voter. (Accessed August 2023) Diabetes rates by country. [Online]. Available: https://wisevoter.com/country-rankings/diabetes- rates-by-country/
3. Statista. (Accessed August 2023) Number of diabetics worldwide. [Online]. Available: https://www.statista.com/statistics/271442/number- of-diabetics-worldwide/
4. U. Rajendra Acharya, K. S. Vidya, D. N. Ghista, W. J. E. Lim, F. Molinari, and M. Sankaranarayanan, "Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method," Knowledge-Based Systems, vol. 81, pp. 56–64, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705115000374
5. "Introduction," Diabetes Care, vol. 25, no. suppl1, pp.s1 − −s2, 012002.[Online].Available : http://doi.org/10.2337/diacare.25.2007.S1
6. R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," International Journal of Environmental Research and Public Health, vol. 18, no. 14, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/14/7346.
7. I. Rodríguez-Rodríguez, I. Chatzigiannakis, J.-V. Rodríguez, M. Maranghi, M. Gentili, and M.- Zamora-Izquierdo, "Utility of big data in predicting short-term blood glucose levels in type 1 diabetes mellitus through machine learning techniques," 2019. [Online]. Available: https://www.mdpi.com/1424- 8220/19/20/4482
8. J. L. Ruiz, J. L. Sherr, E. Cengiz, L. Carria, A. Roy, G. Voskanyan, W. V. Tamborlane, and S. A. Weinzimer, "Effect of insulin feedback on closed-loop glucose control: A crossover study," Journal of Diabetes Science and Technology, vol. 6, no. 5, pp. 1123–1130, 2012, pMID: 23063039. [Online]. Available: https://doi.org/10.1177/193229681200600517
9. Kianat, Humayun Salahuddin, & Muhammad Saleem Anjum. (2024). IoT based Intelligent Pollution Monitoring System using Machine Learning Technique. *Journal of Computing & Biomedical Informatics*, *6*(02), 529–537. Retrieved from https://www.jcbi.org/index.php/Main/article/view/469
10. G. Castillo Estrada, L. del Re, and E. Renard, "Nonlinear gain in online prediction of blood glucose profile in type 1 diabetic patients," in 49th IEEE Conference on Decision and Control (CDC), 2010, pp. 1668–1673.
11. N. Nuryani, S. H. Ling, and H. Nguyen, "Electrocardiographic signals and swarm-based support vector machine for hypoglycemia detection," Annals of Biomedical Engineering, vol. 40, pp. 934–945, 10 2011.
12. C. Marling, L. Xia, R. Bunescu, and F. Schwartz, "Machine learning experiments with noninvasive sensors for hypoglycemia detection," in Proceedings of the IJCAI Workshop on Knowledge Discovery in Healthcare Data, New York, NY, USA, vol. 10, 2016.
13. M. Hussain, W. Sharif, M. R. Faheem, Y. Alsarhan, and H. A. Elsalamony, "Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model", Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29779–29786, Dec. 2025.
14. C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short- time prediction of glucose concentration," IEEE transactions on biomedical engineering, vol. 59, no. 6, pp. 1550–1560, 2012.
15. M. A. Makroum, M. Adda, A. Bouzouane, and H. Ibrahim, "Machine learning and smart devices for diabetes management: Systematic review," Sensors, vol. 22, no. 5, p. 1843, 2022.
16. Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A machine learning approach to predicting diabetes complications," in Healthcare, vol. 9, no. 12. MDPI, 2021, p. 1712.
17. R. Jagannathan, J. S. Neves, B. Dorcely, S. T. Chung, K. Tamura, M. Rhee, and M. Bergman, "The oral glucose tolerance test: 100 years later," Diabetes, Metabolic Syndrome and Obesity, pp. 3787–3805, 2020.
18. S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: review and case study," Applied Sciences, vol. 9, no. 21, p. 4604, 2019.