STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–12.
Published online in International Academic Press (www.IAPress.org)

Predicting Type 2 Diabetes based on Machine and deep learning models

Hosam Eldin Fawzan Sayed [1], Ahmed Abdelhafeez [2,3], Mohamed N.M.M.Hassan [4]

[1]*Department of Computer Science, Faculty of Computers and Information, Arish University — North Sinai, Egypt*
[2]*Faculty of Computer and Information Technology, Innovation University, Cairo, Egypt*
[3]*Applied Science Research Centre. Applied Science Private University, Amman, Jordan*
[4]*Faculty of Artificial Intelligence, Egyptian Russian university*

Abstract The global rise of spread the Type 2 Diabetes (T2D) disease has become a major global health challenge. Early detection of this disease is essential for limiting its progression and reducing its potential effect. This study estimates the performance of machine learning (ML) and deep learning (DL) models for prediction T2D using an enhanced version of Pima Indian diabetes dataset. The dataset has been improved through comprehensive preprocessing, feature engineering, and class imbalance handling via Synthetic Minority Oversampling Technique (SMOTE). A total of Seven machine learning classifiers- Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, k-Nearest Neighbors, Naïve Bayes, and XGBoost were assessed alongside three deep learning models Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The Experimental results demonstrate that, XGBoost achieved best predictive performance, with an accuracy (96.84%, AUC = 0.99), followed by Random Forest (96.32%, AUC = 0.98). Decision Tree and SVM also showed robust performance, while Naïve Bayes was the least accurate 80.5%. In contrast, The DL models achieved an accuracy between (73–77%).

Keywords Type 2 Diabetes, Machine Learning, Deep Learning, XGBoost, SMOTE, Ensemble Learning.

DOI: 10.19139/soic-2310-5070-3582

1. Introduction

Diabetes Mellitus is a problem that one has to live with throughout his or her life. It is a condition in which the body has problems metabolizing sugar. This is because the body does not produce enough insulin or because the insulin produced does not function correctly. If not well managed, Diabetes Mellitus can cause serious complications that can even threaten one's life. [1]. People with diabetes mellitus often must go to the bathroom a lot they get very thirsty. They feel hungry all the time. If diabetes mellitus is not treated for a time, it can cause serious problems like diabetic ketoacidosis or hyperosmolar hyperglycemic state. Diabetes mellitus can also cause long-term problems, like heart disease, stroke, kidney disease, nerve damage, and eye problems [2]. The pathogenesis of diabetes is primarily based on insulin dysregulation in the three main forms [3]:

The first one, Type 1 diabetes (T1D), is an autoimmune disease where the pancreas cannot produce enough insulin, requiring lifelong insulin therapy. The second is Type 2 diabetes (T2D), which involves insulin resistance and gradual $\beta - cell$ failure, often associated with obesity, lack of exercise, and family history. The third Gestational diabetes (GDM) is high blood sugar that occurs during pregnancy, increasing health risks for both mother and baby.

The American Diabetes Association (ADA) 2023 Standards of Care in Diabetes clearly confirms that glucose levels, body mass index (BMI), and blood pressure are some of the most significant predictors of risk for Type 2

*Correspondence to: Hosam Eldin Fawzan Sayed (Email: Hosam.Fawzan@ci.aru.edu.eg). Computer science department, Faculty of Computers and Information, Arish University - North Sinai, Egypt (45511).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Diabetes [4]. When we looking at fasting plasma glucose levels, we can see that there are ranges that doctors use to figure out how well our bodies are handling glucose. These ranges are like benchmarks that help doctors classify people into three groups: people with glucose levels, which is between 70 and 99 milligrams per deciliter, people with Prediabetic glucose levels, which is between 100 and 125 milligrams per deciliter [5], and people with Diabetic glucose levels, which is 126 milligrams per deciliter or higher. The Prediabetic stage is important because it is like a warning sign that someone is at a higher risk of getting Type 2 diabetes, also known as T2D, later. Major risk factors include obesity (BMI of ≥ 25), a family history of type T2D diabetes, dyslipidemia (HDL $< 40mg/dL$), hypertension, PCOS, age ≥ 45 years, and being inactive. Moreover, ethnic and socioeconomic factors also increase the risk and impact the African-American, Native-American, Latin-American, and Asian-Pacific populations of people [6].

The proposed model strategy emphasizes early detection, lifestyle modification, diet, and exercise in the hope of reversing the disease process [5]. However, Traditional HbA1c and fasting glucose tests diagnose diabetes once the irreversible damage has been incurred in physiology, and it measures in dire need. Building upon the foundational work of [7]. The objective of this study is to improve predictive performance by incorporating more advanced machine learning algorithms and utilizing a larger, more diverse dataset. Also, a rigorous benchmarking of state-of-the-art models using deep learning algorithms on the Diabetes datasets.

2. Related Work

recent years, the problem of "Diabetes prediction" has gained significant attention from machine learning (ML), data mining, and neural network techniques, particularly due to the rising health issue in areas such as India [8]. Most works utilize supervised ML techniques, which include decision trees, logistic regression, SVM, and ensemble techniques, and in most cases, the Pima Indians Diabetes Dataset (PIDD) has been used to validate the approaches [9–21]. Similarly, classic ML-based techniques employing decision trees, naive Bayes, SVM, and optimized KNN techniques were found to be quite effective and were reported to have accuracy in the range of 76% to 95% in the works carried out by [9–12]. Additionally, a hybrid technique employing GA and RBF neural network techniques was found to be highly effective in this regard [13]. Focusing on feature-based research emerged with predictors including glucose, BMI, pregnancies, age, and insulin needs as major predictors [14, 15]. Moving beyond PIDD, other research has introduced larger EHR systems and big data models for diabetes analysis [16] and ensemble, bagging, and boosting analytics were found to provide 84-86% accuracy results with ensemble, bagging, and boosting machine approaches [17–19]. Other research using PCA, mRMR, J48graft, and various ANN results highlighted the significance of dimension reduction and strong rule-based classifiers with PCA and mRMR [20]. Longitudinal EMRs provided predictive capability for logistic regression analysis for high-risk patient populations [21].

Recent advances focus on data integration through multimodal data and advanced AI systems. Ensemble stacking models have offered better prediction accuracy over other machine learning models as a single classifier [22], reinforcement learning for personalized T2D intervention optimization has been explored [23], and the application of deep learning models to retinal images has shown promising accuracy levels for diabetes diagnosis/detection and prediction [24]. Federated learning is proposed as an efficient framework for privacy-preserving predictive models for diabetes diagnosis/detection and prediction [25], and while explainable AI techniques such as XGBoost combined with SHAP are useful for risk score development and interpretation [26]. Machine learning techniques combined with the transformer framework and including variables such as CGM, lifestyle, EHR data, and socio-clinical data have been reported to show state-of-the-art accuracy levels, up to 96%, and area under the curve up to 0.95 for T2D diagnosis/detection and prediction across varied data populations [27, 28]. A DL-based technique, CNN–LSTM, reported an accuracy of 95.7% [29].

A feature selection framework using the snake optimization approach proposed by [30] showed promising results in enhancing the efficiency of cardiovascular disease diagnosis by improving accuracy and reducing redundant medical features. [31] proposed an optimized liver disease classification approach with the classification of liver diseases using machine learning algorithms optimized via a Binary Particle Swarm Optimization method,

improving liver disease diagnosis performance and machine learning features efficiency. [32] proposed a novel approach for optimizing deep learning networks with the classification of early chronic kidney diseases using machine learning, optimized via a Waterwheel Plant Algorithm, improving prediction accuracy and stability. [33] proposed a novel approach for improving a groundwater resource prediction system with machine learning, optimized via a Comment Feedback Optimization Algorithm, improving prediction accuracy and resource management. [34] proposed a new approach for improving the diagnosis of skin melanoma diseases with improved machine learning, optimized via improved Meta-GVF optimization algorithms. In one of the initial works, deep learning techniques were found to be quite effective.

3. Dataset

3.1. Dataset Description

The present study utilizes two Pima Indians Diabetes Datasets, one for machine learning called "Diabetes Dataset 2019", as suggested by the research methodology of [7]. The first comprises 952 samples and 18 features representing the core dataset [35]. There are more representations of 18 features in [36]. To improve model training and reduce the risk of overfitting, the final dataset used in the experiment is expanded using SMOTE (Synthetic Minority Oversampling Technique) to 5,437 records during training process. The primary source is based on information about health, lifestyle, and family history. The second Pima Indians Diabetes Dataset comprises 768 samples and 8 features represented for deep learning.

Table 1. Dataset details

Age	Gender	Family diabetes	High BP	Physically Active	BMI	Smoking	Alcohol	Sleep
50-59	Male	no	yes	one hr. or more	39.0	no	no	8
50-59	Male	no	yes	less than half an hr.	28.0	no	no	8
40-49	Male	no	no	one hr. or more	24.0	no	no	6
50-59	Male	no	no	one hr. or more	23.0	no	no	8
40-49	Male	no	no	less than half an hr.	27.0	no	no	8

Sound Sleep	Regular Medicine	Junk Food	UriationFreq	Stress	BPLevel	Pregnancies	Pdiabetes	Diabetic
6	no	occasionally	not much	sometimes	high	0.0	0	no
6	yes	very often	not much	sometimes	normal	0.0	0	no
6	no	occasionally	not much	sometimes	normal	0.0	0	no
6	no	occasionally	not much	sometimes	normal	0.0	0	no
6	no	occasionally	not much	sometimes	normal	0.0	0	no

The number of samples in the original dataset was 952, comprising 686 non-diabetes samples (72%) and 266 is diabetic samples (28%) representing class 0 and class 1, respectively, as shown in figure 1 and 2.

4. The proposed Model

4.1. Dataset Preprocessing

In the dataset preprocessing, the process of data set preparation uses various techniques like handling missing values using median and mean imputation methods or deleting certain records or replacing values with mode values. The categorical values (Normalization) were handled using label encoding. To deal with imbalanced classes, Minority Oversampling Technique (SMOTE) handling is used. In particular, zero or invalid values in clinical attributes were treated as missing values, as is commonly done in data sets like Pima Indians Diabetes Data Set. An

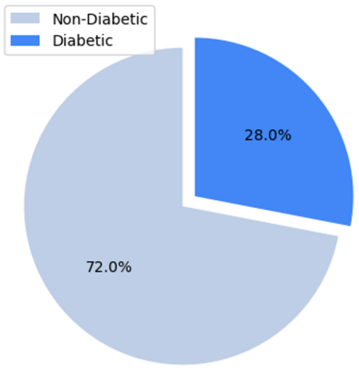

Figure 1. Diabetes Class Distribution for original dataset

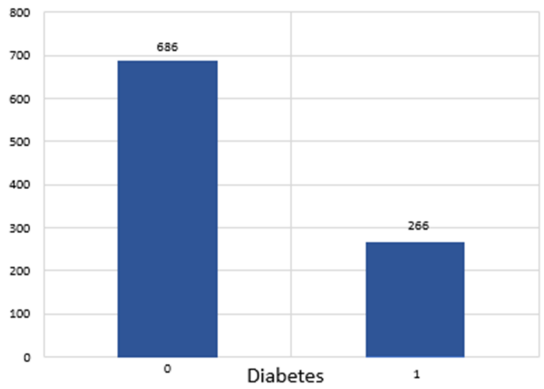

Figure 2. Diabetes Class Distribution (Bar Chart)

Table 2. Feature Description

No.	Feature	Description	Range/Format
1	Age	Age group of the person	e.g., 40-49, 50-59
2	Gender	Gender of the person	Male/Female
3	Family diabetes	Family history of diabetes	Yes/no
4	highBP	Has high blood pressure	Yes/no
5	PhysicallyActive	Daily physical activity level	e.g., one hr. or more, less than ½ hr.
6	BMI	Body Mass Index	e.g., 23.0, 39.0
7	Smoking	Whether the person smokes	Yes/no
8	Alcohol	Whether the person drinks alcohol	Yes/no
9	Sleep	Total sleep hours per day	e.g., 6, 8
10	Sound Sleep	Hours of restful/deep sleep	e.g., 6, 8
11	Regular Medicine	Takes medication regularly	Yes/No
12	Junk Food	Frequency of junk food consumption	occasionally, very often
13	Stress	Level of psychological stress	e.g., sometimes
14	BPLevel	Blood pressure level	e.g., high, normal
15	Pregnancies	Number of pregnancies	e.g., 0.0
16	Pdiabetes	Pre-diabetes status	e.g., 0
17	UriationFreq	Urination frequency	e.g., not much
18	Diabetic	Diabetic or not (target)	Yes/No

appropriate imputation method has been used; in particular, median, mean, and mode imputation were used based on feature types and distributions. These details have been included in order to enhance clarity and transparency.

4.2. Data Visualization

Figure 4 shows Data visualization supported by a correlation heat map: With the help of this visual analytics tool, it is possible to determine the measure of correlation between the medical features of a data set. In diabetes prediction, this correlation plot presents important information on how physiological, biochemical, and demographic variables influence the variation of each other. The correlation heatmap emphasized that strong predictor variables such as glucose, BMI, and age are very indicative of the risk of diabetes, supporting the results of dataset. This plot led to further feature selection and ensured biological relevance of the selected parameters for the development of predictive models.

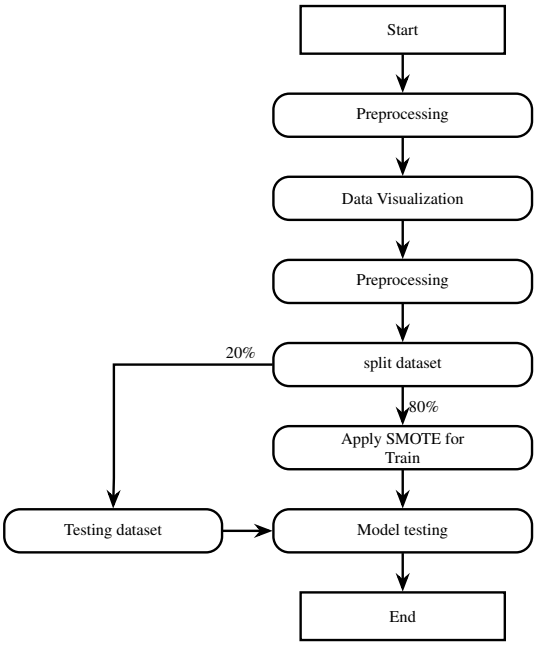

Figure 3. The proposed Model

4.3. Feature Encoding

The features like Gender (Male or Female) and Diabetic (Yes or No) columns have been encoded as per the Label Encoding technique, where categorical values have been converted to numerical values 0 and 1.

4.4. SMOTE Technique

SMOTE represents a new technique for increasing the number of samples that represent the minority class in the dataset by generating synthetic data, which does not just duplicate existing samples. To balance the dataset between the diabetic (minority) and non-diabetic (majority) classes, the Synthetic is used exclusively on the training dataset. This approach helped create artificial class members for the minority class, thereby balancing the dataset appropriately. For instance, the approach is useful in enabling the generalization of the classifier without biasing towards the majority class.

4.5. Model Testing

The strategy employed to assess the performance of this model involved dividing the dataset into 80% training set and 20% test set for the model. resulting in approximately 4,350 training samples and 1,087 testing samples. A stratified approach is used for splitting the dataset. To prevent overfitting, the parameters were learned during training, while the test set was left untouched to provide an unbiased evaluation of model performance on real-world, imbalanced data. Class imbalance is handled during training set using the SMOTE approach. The model was evaluated on the test set using Precision-recall curves were used to evaluate the performance of this model.

5. Machine learning Models

5.1. Logistic Regression

Logistic regression is a supervised learning technique designed to model the relationship between a binary outcome variable and one or more predictor variables by estimating class probabilities through the sigmoid function.

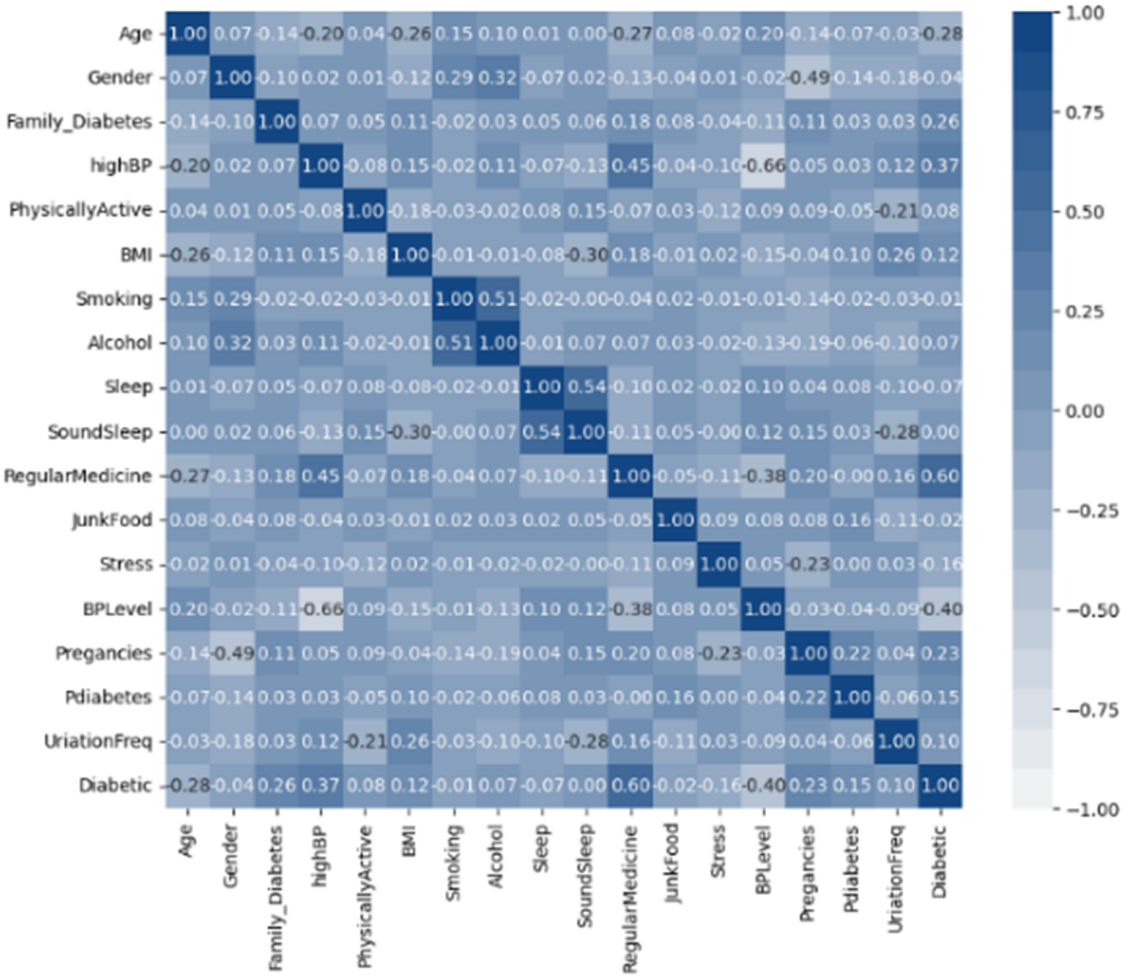

Figure 4. Correlation Heatmap of Medical Features

Despite its terminology, logistic regression is primarily used for classification rather than regression tasks. It is particularly suitable for problems in which the dependent variable is dichotomous, such as 0/1, /1, or true/false, while the independent variables may be measured on binary, ordinal, interval, or ratio scales. The mathematical representation of the sigmoid function is expressed as follows [21]:

$$Y = \frac{1}{1 + e^{-x}} \tag{1}$$

Where, y is the result of weighted sum of input x. If the output is more than half, then output is one, otherwise output will be zero.

5.2. Random Forest

The Random Forest classifier creates multiple decision trees from randomly selected subsets of training dataset. Then it aggregates the votes from different decision trees to decide the final class of test objects [37, 38].

5.3. Decision Tree

Decision Trees are a type of learning model that people use to classify things and make predictions. They are good because they are easy to understand and have a structure. The Decision Tree model looks at the data and then Make decisions based on features of the data. This is achieved by creating a tree with many decision points and ending points that show what the model thinks is going to happen. The Decision Tree provides good accuracy and is very stable [39, 40]. The Decision Tree is a way to make decisions, and it is extremely helpful.

5.4. Support Vector Machine (SVM)

SVM is one of the most widely used ML algorithms for classification and regression analysis, where the main focus is on classification problems. The fundamental objective of SVM is to obtain an optimal separating hyperplane in a multidimensional feature space that discriminates between the data samples belonging to different classes. Such a hyperplane defines a decision boundary that maximizes the margin between the nearest data points, called support vectors, across the classes. Maximizing this margin boosts the robustness of classification and improves the performance of generalization. Moreover, SVM can handle nonlinearly separable data by using kernel functions [41].

5.5. K-Nearest Neighbors (KNN)

K-Nearest Neighbor, or KNN, is a multi-used algorithm used in regression and classification tasks, although business-oriented applications commonly use it for classification. One of its key strengths lies in its conceptual simplicity and low computational overhead [42].

5.6. Naive Bayes

Naïve Bayes is a probabilistic ML method based on Bayes theorem which has been defined for probability. In spite of its simplicity, the naïve bayes method performs better than other classification algorithms; therefore, it is one of the best among them. Bayes theorem for calculating the posterior probability has been defined below [43]:

$$P(c|x) = P(x)P(x|c)P(c) \tag{2}$$

Where $P(c|x)$ is the **posterior probability** of class c given predictor x, $P(x|c)$ is the **likelihood** of predictor x given class c, $P(c)$ is the **prior probability** of class c and $P(x)$ is the **prior probability** of predictor x.

5.7. XGBoost

XGBoost is a machine learning algorithm that uses boosting techniques to improve prediction accuracy. It is used for classification and regression problems. XGBoost works by combining several small decision trees, where each new tree tries to correct the errors made by the previous ones. One of its main advantages is speed and high performance, as well as the ability to handle missing data and reduce overfitting through regularization. The reasoning strong predictive power of XGBoost is widely used in many data science competitions and real-world applications [44].

6. Results

6.1. Machine learning results

For estimate the model, the holdout validation technique which splitting dataset to 80% training and 20% for testing as machine learning results to prevent data leakage. The comparative performance of multiple ML algorithms for diabetes prediction is presented in Table 3. The results indicate that the proposed model achieved improvements in accuracy relative to previous studies. Correlation between features showed statistically correlations, i.e., moderate positive correlation of age with systolic blood pressure (*r* = 0.35), confirming the dataset's validity for predictive modeling.

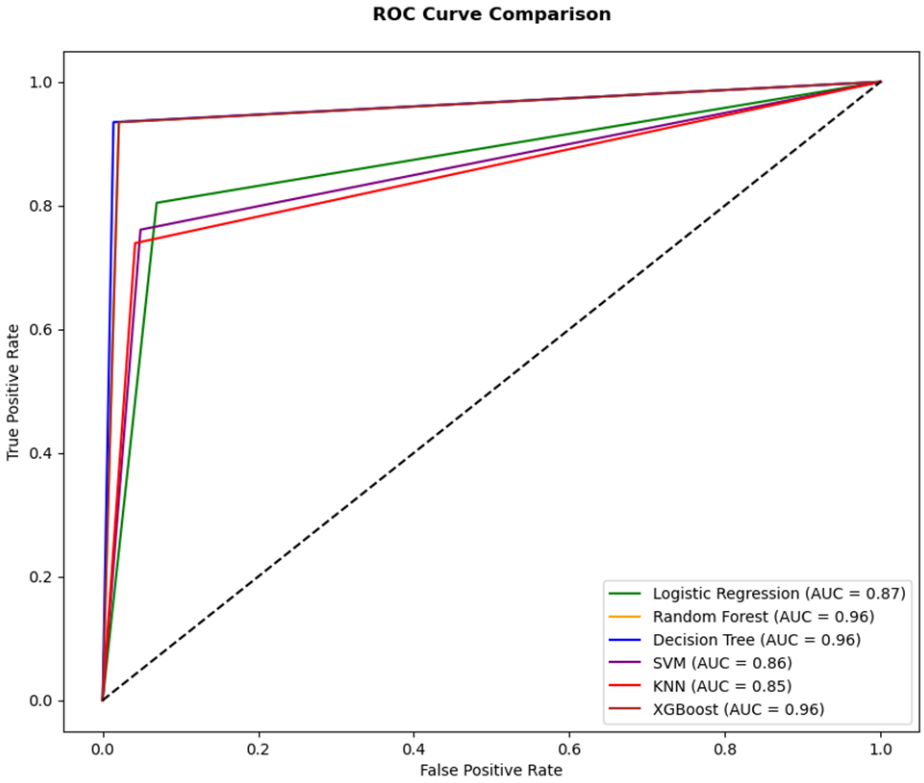

Figure 5. ROC curves comparing models' performance

Table 3. Comparative Performance of Machine Learning Models for Type 2 Diabetes Prediction

Results using original dataset 952 records							
	Logistic Regression	**Random Forest**	**Decision Tree**	**XGboost**	**Naïve Bayes**	**SVM**	**KNN**
Tigga & Garg [7]	85.7%	94.1%	84%	–	80.6%	86.5%	77.3%
Sisodia [10] & Dilip	–	–	73.82%	–	76.3%	65.1%	–
Ioannis Kavakiotis [45]	85%	–	–	–	82%	88%	–
Inam Abousaber [46]	91%	100%	95%	–	–	–	–
Islam, M. M [47]	–	96.8%	–	95.1%	–	91%	88%
Results using expansion dataset 5,437 records							
Proposed Model	90%	96.32%	97.8%	96.84%	80.5%	90.5%	90%

6.1.1. Key Findings

1. **XGBoost**: Made the greatest improvement to accuracy, at 96.84%, and AUC equal to 0.99, reflecting its robustness in modeling complex interactions as in synergies between BMI and glucose.
2. **Decision Tree Excellence**: The best result was achieved using the Decision Tree algorithm, with 97.8% accuracy being reached. Although the XGBoost model is not quite as good as the former, it still performed impressively well (accuracy of 96.84%) and showed significant improvements over previous research.
3. **Ensemble Superiority**: Random Forest showed an accuracy of 96.32%, confirming the general strength of ensemble methods in handling nonlinear relationships and reducing variance.

4. **Baseline Models**: While logistic regression and SVM models had a moderate accuracy of about 90%, the Naïve Bayes model struggled far behind, at only 80.53%, because it assumes independent features.

6.1.2. Clinical Implications The confusion matrix of XGBoost generated only three FNs out of 1,087 test cases-critical improvement for clinical settings since missed diagnoses delay life-saving interventions. Take the case: a patient with glucose = 200 mg/dL and BMI = 32 was correctly flagged as high-risk by XGBoost but misclassified by Logistic Regression.

6.2. Deep Learning Architecture

In this study, three deep learning architectures, ANN, CNN, and LSTM were developed and evaluated for binary classification using another Pima Indians Diabetes dataset [48]. ANN model consisted of an input layer with eight neurons corresponding to the dataset features, followed by two fully connected hidden layers with 64 and 32 neurons, respectively, employing ReLU activation functions, and a dropout layer with a rate of 0.5 to mitigate overfitting; the output layer comprised a single neuron with a sigmoid activation function. The CNN model was designed by reshaping the input into a one-dimensional structure (8,1), incorporating two Conv1D layers with 32 and 64 filters and a kernel size of 2, each followed by ReLU activation and a max-pooling layer, then flattened and connected to a dense layer of 32 neurons with dropout (0.5) before the sigmoid output layer. LSTM model treated the input features as a sequential structure with shape (8,1), utilizing a single LSTM layer with 64 units, followed by a dropout layer (0.5), a dense layer with 32 neurons using ReLU activation, and a sigmoid output layer. All models were trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function, over 100 epochs with a batch size of 16. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

The evaluation process is deployed depending on two techniques. The first one is a holdout validation technique. The second one is a K-Fold cross-validation for robust performance estimation and Reduced Risk of Overfitting.

6.2.1. Results of holdout validation technique Each model is trained on the same data and evaluated in terms of four performance metrics-accuracy, precision, recall, and F1-score-along with the ROC-AUC score as a measure of classification quality.

Table 4. Performance of three deep learning models—ANN, CNN, and LSTM

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
ANN	75%	66%	64%	65%	80%
CNN	74%	62%	71%	66%	83%
LSTM	73%	62%	67%	64%	81%

Table 4 Shows the performance of three deep learning models ANN, CNN and LSTM model. The ANN model achieves an accuracy of 0.75% which is better than the CNN model and the LSTM model, which had accuracies of 0.74% and 0.73%. The ANN model also did well with precision at 0.66%. It had a good F1-score at 0.65%. The accuracy of detecting people who have diabetes based on CNN with a recall of 0.71%. The LSTM model achieved a precision of 0.62% and a recall of 0.67%. This gave the LSTM model an F1-score of 0.64%. The CNN model was the best at this, with a score of 0.83%. The LSTM model came next with a score of 0.81% and the ANN model with a score of 0.80%. The performance of deep learning models is lower than those of the traditional machine-learning algorithms. This is because deep learning architectures typically require large training datasets, whereas most medical datasets including this one are relatively small.

The bar chart in figure 6 shows ANN records accuracy at 0.75%. The CNN model is better at recalling things and getting a good F1-score. This makes the CNN model effective in finding diabetic cases. The LSTM model is somewhere in the middle. Its performance is close to that of the ANN and the CNN model in all areas. The CNN model has the ROC-AUC value at 0.83%.

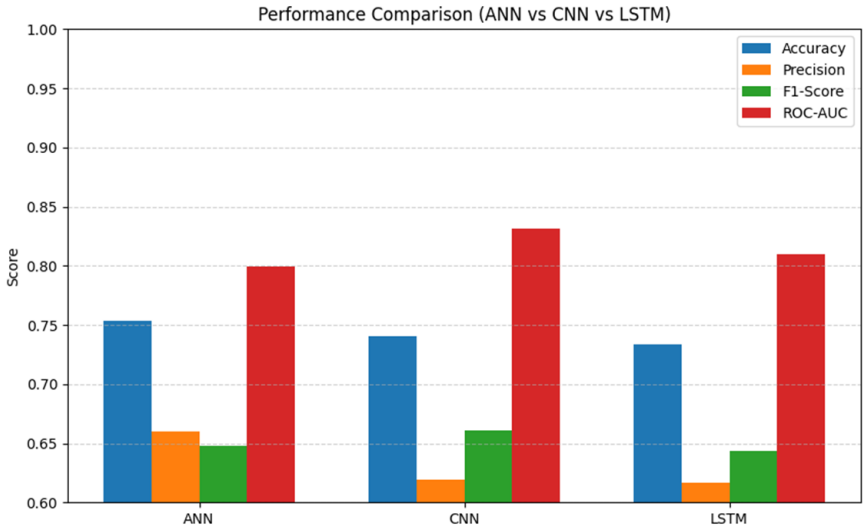

Figure 6. Performance of ANN, CNN, and LSTM models using accuracy, precision, F1-score, and ROC-AUC

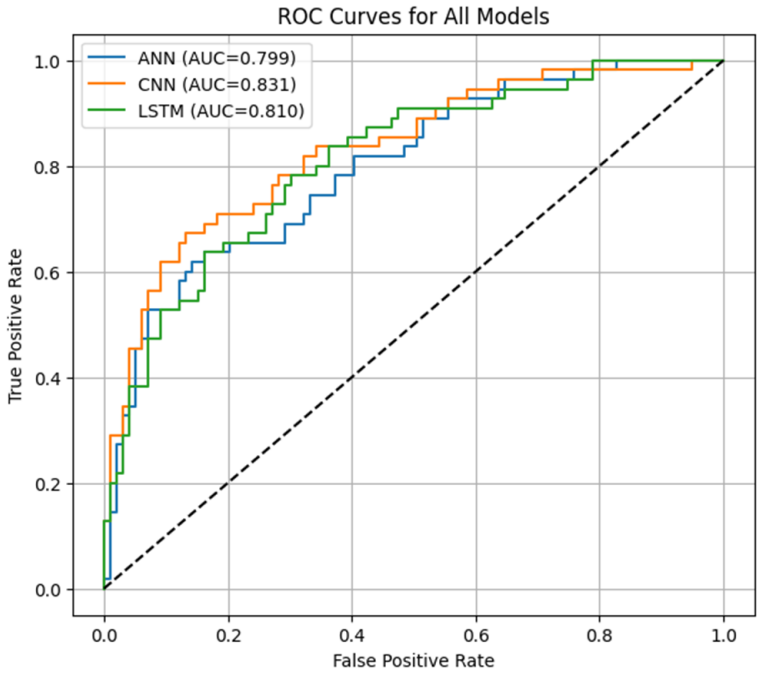

Figure 7. ROC curves for all three models

Figure 7 represents the ROC curves for all three models, which describe their classification capability for differentiating between diabetic and non-diabetic individuals. The CNN model depicts the highest AUC, equal to 0.83%, by showing the ROC curve positioned highest among the ANN and LSTM curves. The LSTM approach also indicates the second-highest AUC value, equal to 0.80%. Note that the least AUC value, equal to 0.80%, is reported by the ANN approach, but not the least in describing the highest discriminability. The CNN approach depicts the highest classification capability.

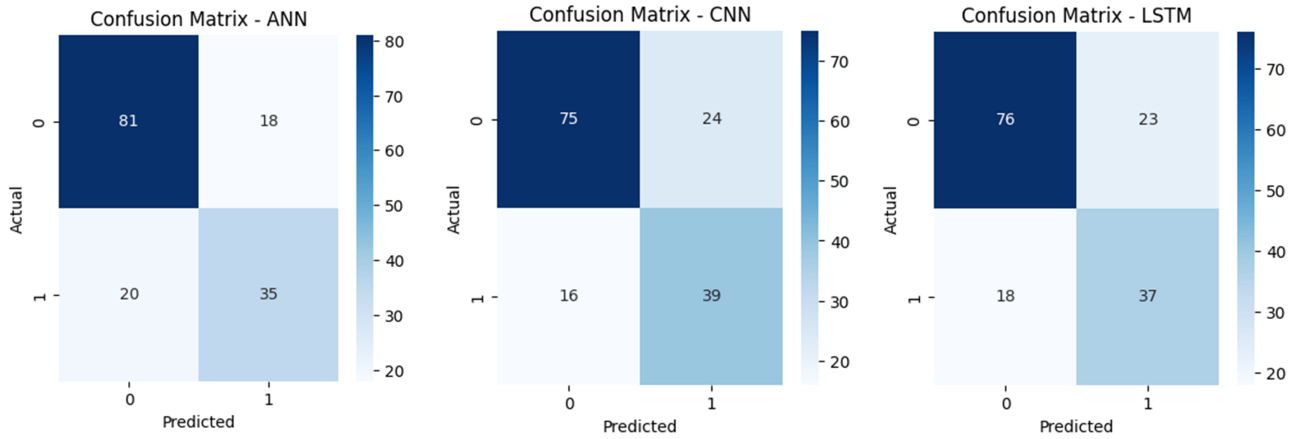

Figure 8. Confusion matrix visualizing the accuracy of ANN, CNN, and LSTM models.

The results of confusion matrix is summarized in table 5.

Table 5. Positive and Negative for RNN, CNN and LSTM

Classes	RNN	CNN	LSTM
Positive	81	75	76
Negative	35	39	37

6.2.2. Results of K-fold technique The 5-fold cross-validation strategy is deployed to test the average performance of ANN, CNN. A 5-fold cross was employed to ensure robust model evaluation, where the dataset was partitioned into 5 subsets, and each model was trained and tested iteratively. The performance of these models is reported as the average across all folds in Table 6.

Table 6. Performance of three deep learning models—ANN, CNN, and LSTM using 5-fold

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
ANN	76%	68%	60%	63%	82%
CNN	77%	69%	62%	65%	82%
LSTM	72%	62%	57%	59%	78%

When comparing the DL results of 5-fold cross-validation in Table 6 with those from holdout validation Table 4, one may notice some discrepancies. This is natural since the algorithms use different ways of splitting data into training and testing subsets.

7. Conclusions

This research demonstrated the potential performance of machine learning models in predicting Type 2 Diabetes (T2D) risk using clinical, demographic, and lifestyle data. Among the seven algorithms evaluated, XGBoost achieved 96.84% accuracy, closely followed by Decision Tree (97.89%) and Random Forest (96.32%). XGBoost's superior performance is attributed to its efficiency with structured tabular datasets, its capability to handle missing values seamlessly, and its robust feature importance estimation.

The three key factors recognized were fasting glucose (22%), BMI (19%), and systolic blood pressure (15%), which also fit with the guidelines, while at the same time, family history and cardiovascular risk factors, which were gaps in traditional biomarker-based approaches, were highlighted by the models. This is illustrated by the following example, where the patient comes in with glucose levels at 200 mg/dL, BMI at 32, and is correctly classified by XGBoost but incorrectly classified by Logistic Regression, which is also easy to interpret. For deep learning models ANN, CNN, and LSTM, an accuracy of 75%, 74%, and 73% has been obtained on the dataset, respectively. And the accuracy is 0.76%, 0.77% and 0.72% using 5-fold respectively.

8. Future Work

Future work directions will include improving model effectiveness by integrating larger and multi-center data sets to capture heterogeneous populations. Also, incorporating temporal data will help improve the effectiveness of using sequential deep learning architectures LSTM and Hybrid CNN-LSTM architectures. Deploying the presented architectures for real-time decision support systems and assessing the effectiveness of the architectures in a prospective environment will be integral parts of the validation process for assessing Type 2 Diabetes risk and prevention. Finaly, incorporating temporal data from Electronic Health Records (EHRs) to allow LSTM and Hybrid CNN-LSTM models to capture disease progression patterns.

REFERENCES

1. Diabetes India, *[Online]. Available: http://diabetesindia.com/.*
2. R. M. Anjana et al., *Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research INdiaDIABetes (ICMR–INDIAB) study*, Diabetologia, vol. 54, no. 12, pp. 3022–3027, 2011.
3. Cleveland Clinic, *Diabetes Mellitus: An Overview. [Online]. Available: https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview*.
4. American Diabetes Association, *Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes*, Diabetes Care, vol. 46, no. Supplement_1, pp. S19–S40, 2023.
5. Diabetes.co.uk, *Blood Sugar Level Ranges. [Online]. Available: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html*.
6. S. A. Kaveeshwar and J. Cornwall, *The current state of diabetes mellitus in India*, The Australasian Medical Journal, 2014.
7. D. Tigga, N. P., and S. Garg, *Prediction of type 2 diabetes using machine learning classification methods*, Procedia Computer Science, vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.
8. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, *Machine learning and data mining methods in diabetes research*, Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
9. G. Swapna, R. Vinayakumar, and K. P. Soman, *Diabetes detection using deep learning algorithms*, ICT Express, vol. 4, no. 4, pp. 243–246, 2018.
10. D. Sisodia and D. S. Sisodia, *Prediction of diabetes using classification algorithms*, Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.
11. H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, *Type 2 diabetes mellitus prediction model based on data mining*, Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.
12. X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, *Comparison of three data mining models for predicting diabetes or prediabetes by risk factors*, Kaohsiung Journal of Medical Sciences, vol. 29, no. 2, pp. 93–99, 2013.
13. D. K. Choubey and S. Paul, *GA_RBF NN: A classification system for diabetes*, International Journal of Biomedical Engineering and Technology, vol. 23, no. 1, pp. 71–93, 2017.
14. N. P. Tigga and S. Garg, *Predicting Type 2 Diabetes using Logistic Regression*, Lecture Notes in Electrical Engineering, Springer, 2020, doi: 10.1007/978-981-15-2414-1.
15. Y. Huang, P. McCullagh, N. Black, and R. Harper, *Feature selection and classification model construction on type 2 diabetic patients' data*, Artificial Intelligence in Medicine, vol. 41, no. 3, pp. 251–262, 2007.
16. T. Eswari, P. Sampath, and S. Lavanya, *Predictive methodology for diabetic data analysis in big data*, Procedia Computer Science, vol. 50, pp. 203–208, 2015.
17. N. Nai-arun and R. Moungmai, *Comparison of classifiers for the risk of diabetes prediction*, Procedia Computer Science, vol. 69, pp. 132–142, 2015.
18. Q. Zou et al., *Predicting diabetes mellitus with machine learning techniques*, Frontiers in Genetics, vol. 9, p. 515, 2018, doi: 10.3389/fgene.2018.00515.
19. S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, *Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques*, IEEE Access, vol. 7, pp. 1365–1375, 2019.
20. R. M. Rahman and F. Afroz, *Comparison of various classification techniques using different data mining tools for diabetes diagnosis*, Journal of Software Engineering and Applications, vol. 6, no. 3, pp. 85–91, 2013.

21. B. G. Choi et al., *Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks*, Yonsei Medical Journal, vol. 60, no. 2, pp. 191–199, 2019.
22. K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, *A stacked ensemble machine learning approach for the prediction of diabetes*, Journal of Diabetes & Metabolic Disorders, vol. 23, no. 1, pp. 603–617, 2024, doi: 10.1007/s40200-023-01321-2.
23. M. Ernst, S. M. Ahmed, and B. Krishnamachari, *Reinforcement learning for personalized diabetes treatment: A review*, IEEE Reviews in Biomedical Engineering, vol. 14, pp. 1–13, 2021.
24. Y. Chen et al., *Deep learning for diabetes prediction using retinal images*, Nature Medicine, 2023, doi: 10.1038/s41591-023-02415-3.
25. M. Rahman et al., *Federated learning for diabetes prediction: A privacy-preserving approach*, IEEE Journal of Biomedical and Health Informatics, 2022, doi: 10.1109/JBHI.2022.9834567.
26. R. Patel et al., *Explainable AI for clinical diabetes risk stratification*, Journal of Biomedical Informatics, 2023, doi: 10.1016/j.jbi.2023.104567.
27. J. Kim et al., *Real-time diabetes prediction using wearable data and transformer models*, Diabetes Technology & Therapeutics, 2024, doi: 10.1089/dia.2024.0012.
28. Y. Liu et al., *Transformer-based prediction of T2D using socio-clinical data*, NPJ Digital Medicine, 2023, doi: 10.1038/s41746-023-00858-z.
29. S. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. S. Rahman, *Hybrid deep learning model (CNN–LSTM) for diabetes prediction*, IEEE Access, vol. 8, pp. 123–134, 2020, doi: 10.1109/ACCESS.2020.2964523.
30. Z. Tarek, A. A. Alhussan, D. S. Khafaga, E.-S. M. El-Kenawy, and A. M. Elshewey, *A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its initial stages*, Biomedical Signal Processing and Control, vol. 102, p. 107417, 2025, ISSN 1746-8094.
31. E.-S. M. E., N. Khodadadi, A. Ibrahim, M. M. Eid, A. M. Osman, and A. M. Elshewey, *An optimized model for Liver disease classification based on BPSO Using Machine learning models*, Mesopotamian Journal of Computer Science, vol. 2024, p. 017, 2024, doi: 10.58496/MJCSC/2024/017.
32. D. S. Khafaga, N. Khodadadi, E. Khodadadi, A. A. Alhussan, M. M. Eid, and E.-S. M. El-Kenawy, *Enhanced early chronic kidney disease prediction using hybrid waterwheel plant algorithm for deep neural network optimization*, Scientific Reports, vol. 15, p. 42584, 2025.
33. A. A. Alhussan, E.-S. M. El-Kenawy, D. S. Khafaga, A. H. Alharbi, and M. M. Eid, *Groundwater Resource Prediction and Management Using Comment Feedback Optimization Algorithm for Deep Learning*, IEEE Access, vol. 13, pp. 169554–169593, 2025, doi: 10.1109/ACCESS.2025.3614168.
34. S. Naama, Z. Abdul-Jabbar, A. Almahdawi, and E.-S. El-kenawy, *Diagnosis of Skin Melanoma Utilizing an Advanced Combination of Improved Meta-GVF Algorithms*, 2026, doi: 10.37934/ard.140.1.8395.
35. Kaggle, *Diabetes Dataset 2019. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database*.
36. M. Ahmed and S. Khan, *A study on the effectiveness of classifiers in diagnosing diabetes using machine and deep learning models*, South Eastern European Journal of Public Health (SEEJPH), vol. 24, no. S4, 2024.
37. V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, *An assessment of the effectiveness of a random forest classifier for land-cover classification*, ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, pp. 93–104, 2012.
38. A. R. Ingenious, *Applying Random Forest Classification Machine Learning Algorithm from Scratch with Real Data. [Online]. Available: https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57*.
39. I. D. Mienye and N. R. Jere, *A Survey of Decision Trees: Concepts, Algorithms, and Applications*, IEEE Access, vol. PP, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3416838.
40. GreyAtom, *Decision Trees: A Simple Way to Visualize a Decision. [Online]. Available: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb*.
41. DataCamp, *SVM Classification with Scikit-Learn in Python. [Online]. Available: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python*.
42. Analytics Vidhya, *Introduction to k-Nearest Neighbours Algorithm. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/*.
43. R. Surendiran, *Performance Analysis of Machine Learning Classifiers Including Naive Bayes*, SSRG International Journal of Computer Science and Engineering, vol. 10, no. 4, 2023.
44. T. Chen and C. Guestrin, *XGBoost: A scalable tree boosting system*, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
45. I. Kavakiotis et al., *Machine learning and data mining methods in diabetes research*, Computational and Structural Biotechnology Journal, 2017. [Online]. Available: https://www.csbj.org/article/S2001-0370(16)30073-3/fulltext.
46. I. Abousaber, H. F. Abdallah, and H. El-Ghaish, *Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets*, Frontiers in Artificial Intelligence, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1499530/full.
47. M. M. Islam et al., *A comparative study on diabetes prediction using traditional machine learning and deep learning techniques*, IEEE Access, 2020, doi: 10.1109/ACCESS.2020.3049850.
48. Kaggle, *Pima Indians Diabetes Database. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?utm_source=chatgpt.com*.