NIIT UNIVERSITY
THE UNIVERSITY OF THE FUTURE

**Project ID : 24**

# R&D Project Report

**On**

# Developing an AI-driven Machine Learning Model for Detection of Non-Infectious Diseases

**Academic Year: 2024-25**

**Faculty Mentor: Dr. Vikas Upadhyaya**

**Submitted By**


| | | |
|---|---|---|
| Agnishwar Raychaudhuri | BT22GDS056 | B. Tech DS |
| Khushi Singh | BT22GCS089 | B. Tech DS |
| Geetika Agrawal | BT22GCS106 | B. Tech CSE |
| Jagriti Singh | BT22GCS145 | B. Tech CSE |

# TABLE OF CONTENTS

# LIST OF FIGURES & TABLES

# INTRODUCTION

Diabetes mellitus is a complex metabolic disorder that has emerged as one of the 21st century's most pressing health challenges. Commonly referred to as diabetes, is a chronic metabolic disorder caused by high blood glucose levels (hyperglycemia) due to the body's inability to produce or effectively use insulin. Insulin is a hormone produced by the β-cells of pancreas. It regulates blood sugar by facilitating the absorption of glucose into cells for energy production. When this process is impaired, glucose accumulates in the bloodstream, leading to various health complications that affect almost every organ system and significantly influencing quality of life. [38]From ancient Egyptian physicians who first documented its sweet-tasting symptoms to modern-day researchers racing to find innovative treatments, diabetes has commanded the attention of medical professionals for millenia. Today, it affects hundreds of millions globally, crossing geographical boundaries and socioeconomic divisions, making it not just a medical condition but a critical public health priority that demands our urgent attention and understanding.

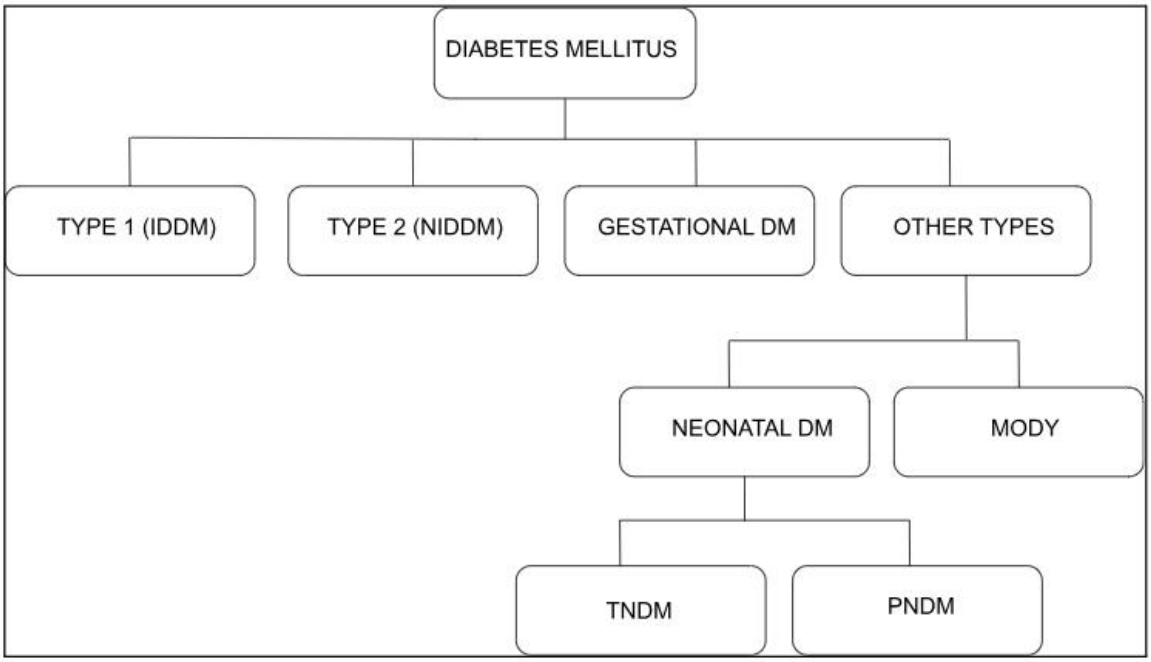


*Fig. 1: Classification of diabetes types*

Types of Diabetes : A Comprehensive Overview

The landscape of diabetes is diverse, with each type presenting its own unique characteristics, challenges and management approaches :

1. Type 1 Diabetes (T1D)

T1D also known as insulin-dependent diabetes mellitus (IDDM), occurs due to the autoimmune damage of the β-cells. This leads to suppression or cessation of the body's capacity to produce insulin thereby creating an absolute insulin deficiency. The onset typically occurs before the age of 30, with a peak in adolescence, though it can develop at any age. Patients require careful monitoring as they face a risk of diabetic ketoacidosis, a severe complication that occurs when the body, lacking insulin, begins breaking down fat too rapidly. Management involves multiple daily insulin injections or insulin pump therapy, coupled with regular blood glucose monitoring and carbohydrate counting.

2. Type 2 Diabetes (T2D)

T2D, also known as non-insulin-dependent diabetes mellitus (IDDM), accounts for approximately 90 % of all diabetes cases. It develops through a complex interplay of genetic susceptibility and environmental factors. The condition progresses through several stages, it first begins with insulin resistance, where cells become less responsive to the effect of insulin. Initially, the pancreas compensates by producing excess insulin, but gradually over time, this compensation fails, leading to declining insulin production. T2D poses risks such as obesity(particularly central adiposity), physical inactivity, poor diet, advancing age, and certain ethnicities. The progression can be influenced by interventions at various stages, from lifestyle modifications in pre-diabetes to combinations of oral medications, injectable incretin-based therapies, eventually insulin therapy in advanced stages.

3. Gestational Diabetes (GDM)

GDM emerges during pregnancy due to placental hormones that create insulin resistance, challenging the ability of the mother to maintain normal blood glucose levels. This condition affects 2-10% of pregnancies approximately and requires careful monitoring as it can have an impact on both maternal and fetal health. Complications may include macrosomia (excessive fetal growth), increased risk of cesarean delivery and neonatal hypoglycemia. Women with GDM face a 35-60% chance of developing type 2 diabetes within 10-20 years post-pregnancy, necessitating regular screening after delivery.

4. Other Specific Types

Other specific distinct forms of diabetes include

a. Monogenic Diabetes : Including Maturity Onset Diabetes of the Young (MODY), characterized by genetic defects in the functionality of the β-cells. At least 14 different types of monogenic diabetes have been identified, each with specific genetic mutations affecting insulin production or glucose regulation

b. Neonatal Diabetes Mellitus (NDM) : NDM is a rare form of diabetes diagnosed within six months of life, presenting in two primary forms : Transient Neonatal Diabetes Mellitus (TNDM) and Permanent Neonatal Diabetes Mellitus (PNDM). TNDM typically resolves within the first few months but may recur later, often resulting in genetic abnormalities.
PNDM on the other hand is a lifelong condition requiring ongoing treatment.

In recent years, Artificial intelligence (AI) and machine learning (ML) have revolutionized diabetes prediction by analyzing vast datasets, identifying patterns, and delivering highly accurate predictions beyond traditional methods. These ML models integrate diverse data sources, from basic demographics to detailed medical histories, enabling a more comprehensive and precise approach to detecting diabetes.

AI and ML play a transformative role in diabetes prediction by enabling early detection through the analysis of vast datasets such as BMI, blood glucose, skin thickness etc. These AI-ML models not only predict the likelihood of developing diabetes but also offer great help to health professionals in minimizing its risk and offering personalized treatments. From traditional methods to Deep learning models the prediction of diabetes has come a long way. Techniques like explainable AI have now become a crucial part in the medical field in order to understand which factors are most influential in these predictions.Deep Learning models now use medical images to help detect early signs of diabetic retinopathy. Despite the varying accuracy across different models and datasets, AI-ML models in diabetes prediction signify a promising future for improved real-time diagnosis and monitoring. This literature review contains papers published from 2018 onwards.

# LITERATURE REVIEW

This literature review provides a comprehensive overview of the existing research and advancements in the field of AI and machine learning to predict diabetes, highlighting the key methodologies, datasets, features and results. The study transitions from reviewing and analyzing traditional methodologies to deep learning techniques developed by researchers and professionals to predict diabetes mellitus.

Three supervised machine learning algorithms—Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN)—were used by the authors of [1] to predict diabetes. The algorithms used the following features: age, diabetes, skinfold thickness, BMI, blood pressure, insulin levels, and diabetes spectrum function.SVM was demonstrated to be particularly good for binary classification tasks because of its capability to produce optimal hyperplanes, but LR produced easy-to-understand outcomes for binary result prediction. Because neural networks follow the patterns of the brain in forming networks, combining all these patterns proved to increase accuracy.

In [2] the authors use a Deep Belief Neural Network (DBN) model for diabetes prediction, demonstrating superior performance compared to traditional classifiers through a three-phase methodology incorporating pre-processing, pre-training DBN, and fine-tuning, with PCA feature selection and automated learning capabilities for pattern detection in complex medical data; however, the study has limitations including its reliance on a single gender-specific dataset of 768 Pima Indian females, lack of cross-validation results, and absence of hyperparameter optimization discussion. The model's architecture consists of input, three hidden layers [500 500 1000], and output layers, utilizing ReLU and sigmoid activations, with RBM training for 10 epochs and Gaussian Distribution weight initialization, processing 8 input attributes and implementing neural network classification with a [500 500 1000 2] topology, SGD, and specific parameters (0.01 learning rate, 0.5 momentum). Performance metrics show DBN achieving superior results (Recall: 1.0, Precision: 0.6791, F1: 0.808) compared to other methods including Naïve Bayes (Recall: 0.759, Precision: 0.763, F1: 0.760), RBF-NN (Recall: 0.761, Precision: 0.756, F1: 0.757), Decision Tree (Recall: 0.738, Precision: 0.735, F1: 0.736), Logistic Regression (Recall: 0.73, Precision: 0.73, F1: 0.73), Random Forest (Recall: 0.71, Precision: 0.72, F1: 0.72), and SVM (Recall: 0.424, Precision: 0.651, F1: 0.513).

The authors of [3] analyze various machine learning techniques for diabetes mellitus prediction, emphasizing early detection's importance and exploring real-time data integration possibilities, while acknowledging challenges in model interpretability and dataset limitations. The study utilized a dataset of 200 patients from Chittagong, Bangladesh, with 16 attributes and implemented four algorithms with varying performance metrics: C4.5 Decision Tree performed best (Accuracy: 73.5%, Precision: 72%, Recall: 74%, Specificity: 72%, F1: 72%), followed by SVM (Accuracy: 70%, Precision: 72%, Recall: 68%, Specificity: 74%, F1: 70%), KNN (Accuracy: 68%, Precision: 70%, Recall: 66%, Specificity: 72%, F1: 68%), and Naive Bayes (Accuracy: 65%, Precision: 67%, Recall: 63%, Specificity: 69%, F1: 65%), though the study's generalizability is limited by its reliance on a single, geographically specific dataset.

It [4] demonstrates the effectiveness of Kernel-based Support Vector Machines (SVM) for diabetes classification, implementing linear, polynomial, and radial kernels to handle various data distributions, while

emphasizing the importance of early detection and diagnosis for improved patient outcomes and reduced healthcare costs; however, it faces challenges in model interpretability and dataset generalizability limitations. Using a structured workflow of data collection, preprocessing, and model evaluation on 332 records with seven input variables, the study split data into 70% training and 30% testing sets, achieving impressive results across different kernel implementations: Linear Kernel SVM showed perfect performance (Accuracy: 100%, Sensitivity: 1.0, Specificity: 1.0), followed by Radial Kernel SVM (Accuracy: 99%, Sensitivity: 0.98, Specificity: 1.0), and Polynomial Kernel SVM (Accuracy: 90%, Sensitivity: 1.0, Specificity: 0.87).

In [5] the authors explore various machine learning algorithms for diabetes prediction, emphasizing early detection's importance and the potential integration with real-time healthcare data collection, while acknowledging challenges in model interpretability and dataset limitations. Following a structured workflow incorporating data collection, preprocessing, model selection, training, and evaluation using attributes like glucose level, blood pressure, BMI, age, and insulin levels, the study compared multiple algorithms with varying performance metrics: KNN (K=10) achieved the highest accuracy (76%, Precision: 0.76, Sensitivity: 0.73, F1: 0.75), followed by SVM (Accuracy: 75%, Precision: 0.73, Sensitivity: 0.74, F1: 0.73), Naive Bayes (Accuracy: 74%, Precision: 0.74, Sensitivity: 0.74, F1: 0.74), and both Decision Tree and Random Forest showing similar results (Accuracy: 71%, with slight variations in other metrics), though the authors acknowledge the need for validation on larger, more diverse datasets to improve generalizability.

The study [7] used Decision Tree, K-nearest neighbor(KNN), Naive Bayes, and Random Forest on the PIMA dataset after a series of thorough preprocessing that included handling incomplete data and standardization. RF showed the best results with 86% accuracy and was found to be very well performing with noise and missing data. This study also proposed a cross of machine-learning models with real-time data collection from IoT sensors for enhanced healthcare applications.
The authors of [9] present a comparative analysis of machine learning algorithms for diabetes prediction using the PIMA Indian Diabetes dataset of 769 samples with 8 features, implementing comprehensive data preprocessing including null value checking, cleaning, and outlier removal. While the study demonstrates strong potential for early diabetes diagnosis and treatment planning through clear data visualizations and correlation analysis, it faces limitations including limited dataset size, lack of feature importance discussion, absence of cross-validation results, and no external validation dataset. Using an 80-20 train-test split, the comparative results show Random Forest achieving the highest accuracy at 78.57%, marginally outperforming Linear SVM (77.92%) and K-NN (77.27%), though the research acknowledges the need for addressing class imbalance and improving model interpretability for practical implementation.

To address the use of an ensemble model, the paper **[10]** highlighted the exceptional performance of Random Forest in diabetes prediction, emphasizing its ability to handle complex healthcare data. Ensemble models, like RF, integrate numerous decision trees to enhance prediction accuracy and also reduce overfitting risk. The authors performed intensive preprocessing, including normalization, handling class imbalance, and feature importance evaluation. They used measures such as precision, recall, F1-score, and accuracy to evaluate the performance of the model. This study highlighted the effectiveness of ensemble models, particularly random forests, in addressing the complexities of diabetes prediction and improving the predictive power of ML models in medical diagnostics.

In [11] the authors investigate early-stage diabetes mellitus risk prediction using machine learning algorithms on the PIMA, focusing on a hybrid stacking model. The study meticulously preprocessed the dataset, addressing missing and zero values, and performed extensive data visualization and analysis. Individual algorithm performances were evaluated, with accuracies ranging from 77.27% (Naive Bayes) to 88.31% (KNN). The proposed hybrid model, combining KNN, SVM, and Decision Tree with Logistic Regression as a meta-learner, achieved a superior accuracy of 90.62%, outperforming all individual models. However, the study is limited by its reliance on a single dataset, lack of external validation, and absence of discussions on overfitting mitigation, feature importance, and comparisons with deep learning approaches.

The authors of the study [12] used three supervised machine learning algorithms on the PIMA Indian diabetes dataset: logistic regression, random forest (RF), and decision tree (DT). This study highlighted the significance of early diagnosis for effective diabetes control, with logistic regression having the highest accuracy of 76%, followed by RF with 75%. The results showed the best performance of logistic regression in handling the binary distribution function, while also validating the robustness of random forests to noise and their ability to represent relationships.

In [13] the authors assess the performance of Logistic Regression (LR) and Random Forest (RF) for diabetes prediction using a dataset of 520 individuals from a Sylhet, Bangladesh hospital. Employing an 80-20 train-test split and 10-fold cross-validation, the research aimed to demonstrate the practical application of machine learning in early diabetes detection. Random Forest achieved a high accuracy of 99.03%, significantly surpassing Logistic Regression's 94.23%. While the study highlights the potential of machine learning in healthcare, its limitations include the small, geographically limited dataset and the evaluation of only two models, which may restrict the generalizability of the findings.
The authors of [6] indicated a move toward the use of deep learning models in the prediction of diabetes. This study used continuous oscillation deep neural networks to reduce overfitting and improve the prediction of blood glucose outcomes.

The study [8] showed significant results obtained by comparing five groups, namely Naïve Bayes, random forest, logistic regression, neural network, and SVM, using the PIMA dataset. Logistic regression showed the best performance with 77.2% accuracy, which was effective in classifying binary tasks, and showed that preliminary techniques such as process control hold utter importance.

The authors of [14] evaluated seven machine learning algorithms for diabetes prediction using the Pima Indian Diabetes Dataset, which comprises 768 female patients. The research employed thorough data preprocessing, including cleaning, feature selection, and scaling, and assessed model performance using accuracy, sensitivity, specificity, F1-score, ROC and AUC. Notably, Support Vector Machine (SVM) and Decision Tree (DT) achieved perfect scores across all metrics (1.0000), while Logistic Regression reached 98.16% accuracy. Gradient Boosting and Random Forest both achieved 94.11% accuracy, K-Nearest Neighbor 90%, and Naive Bayes 89.74%. While the study demonstrates high accuracy, the dataset's limited demographic and size, along with the lack of external validation, raise concerns about the models' generalizability. The unusually perfect performance of SVM and DT also warrants further investigation.

The research paper [15] analyzed the use of machine learning algorithms for predicting diabetes by focusing on improving accuracy through data pre-processing and the use of SVM, RFC, and DNN algorithms. The study used data from the National Institute of Diabetes and Digestive and Kidney Diseases. The authors pre-processed the data with dummy variables and PCA, and achieved the highest accuracy of 89% with DNN, exceeding SVM and RFC. Despite emphasizing the significance of data pre-processing, this study identified various gaps, such as the limited exploration of ensemble techniques and relies on a single dataset, which may not fully represent the diverse population varying in demographics, lifestyle factors, and healthcare access to individuals.

This research paper [16] explored diabetes prediction using machine learning on a large dataset of 70,000 clinical records, and the Pima Indian Diabetes Database. The study implemented Logistic Regression, Random Forest, SVM, and KNN, emphasizing the importance of early diabetes detection. The methodology included data preprocessing, model training, and performance evaluation using various metrics. However, the interpretability of complex models and the reliance on specific datasets pose limitations to generalizability. On the 70,000 patient dataset, Random Forest achieved the highest accuracy at 79%, followed by SVM (77%), Logistic Regression (76%), and KNN (69%). On the Pima dataset, Random Forest again performed best with 80% accuracy, followed by SVM (77%), and Logistic Regression and KNN both at 73%.

This review paper [17] examines supervised and unsupervised learning techniques for diabetes prediction, detailing the KDD process and comparing algorithms like Decision Trees, SVM, Naive Bayes, and K-means. The authors highlight the potential of combining supervised and unsupervised methods, such as SVM with K-means, to enhance prediction accuracy. However, the study's reliance on datasets like the Pima Indian Diabetes dataset limits generalizability, and it lacks in-depth algorithm performance comparisons across diverse datasets. The paper provides a comparative table of algorithm performance, including AdaBoost (98.8%), Random Forest (94.10%), XGBoost (88.1%), K-means (78%), Artificial Neural Networks (75.7%), and a combined SVM and K-means (99.64%).

The researchers of [22] examined various ML approaches applied to diabetic datasets to aid in the early diagnosis and management of Diabetes Mellitus. The paper highlighted the use of ML techniques with Big Data Analytics tools such as Hadoop and MapReduce, as well as classifiers like Naïve Bayes, Decision Trees, SVM, KNN, Random Forest, and Gradient Boosting.

The study [18] explored diabetes mellitus prediction using machine learning, emphasizing the effectiveness of XGBoost for early detection. The research compared XGBoost, SVM, Naïve Bayes, Decision Tree, and Random Forest across three models: diabetes prediction, type 1 vs. type 2 classification, and prediabetes prediction. The methodology involved data preprocessing, including handling missing values, encoding categorical data, and feature scaling, followed by an 80:20 train-test split. However, concerns regarding XGBoost's potential for overfitting, the interpretability of complex models, and the limited dataset diversity were noted. In Model A (diabetes prediction), XGBoost achieved 92.5% accuracy. In Model B (type 1 vs. type 2 classification), Random Forest reached 89.2% accuracy. In Model C (prediabetes prediction), SVM attained 85.3% accuracy.

In [19] the authors explore machine learning-based diabetes prediction across Bangladesh, India, and

Germany, highlighting the significant diabetes burden in Bangladesh. The research utilized nine algorithms, including boosting methods like AdaBoost, CatBoost, Gradient Boost, and XGBoost, and employed ADASYN oversampling to address class imbalance. However, data availability, quality, and dataset size disparities posed challenges. The study lacked detailed feature selection analysis. For the Bangladesh dataset (14,401 records), boosting algorithms achieved near-perfect accuracy (99.9-100%). For the PIMA Indian dataset (768 records), CatBoost performed best with 83.1% accuracy. For the German dataset (2,000 records), AdaBoost and CatBoost achieved 99% accuracy.

The study by **[20]** explored machine learning algorithms for early diabetes prediction using the Pima Indians Diabetes Dataset (PIDD). The research employed thorough data preprocessing, including mean imputation for missing values, oversampling for class imbalance, and z-score normalization. Eight classifiers were evaluated, with XGBoost achieving the highest accuracy of 89.07%. However, the study's limitations include reliance on a single, gender-specific dataset, absence of cross-validation, limited hyperparameter optimization, and lack of comparisons with deep learning methods. The dataset's small size, limited feature importance analysis, and absence of external validation also pose concerns. LightGBM (88.28%), Random Forest (88.15%), SVM (85.39%), Logistic Regression (84.86%), KNN (84.07%), Naïve Bayes (82.13%), and Decision Tree (80.12%) also demonstrated varying levels of accuracy.

In **[21]** the authors investigated diabetes prediction in teenagers using machine learning algorithms, focusing on Logistic Regression, KNN, SVM, Random Forest, and XGBoost. However, the interpretability of complex models and the limited dataset from 150 students at Dayananda Sagar University were identified as challenges. The methodology involved data preprocessing, model training, and evaluation using various metrics. Random Forest and XGBoost both achieved the highest accuracy of 96.49%, with similar sensitivity, specificity, F1-score, and AUC values. SVM reached 82% accuracy, Logistic Regression 79%, and KNN 58%.

This systematic overview in [22] explored diabetes mellitus prediction using various machine learning techniques, including SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting. However, challenges related to the interpretability of complex models and the reliance on a single dataset were noted. The research followed a structured workflow, encompassing data preprocessing and model evaluation. The paper compiled accuracy metrics from various studies, highlighting a highest reported accuracy of 99.04% using a 1D CNN, Random Forest achieving 97.5% accuracy, and a basic model using SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting resulting in 77% accuracy.

In **[23]** the authors developed a diabetes prediction method using classification and ensemble learning algorithms, including Random Forest, KNN, Label Encoder, and train-test split, on the PIMA Indian Diabetes Dataset. The research included detailed data preprocessing, addressing missing values and data imbalance through oversampling. However, the study's reliance on a single dataset and the complexity of models like Random Forest were noted as limitations. The paper evaluated KNN, Random Forest, Decision Tree, and SVM using accuracy, precision, recall, and F1-score. Random Forest achieved 98% accuracy across all metrics. Decision Tree reached 96% accuracy, with 95% precision, 98% recall, and 97% F1-score. KNN showed 76.56% accuracy, 78.8% precision, 76.5% recall, and 77.6% F1-score. SVM had 65% accuracy, 63% precision, 97% recall, and 77% F1-score.

Research by [24] demonstrated the use of various ML models for early diabetes diagnosis, including KNN, SVM, Gradient Boosting, Naïve Bayes, and LR. KNN was found to be the best-performing algorithm, with 75% accuracy, and had proven its usefulness when used with Flask for real-time prediction. This study addressed issues such as data quality and sampling bias while indicating the potential of AI to transform healthcare by providing insights into diabetes risk.

The research paper [25] studied the application of ML classifiers for predicting diabetes mellitus. The study utilized five ML models—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and K-nearest neighbors (KNN)—to evaluate their effectiveness in diabetes prognosis. The RF classifier achieved the highest accuracy of 92.23%, demonstrating its ability to model complex, nonlinear relationships, while LR showed a lower accuracy of 74.42%. This study highlighted the potential of ML to enable early diagnosis through intensive data preprocessing, which included handling missing values, feature scaling, and correlation adjustments. The evaluation framework included accuracy, precision, recall, and F1-score measures which identified RF as the most effective model.

Various machine learning models have recently seen a rise in the prediction of diabetes in patients. One such study [26] focused on using genetic algorithm-based feature selection and classification methods to predict diabetes. This study addressed class imbalance using the ADASYN technique and demonstrated significant accuracy improvements with GA-based feature selection. The authors used two diabetic datasets in which the accuracy increased from 84.5% to 90.3% in Diabetic Dataset-1 (DD-1) and from 94.5% to 97.6% in Diabetic Dataset-2 (DD-2).

The research paper [27] explored diabetes prediction using Gaussian Naïve Bayes (GNB) and Artificial Neural Networks (ANN), culminating in an ensemble model. The research achieved high accuracy through thorough data preprocessing, including PCA for dimensionality reduction, and an 80-20 train-test split. The ANN model alone reached 98.7% accuracy, while the ensemble model, combining ANN and GNB via majority voting, achieved 94.1% accuracy with 100% precision. GNB alone achieved 89.6% accuracy. However, the study's reliance on the PIMA Indians dataset, lack of class imbalance handling, limited discussion of real-world implementation challenges, and lack of model interpretability were noted as limitations. ANN achieved 98.7% accuracy, 100% precision, 98.2% F1-score and 96.4% recall. GNB achieved 89.6% accuracy, 85.9% precision, F1-score and recall. The Ensemble model achieved 94.1% accuracy, 100% precision, 91.4% F1-score and 84.2% recall.

This study by the authors of [28] proposed an ARIMA-ELMAN-ANN hybrid model for diabetes prediction, achieving 96.31% overall accuracy. The model combined time-series analysis (ARIMA), recurrent neural networks (ELMAN), and nonlinear modeling (ANN), and utilized F-score based feature selection. The research employed robust data preprocessing to handle missing values. However, the study lacked specifics on dataset size, demographic representation, and class imbalance handling. Methodological limitations included insufficient comparison with other models, limited discussion of model interpretability, and absence of external validation. The hybrid model achieved 96.31% overall accuracy, with approximately 96.43% training accuracy after 100 epochs. Model building times were: ANN (19 seconds), ARIMA with feature selection (12 seconds), and the hybrid model (4.2 seconds). The fitness function reached an optimal value of 0.021 after 18 iterations, and training/validation loss showed continuous improvement.

This paper [29] presented a comprehensive overview of machine learning applications in personalized diabetes prediction, covering various algorithms and their implementations, including traditional and advanced methods. The research included a bibliometric analysis highlighting global research trends and an extensive comparative analysis of algorithm accuracies, ranging from 77.37% to 98.9%. Challenges identified included data imbalance, limited data availability, feature selection difficulties, model generalizability, interpretability, privacy concerns, and lack of standardized validation. The study categorized diabetes types and evaluated algorithms like Logistic Regression, Decision Trees, Random Forest, SVM, Neural Networks, Naïve Bayes, KNN, and Gradient Boosting. The paper compiled accuracy metrics from various studies, with highest reported accuracies of 98.9% (LGBM and Random Forest), 98% (ensemble methods), and 95.83% (RFWBP), and lower range accuracies of 77.37% (SVM) and 80% (Random Forest).

This study[30] developed an explainable AI model for diabetes prediction using a stacking classifier on the PIMA Indian Diabetes dataset. The research employed a comprehensive preprocessing pipeline, including KNN imputation, OCSVM for anomaly detection, and SMOTE+ENN for class imbalance. The stacking classifier, combining KNN, SVM, and XGB with Random Forest as a meta-classifier, achieved 98% accuracy. The integration of LIME provided model interpretability, addressing the "black box" problem. However, the study's reliance on a single, gender-specific dataset, lack of computational overhead discussion, limited XAI technique exploration, and potential dataset bias were noted as limitations. The framework included data preprocessing, ensemble model architecture, and an explainability layer. The model achieved 98% accuracy, 99% precision, 98% recall, and 99% F1-score.

In [31] the authors developed an ensemble deep learning model for diabetes prediction, combining LSTM, DNN, and CNN with a soft voting classifier. The challenges faced were related to the interpretability of complex deep learning models and the reliance on specific datasets were noted. The methodology included data preprocessing, model training, and evaluation using various metrics. The ensemble model achieved 99.81% accuracy, 99.45% precision, 99.8% sensitivity, and 99.72% F1-score.

This systematic review paper [32] provides a comprehensive overview of machine learning (ML) and deep learning (DL) techniques for diabetes mellitus detection and management. It analyzes traditional methods like SVM and KNN, as well as advanced approaches like ANN and CNN, documenting their performance metrics. The paper notes accuracy ranges from 68% for retinopathy models to 99.78% for diabetes detection using neural networks and SVM. The review evaluates various ML algorithms, highlighting performance metrics such as 99.78% accuracy using SVM and ANN, 98% using LSTM, 98.07% using KNN, and 96% using Random Forest and Fuzzy Neural Network.

The study [33] utilized LightGBM with SMOTE analysis to classify diabetic patients using the PIMA Indian Diabetes dataset. The research employed ANOVA for feature selection and SHAPE for model interpretability. However, the study's reliance on a dataset limited to Pima Indian females and the use of ANOVA for feature selection were noted as limitations. The model achieved 72% accuracy, 68% precision, 72% recall, and 70% F1-score.

The authors of [34] utilized a hybrid Grey Wolf and Dipper Throated Optimization (GWDTO) algorithm for feature selection, combined with a Convolutional Autoencoder (Conv-AE) for diabetes prediction using

the PIMA Indian Diabetes Dataset. The research employed Min-Max scaling for data preprocessing and achieved 99.10% accuracy, outperforming traditional techniques. However, limitations included reliance on a single dataset, limited discussion of class imbalance, lack of cost-benefit analysis, absence of cross-validation, and insufficient comparison with other optimization techniques. The study also lacked analysis of model interpretability and scalability. The GWDTO-ConvAE method achieved 99.10% accuracy, 97.32% precision, 97.31% recall, 97.42% F1-score, and 97.34% specificity.

In **[35]** the authors explored machine learning and deep learning approaches for diabetes prediction on the PIMA Indian Diabetes Dataset, using AdaBoost, XGBoost, and RNNs.The research incorporated IQR for outlier detection and employed detailed preprocessing, including Min-Max scaling. However, the study's reliance on a single dataset and the underperformance of AdaBoost and XGBoost were noted limitations. The RNN model achieved the highest accuracy. IQR with XGBoost resulted in 70.8% accuracy, 58.1% precision, 65.5% recall, 61.5% F1-score, and 76.7% ROC-AUC. IQR with AdaBoost yielded 73.4% accuracy, 62.5% precision, 63.6% recall, 63.1% F1-score, and 78.6% ROC-AUC. IQR with RNN achieved 90.3% accuracy, 88.5% precision, 83.6% recall, 85.9% F1-score, and 85% ROC-AUC.

The paper **[36]** explored diabetes prediction using ensemble learning and LIME for interpretability on the Diabetes Prediction Dataset. The research utilized various ML algorithms, including Random Forest, SVM, Naive Bayes, Decision Tree, Neural Networks, and K-means clustering, and employed RFE for feature selection. Detailed data preprocessing and EDA were conducted. However, limitations included reliance on a specific dataset, potential class imbalance issues, and limited discussion of real-world implementation. The Neural Network model achieved the highest accuracy of 97.21%. Performance metrics for other models were: Logistic Regression with RFE (95.99% accuracy), SVM with RFE (96.30% accuracy), Random Forest with RFE (97.06% accuracy), Gradient Boosting with RFE (97.25% accuracy), Voting Classifier with RFE (97.13% accuracy), Naive Bayes (92.30% accuracy), Decision Tree (85.56% accuracy), and K-means Clustering (91.44% accuracy).

# DESCRIPTION OF THE DATASET

For our R&D, we have worked with the **PIMA Indian Diabetes Dataset**, a well-known benchmark in medical machine learning. This dataset, provided by the **National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)**, has been widely used for diabetes prediction. It specifically focuses on diagnosing diabetes in **Pima Indian women.**

**Features in the Dataset**

| Feature | Description |
|---|---|
| Pregnancies | Number of times the patient has been pregnant |
| Glucose | Plasma glucose concentration (mg/dL) during an oral glucose tolerance test |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-hour serum insulin (mu U/ml) |
| BMI | Body Mass Index (weight in kg / (height in m)$^2$) |
| DiabetesPedigreeFunction | A function measuring diabetes likelihood based on family history |
| Age | Age of the patient (years) |
| Outcome | Target variable (0 = Non-diabetic, 1 = Diabetic) |

*Table 1 : Features of the PIMA Dataset*

# INITIAL FINDINGS

For our initial analysis, we trained the dataset using three different approaches:

1. **Combined Dataset (PIMA + Frankfurt Medical College)** – We merged the PIMA Diabetes Dataset with a dataset from the Medical College of Frankfurt to enhance diversity and improve model generalization.
2. **Augmented Dataset with SMOTE** – We applied data augmentation techniques to the combined dataset and addressed class imbalance using Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples for the minority class to ensure a more balanced dataset.
3. **Combined Dataset** – We created a final dataset by combining the original datasets (PIMA + Frankfurt) with the augmented data, using both real and synthetic samples hoping to improve model performance.

After generating the datasets, we proceeded to evaluate them using the models referenced in the literature review. To ensure a comprehensive analysis, we first validated these models on the original dataset, assessing their performance based on **Accuracy, Precision, Recall, and F1 Score**. We then tested the two newly created datasets—**the augmented dataset and the combined dataset**—with the same models to compare their effectiveness and measure improvements in predictive performance.

**Accuracy** – Measures the overall correctness of a model by calculating the proportion of correctly classified instances out of all instances.

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

Best used when the dataset is **balanced**, as it gives an equal weight to all classes.

**Precision** – Indicates how many of the predicted positive instances are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Important in cases where **false positives** need to be minimized.

**Recall (Sensitivity or True Positive Rate)** – Measures how well the model identifies actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Useful when **false negatives** are costly.

**F1 Score** – A harmonic mean of Precision and Recall, balancing the two metrics when there's an **imbalance in the dataset**.

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

Ideal when both **false positives and false negatives** have significant consequences.

| Model | PIMA Dataset (Performance Parameters) | | | | |
|---|---|---|---|---|---|
| Method | Accuracy | F1 | Recall | Precision | References |
| Logistic Regression | 0.73 - 0.9816 | 0.53 - 0.9816 | 0.44 - 0.9816 | 0.67 - 0.8953 | [2],[8],[11],[12],[13],[14],[16],[20],[21],[22],[24],[25],[36] |
| Random Forest | 0.71 - 0.9903 | 0.51 - 0.98 | 0.59 - 0.98 | 0.70 - 0.98 | [2],[5],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17],[18],[20],[21],[22],[23],[24],[25],[29],[32],[36] |
| Decision Tree | 0.6818 - 1 | 0.57 - 98 | 0.62 - 1 | 0.60 - 95 | [2],[3],[5],[6],[7],[10],[11],[12],[14],[20],[22],[23],[24],[25],[36] |
| XGBoost | 0.708 - 0.925 | 0.615 - 0.905 | 0.655 - 0.9 | 0.581 - 0.91 | [17],[18],[19],[20],[21],[35],[36] |
| AdaBoost | 0.734 - 0.991 | 0.631 - 0.71 | 0.636 - 0.73 | 0.625 - 0.71 | [10],[17],[35],[37] |
| ANN | 0.9034 - 0.987 | 0.8595 - 0.982 | 0.8345 - 0.964 | 0.8805 - 1 | [6],[27],[28],[32] |
| KNN | 0.58 - 0.882 | 0.60 - 865 | 0.68 - 876 | 0.52 - 0.888 | [3], [5],[7],[10],[11],[16],[20],[21]],[22],[23],[24],[25],[32], [37] |
| Naive Bayes | 0.65 - 0.923 | 0.616 - 0.8129 | 0.645 - 0.8974 | 0.597 - 0.8444 | [2],[3],[5],[6],[7],[8],[11],[14],[20],[24],[36],[37] |
| Light GBM | 0.8828 - 0.989 | 0.83 - 0.889 | 0.796 - 0.88 | 0.87 - 0.933 | [20],[29] |
| Gradient Boosting | 0.77 - 0.941 | 0.9411 | 0.9411 | NA | [14],[22],[36] |
| Cat Boost | 0.831 - 0.9819 | 0.813 - 0.9730 | 0.816 - 0.9626 | 0.81 - 0.9836 | [19],[26] |

*Table 2 : Performance of various ML Models in the Literature Review*

| Model | Original Dataset | | | | Augmented Dataset | | | | Combined Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Accuracy | F1 | Recall | Precision | Accuracy | F1 | Recall | Precision | Accuracy | F1 | Recall | Precision |
| Logistic Regression | 0.7742 | 0.6271 | 0.5502 | 0.7291 | 0.7617 | 0.7605 | 0.7592 | 0.7617 | 0.7645 | 0.6699 | 0.6107 | 0.7419 |
| Random Forest | 0.9819 | 0.973 | 0.9626 | 0.9836 | 0.91 | 0.91 | 0.9192 | 0.901 | 0.9916 | 0.9894 | 0.9867 | 0.992 |
| Decision Tree | 0.9838 | 0.9756 | 0.9626 | 0.989 | 0.8625 | 0.8615 | 0.8636 | 0.8593 | 0.9706 | 0.9628 | 0.9602 | 0.9653 |
| XGBoost | 0.9061 | 0.74 | 0.735 | 0.735 | 0.8675 | 0.87 | 0.87 | 0.87 | 0.8994 | 0.9 | 0.9 | 0.89 |
| AdaBoost | 0.991 | 0.71 | 0.71 | 0.71 | 0.91 | 0.91 | 0.91 | 0.91 | 0.9518 | 0.95 | 0.95 | 0.95 |
| ANN | 0.944 | 0.77 | 0.77 | 0.77 | 0.8985 | 0.91 | 0.91 | 0.905 | 0.9724 | 0.97 | 0.97 | 0.97 |
| KNN | 0.866 | 0.8672 | 0.8664 | 0.8687 | 0.8233 | 0.8296 | 0.8629 | 0.7988 | 0.9686 | 0.9601 | 0.9679 | 0.9525 |
| Naive Bayes | 0.777 | 0.6365 | 0.6022 | 0.675 | 0.765 | 0.7573 | 0.7358 | 0.7801 | 0.7431 | 0.6495 | 0.6021 | 0.705 |
| Light GBM | 0.9531 | 0.9297 | 0.9198 | 0.9399 | 0.8825 | 0.8828 | 0.8939 | 0.8719 | 0.9686 | 0.96 | 0.9549 | 0.9651 |
| Gradient Boosting | 0.8809 | 0.8125 | 0.7647 | 0.8667 | 0.8175 | 0.8224 | 0.8535 | 0.7934 | 0.8878 | 0.852 | 0.817 | 0.8902 |
| Cat Boost | 0.9819 | 0.973 | 0.9836 | 0.9626 | 0.9275 | 0.9284 | 0.9495 | 0.9082 | 0.9948 | 0.9934 | 0.992 | 0.9947 |

*Table 3 : Verification of existing ML Models*

# TIMELINE

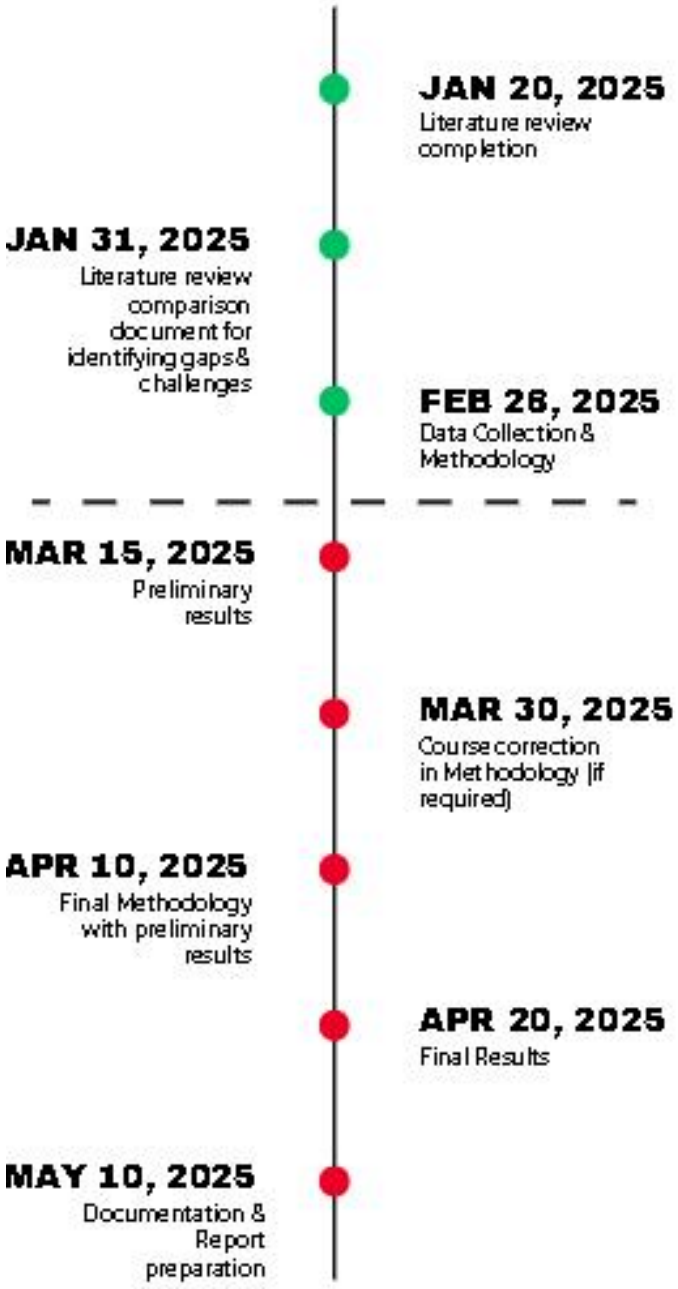


*Fig. 2 : Timeline of the Milestones*

# REFERENCES

[1] Joshi, T. N., & Chawan, P. M. (2018). Diabetes Prediction Using Machine Learning Techniques. *Journal of Engineering Research and Application Www.Ijera.Com*, *8*, 2248–9622. https://doi.org/10.9790/9622-0801020913

[2] Prabhu, P., & Selvabharathi, S. (2019). Deep belief neural network model for prediction of diabetes mellitus. *2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)*, 138–142. https://doi.org/10.1109/ICISPC.2019.8935838

[3] Faruque, M. F., Asaduzzaman, & Sarker, I. H. (2019). Performance analysis of machine learning techniques to predict diabetes mellitus. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1-4. https://doi.org/10.1109/ECACE.2019.8679365

[4] Vijayan, V. V., & Anjali, C. (2016). Prediction and diagnosis of diabetes mellitus - A machine learning approach. *2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*, 122–127. https://doi.org/10.1109/RAICS.2015.7488400

[5] Lyngdoh, A. C., Choudhury, N. A., & Moulik, S. (2021). Diabetes Disease Prediction Using Machine Learning Algorithms. *Proceedings - 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2020*, 517–521. https://doi.org/10.1109/IECBES48179.2021.9398759

[6] Bhargava, R., & Dinesh, J. (2021). Deep Learning based System Design for Diabetes Prediction. *2021 International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON 2021*. https://doi.org/10.1109/SMARTGENCON51891.2021.9645906

[7] Zaman, S. M. T., Paul, S. K., Paul, R. R., & Hamid, M. E. (2021). Detecting diabetes in human body using different machine learning techniques. *2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 1–6. https://doi.org/10.1109/IC4ME253898.2021.9768501

[8] Saxena, R., Sharma, S. K., & Gupta, M. (2021). Analysis of machine learning algorithms in diabetes mellitus prediction. *Journal of Physics: Conference Series*, *1921*(1). https://doi.org/10.1088/1742-6596/1921/1/012073

[9] Pal, M., Parija, S., & Panda, G. (2021, August 5). Improved prediction of diabetes mellitus using machine learning based approach. *2nd International Conference on Range Technology, ICORT 2021*. https://doi.org/10.1109/ICORT52730.2021.9581774

[10] Xu, X., Huang, X., Ma, J., & Luo, X. (2021). Prediction of Diabetes with its Symptoms Based on Machine Learning. *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering, CSAIEE 2021*, 147–156. https://doi.org/10.1109/CSAIEE54046.2021.9543343

[11] Samet, S., Laouar, M. R., & Bendib, I. (2021). Diabetes mellitus early stage risk prediction using machine learning algorithms. *5th International Conference on Networking and Advanced Systems, ICNAS 2021*. https://doi.org/10.1109/ICNAS53565.2021.9628955

[17] Sivaraman, M., & Sumitha, J. (2022). Diabetes Prediction based on Supervised and Unsupervised Learning Techniques - A Review. *3rd International Conference on Smart Electronics and Communication, ICOSEC 2022 - Proceedings*, 1292–1296. https://doi.org/10.1109/ICOSEC54921.2022.9952107

[18] Omoora, E. S., Altaweil, H. A., Nagem, T., & Bozed, K. A. (2023). Diabetes Mellitus Prediction Based on Machine Learning Techniques. *2023 IEEE 11th International Conference on Systems and Control, ICSC 2023*, 225–231. https://doi.org/10.1109/ICSC58660.2023.10449831

[19] Shampa, S. A., Islam, M. S., & Nesa, A. (2023). Machine Learning-based Diabetes Prediction: A Cross-Country Perspective. *2023 International Conference on Next-Generation Computing, IoT and Machine Learning, NCIM 2023*. https://doi.org/10.1109/NCIM59001.2023.10212596

[20] Sarkar, P., & Pawar, S. (2023). Machine Learning based Early Predication and Detection of Diabetes Mellitus. *International Conference on Artificial Intelligence for Innovations in Healthcare Industries, ICAIIHI 2023*. https://doi.org/10.1109/ICAIIHI57871.2023.10489259

[21] Vrindavanam, J., Haarika, R., MG, S., & Kumar, K. S. (2023). Diabetes prediction in teenagers using machine learning algorithms. *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 343-347. IEEE.

[22] Krishna Manaswini, T., Nayak, P., Harshitha, V. S., & Barlapudi, S. (2023). Predictions of Diabetic Mellitus using ML Techniques: A Systematic Overview. *International Conference on Sustainable Computing and Smart Systems, ICSCSS 2023 - Proceedings*, 43–47. https://doi.org/10.1109/ICSCSS57650.2023.10169244

[23] Parimala, G., Kayalvizhi, R., & Nithiya, S. (2023). Diabetes Prediction using Machine Learning. *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*. https://doi.org/10.1109/ICCCI56745.2023.10128216

[24] Sehgal, D., Gautam, B., Kaur, I., Singh, A., Sharma, V., & Kumar, N. (n.d.). *AI-Driven Early Diabetes Prediction*. https://doi.org/10.1109/AIC.2024.7

[25] Kumar, A., Gill, A. S., Singh, J. P., & Ghosh, D. (2024). A Comprehensive and Comparative Examination of Machine Learning Techniques for Diabetes Mellitus Prediction. *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*. https://doi.org/10.1109/ICCCNT61001.2024.10725693

[26] Ravi Kiran, T. S., Srisaila, A., Shankar, G. S., Sowjanya, B., & Lakshmanarao, A. (2024). Machine Learning Approach for Diabetes Prediction using Genetic Algorithm based Feature selection. *2024 3rd International Conference for Innovation in Technology, INOCON 2024*. https://doi.org/10.1109/INOCON60754.2024.10511558

[27] Chandra, T. B., Reddy, A. S., Adarsh, A., Jabbar, M. A., & Jyothi, B. N. (2024). Diabetes Prediction Using Gaussian Naive Bayes and Artificial Neural Network. *International Conference on Distributed Computing and Optimization Techniques, ICDCOT 2024*. https://doi.org/10.1109/ICDCOT61034.2024.10516226

[28] Senthil, J., Akbar, S., Akiladevi, N., Praveena, S., Ravindar, K., & Banupriya, V. (2024). Early Prediction of Diabetes and its Risk Factors based on ARIMA-ELMAN ANN Network. *2nd International Conference on Intelligent Data Communication Technologies and Internet of Things, IDCIoT 2024*, 1376–1381. https://doi.org/10.1109/IDCIoT59759.2024.10468045

[29] Kaur, I., & Ali, A. (2024). An In-Depth Exploration of Machine Learning Algorithms and Performance Evaluation Approaches for Personalized Diabetes Prediction. *Proceedings - 2024 International Conference on Emerging Innovations and Advanced Computing, INNOCOMP 2024*, 532–538. https://doi.org/10.1109/INNOCOMP63224.2024.00093

[30] Aruna Devi, B., & Karthik, N. (2024). Explainable Artificial Intelligence for Prediction of Diabetes using Stacking Classifier. *Proceedings of CONECCT 2024 - 10th IEEE International Conference on Electronics, Computing and Communication Technologies*. https://doi.org/10.1109/CONECCT62155.2024.10677165

[31] Aouamria, S., Boughareb, D., Nemissi, M., Kouahla, Z., & Seridi, H. (2024). International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING An Ensemble Deep Learning Model for Diabetes Disease Prediction. In *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE* (Vol. 2024, Issue 4). www.ijisae.org

[32] Katiyar, N., Thakur, H. K., & Ghatak, A. (2024). Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, *9*. https://doi.org/10.1016/j.prime.2024.100661

[33] Singh Gill, K., Anand, V., Chauhan, R., & Pokhariya, H. S. (2024). Using Machine Learning-based SMOTE Analysis with the Light GBM Classification Method to Classify Diabetic Patients. *2024 3rd International Conference for Innovation in Technology, INOCON 2024*. https://doi.org/10.1109/INOCON60754.2024.10511914

[34] Dattangire, R., Pamulaparthyvenkata, S., Mandvikar, S., Balakrishnan, A., & Chintale, P. (2024). AI/ML-Based Diabetes Application using Hybrid Grey Wolf and Dipper Throated Optimization Algorithm. *International Conference on Intelligent Algorithms for Computational Intelligence Systems, IACIS 2024*. https://doi.org/10.1109/IACIS61494.2024.10721678

[35] Chandra Sekhar Reddy, L., Gottipalli, M., Sravanthi, P., Rajanikanth, J., Yalamarthi, G., & Gurrapu, N. (2024). Bridging Horizons in Diabetes Prediction: A Comparative Exploration of Machine Learning and Deep Learning Approaches in Pima Indian Women. *Proceedings - 2nd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2024*, 386–391. https://doi.org/10.1109/InCACCT61598.2024.10550977

[36] Gilani, S. A. H., Syed, M. H., & Anjum, A. (2024). Effective Diabetes Prediction: Integrating Ensemble Learning with LIME for Robust Results. *2024 International Conference on Frontiers of Information Technology (FIT)*, 1–6. https://doi.org/10.1109/FIT63703.2024.10838461

[37] Bardia, V., & Sophiya, E. (2024). Diabetes Prediction Using Machine Learning Algorithm: A Comparative Analysis. *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, 1973–1979. https://doi.org/10.1109/ICACCS60874.2024.10717264

[38] Rubin, G., & M King, K. (2013). A history of diabetes: from antiquity to discovering insulin. British *Journal of Nursing, www.magonlinelibrary.com, 12*. https://doi.org/10.12968/bjon.2003.12.18.11775