Diabetes Prediction Based on Machine Learning Techniques: A Review

Agnishwar Raychaudhuri[1*], Khushi Singh[1†], Geetika Agrawal[2†]
and Jagriti Singh[2†]

[1*]Data Science, NIIT University, Neemrana, 301705, Rajasthan, India.
[2*]Computer Science, NIIT University, Neemrana, 301705, Rajasthan,
India.

*Corresponding author(s). E-mail(s):
agnishwar.raychaudhuri22@st.niituniversity.in;
Contributing authors: khushi.singh22@st.niituniversity.in;
geetika.agrawal22@st.niituniversity.in; jagriti.singh22@st.niituniversity.in;
[†]These authors contributed equally to this work.

Abstract. Early and precise diabetes diagnosis is vital for effective disease management and complication prevention. With the growing role of machine learning in healthcare, researchers have explored various predictive models to enhance diagnostic precision. This review critically examines decision trees, support vector machines, logistic regression, artificial neural networks (ANNs), and ensemble methods, evaluating their performance based on accuracy, precision, recall, and F1-score. Our analysis indicates that ensemble learning approaches, particularly random forests with gradient boosting, along with ANNs, consistently outperform traditional models, exhibiting superior predictive capabilities. These findings emphasize the potential of machine learning in improving diabetes detection by identifying complex patterns in patient data. Integrating advanced predictive algorithms into diabetes screening can enhance early detection and enable timely medical interventions. As machine learning models continue to evolve, their application in medical diagnostics holds significant promise for bettering diabetes detection and assisting healthcare professionals in selecting the most effective predictive models for clinical use.

Keywords: Artificial Intelligence, Diabetes Prediction, Deep Learning, Healthcare, Machine Learning, Tree- Ensemble

1 Introduction

Diabetes mellitus is a complex metabolic disorder that has emerged as one of the 21st century's most pressing health challenges [38]. 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. [29] 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 millennia. 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.

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

2. Type 2 Diabetes (T2D)

[14]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.

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 foetal health.

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.
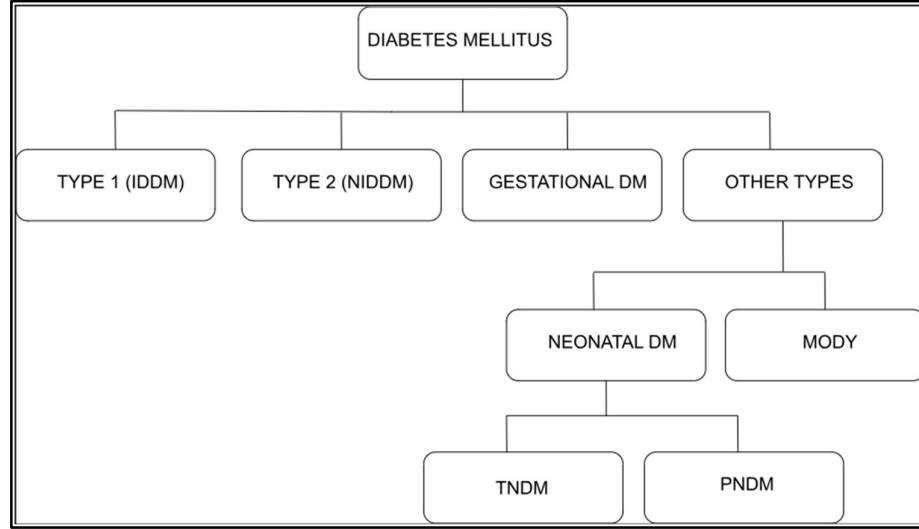
Fig. 1: Classification of Types of Diabetes

2 Machine Learning Methodologies for Diabetes Prediction

Machine learning techniques for diabetes prediction can be classified into several methodological categories, each with distinct approaches to pattern recognition and prediction. Tree-based methods including Decision Trees, Random Forest, and Gradient Boosting techniques create hierarchical decision structures that enable effective classification of diabetes risk factors. These methods are particularly valuable for their interpretability and ability to handle non-linear relationships in medical data.

Probabilistic and distance-based approaches offer complementary strengths in diabetes prediction. Naive Bayes and Logistic Regression models provide statistical frameworks for risk assessment, while K-Nearest Neighbors and Support Vector Machines excel at identifying complex boundaries between diabetic and non-diabetic cases. Neural network methodologies, ranging from traditional Artificial Neural Networks to more sophisticated architectures like Deep Belief Networks, Convolutional Neural Networks, and Recurrent Neural Networks, have demonstrated remarkable capability in capturing intricate patterns in physiological and demographic data. The field increasingly favors ensemble and hybrid methods that strategically combine multiple algorithms to leverage their respective strengths, thereby enhancing predictive performance.

Having established this classification of machine learning methodologies, we now turn to a comprehensive literature review examining how these techniques have been implemented for diabetes prediction across different research contexts.
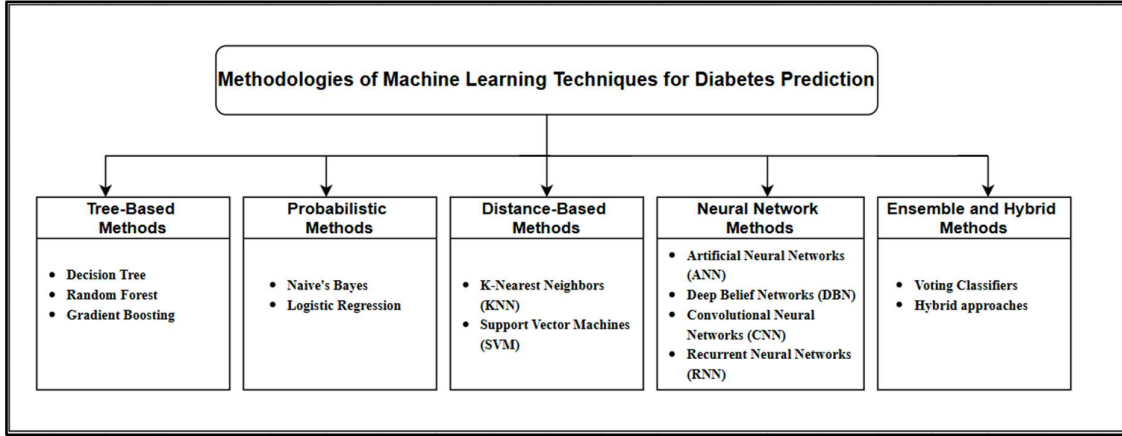
Fig. 2: Classification of Machine Learning Methodologies for Diabetes Prediction

3 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 analysing traditional methodologies to deep learning techniques developed by researchers and professionals to predict diabetes mellitus. The paper [43] explores the application of Support Vector Machine (SVM) modelling for predicting diabetes and pre-diabetes using data from the National Health and Nutrition Examination Survey (NHANES). It develops and validates SVM models for two classification schemes: diagnosed/undiagnosed diabetes vs. non-diabetes and undiagnosed diabetes/pre-diabetes vs. non-diabetes. The study identifies key predictive variables such as family history, BMI, age, waist circumference, and hypertension, achieving an AUC of 83.5% and 73.2% for the two schemes, respectively. A notable strength of this work is its use of simple, non-invasive clinical variables, making it applicable for large-scale screening. Additionally, the study develops a web-based tool, Diabetes Classifier, demonstrating real-world usability. However, limitations include the exclusion of laboratory tests, which may enhance predictive accuracy, and the challenge of generalizing results beyond the NHANES dataset.

The paper [22] explores the feasibility of forecasting Type 2 diabetes (T2D) risk using electronic medical record (EMR) data and machine learning techniques. It evaluates various classifiers, including Random Forest, Support Vector Machine, Naïve Bayes, and Decision Trees, achieving an AUC of over 0.8 for predictions made 180 to 365 days before diagnosis. The study excels in its use of real-world EMR data, making the model practical for clinical integration and early intervention. Additionally, the study employs feature selection techniques to enhance predictive accuracy. However, the result of the study includes a relatively low positive predictive value (~0.24), attributed to class imbalance, and the exclusion of significant risk factors like family history and genetic markers.

The paper [1] presents a boosting ensemble modeling approach for predicting diabetes mellitus using personal and clinical data. The study implements an AdaBoostM1 algorithm combined with a random committee classifier, evaluated on a dataset of 100

records with a prediction accuracy of 81% using 10-fold cross-validation in Weka. A key strength of the work is its integration into a cloud-based clinical decision support system (CDSS), enhancing real-world applicability. Additionally, the ensemble technique improves prediction performance over single classifiers. However, the study is limited by its small dataset, which may affect the generalizability of results.

The paper [2] investigates the prediction of diabetes mellitus using an ensemble machine learning approach combined with the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Using data from the Henry Ford Exercise Testing (FIT) project, which includes 32,555 patients with a five-year follow-up, the study applies multiple classifiers, including Decision Trees, Naïve Bayes, Logistic Regression, and Random Forests. The ensemble method, incorporating Naïve Bayes Tree, Random Forest, and Logistic Model Tree, achieved a high prediction accuracy (AUC = 0.92). A major strength of the study is the effective use of feature selection techniques to identify clinically relevant predictors. Additionally, the application of SMOTE significantly improved model performance in handling imbalanced data. However, limitations include reliance on a specific cohort (patients undergoing exercise stress testing) and potential challenges in generalizing findings to broader populations.

The paper [23] presents a deep learning approach to predict blood glucose levels for diabetic patients based on continuous glucose monitoring (CGM) data. Unlike traditional methods that train models individually for each patient, this study employs a generalized approach, using data from a subset of patients for training and testing on the remaining patients, enhancing its adaptability. The proposed deep neural network, structured hierarchically, outperforms shallow networks and Tikhonov regularization in accuracy, particularly in hypoglycemic and euglycemic ranges, as assessed by the PRED-EGA metric. A key strength of the work is its structured design, which incorporates domain knowledge for feature selection and diffusion geometry techniques to enhance learning. However, limitations include the use of a relatively small dataset (25 patients) and potential challenges in generalizing the model to broader populations, particularly in handling hyperglycemic spikes.

Three supervised machine learning algorithms—Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN)—were compared by the authors of [13] 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 [27] 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; 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, 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 compared to other methods including Naïve Bayes, RBF-NN, Decision Tree, Logistic Regression, Random Forest and SVM.

5

The authors of **[11]** analyzed various techniques for diabetes mellitus prediction, putting emphasis on early detection, 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, followed by SVM, KNN and Naive Bayes.

[40] 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; however, it faced challenges in model interpretability and dataset generalizability limitations. Linear Kernel SVM showed perfect performance, followed by Radial Kernel SVM and Polynomial Kernel SVM

In **[20]** the authors explored various ML algorithms for diabetes prediction, while acknowledging challenges in model interpretability and dataset limitations. The study compared multiple algorithms with varying performance metrics: KNN (K=10) achieved the highest accuracy, followed by SVM Naive Bayes and both Decision Tree and Random Forest showing similar results.

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 by achieving an accuracy of 98.07% with its CNN model, compared to DT model which gave an accuracy of 96.62%, ANN that yielded an accuracy of 90.34% and Naïve Bayes' that gave an accuracy of 76.33%.

The study **[44]** used Decision Tree, K-nearest neighbour (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 study **[32]** 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 **[25]** presented a comparative analysis of machine learning algorithms for diabetes prediction using the PIMA Indian Diabetes dataset. 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. The comparative results show Random Forest achieving the highest accuracy at 78.57%, marginally outperforming Linear SVM and K-NN, 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 **[42]** 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 to handle class imbalance, and feature importance evaluation.

In **[30]** the authors investigated early-stage diabetes mellitus risk prediction using machine learning algorithms on the PIMA dataset, focusing on a hybrid stacking model. Individual algorithm performances were evaluated, with accuracies ranging from 77.27%

6

(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 **[7]** used three supervised machine learning algorithms on the PIMA 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 the work **[21],** the authors assessed the performance of Logistic Regression (LR) and Random Forest (RF) for diabetes prediction using a dataset of 520 individuals from a hospital in Sylhet, Bangladesh. Employing 10-fold cross-validation. 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 **[39]** evaluated seven ML 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, and ROCAUC. Notably, Support Vector Machine (SVM) and Decision Tree (DT) achieved perfect scores across all metrics, while Logistic Regression reached 98.16% accuracy. Gradient Boosting and Random Forest both achieved 94.11% accuracy, K-Nearest Neighbour 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 **[5]** analysed the use of ML algorithms for predicting diabetes by focusing on improving accuracy using 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 **[18]** 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. Random Forest achieved the highest accuracy at 79%, followed by SVM, Logistic Regression, and KNN . On the Pima dataset, Random Forest again performed best with 80% accuracy, followed by SVM (77%), and Logistic Regression and KNN both at 73%.

The review paper **[37]** examined 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 study [24] 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. 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 [35] the authors explored 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 [31] explored ML 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. A total of 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. 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 [41] 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. 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%.

The researchers of [17] 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 highlighted notable accuracy results

from various research, including a highest reported accuracy of 99.04% using a 1-CNN and 97.5% accuracy with Random Forest.

In **[26]** 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. 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 **[33]** 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 **[19]** 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 neighbours (KNN)—to evaluate their effectiveness in diabetes diagnosis. 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%.

Various machine learning models have recently seen a rise in the prediction of diabetes in patients. One such study **[28]** 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 **[8]** explored diabetes prediction using Gaussian Naïve Bayes (GNB) and Artificial Neural Networks (ANN), together in an ensemble model. The research achieved high accuracy through thorough data preprocessing and PCA for dimensionality reduction. 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 is a major limitation.

This study by the authors of **[34]** 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 modelling (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 **[15]** presented a comprehensive overview of machine learning

9

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 paper compiled accuracy metrics from various studies, with the 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 [4] 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 [3] 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. 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 [16] provides a comprehensive overview of ML and DL techniques for diabetes mellitus detection and management. It analyses 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 [36] utilized LightGBM with SMOTE analysis to classify diabetic patients using the PIMA Indian Diabetes dataset. The research employed ANOVA for feature selection and SHAP 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 [10] 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 techniquesThe 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 [9] 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 [12] 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).

Technique	Accuracy	Recall	F1 − Score	Precision	Reference
Support Vector Machine (SVM)	0.65 − 0.8539	0.43 − 0.74	0.53 − 0.77	0.63 − 0.77	[5],[11],[12],[13],[15],[16],[17],[18],[19], [20],[24],[25],[26],[27],[30],[31],[32], [33],[37],[39],[40],[41],[43]
Decision Tree (DT)	0.6818 − 0.9662	0.57 − 0.9545	0.62 − 0.9472	0.60 − 0.9402	[2],[7],[11],[12],[17],[19],[20],[22],[24], [26],[27],[31],[37],[39],[44]
Random Forest (RF)	0.71 - 0.9903	0.51 - 0.98	0.59 - 0.98	0.60 - 0.98	[2],[7],[12],[15],[17],[18],[19],[20],[21], [22],[24],[25],[26],[27],[31],[32],[37], [39],[41],[42],[44]
Logistic Regression	0.73 - 0.9616	0.44 - 0.84	0.53 - 0.77	0.67 - 0.77	[2],[7],[13],[17],[18],[19],[21],[27],[30], [31],[32],[39],[41]
Naive Bayes (NB)	0.65 - 0.90	0.616 - 0.8629	0.645 - 0.8674	0.597 - 0.8644	[11],[12],[20],[30],[37],[39],[44]
K-Nearest Neighbors (KNN)	0.58 - 0.882	0.50 - 0.79	0.60 - 0.83	0.52 - 0.888	[11],[16],[17],[18],[19],[20],[26],[30], [31],[33],[41],[44]
AdaBoost	0.73 - 0.97	0.64 - 0.98	0.63 - 0.98	0.63 - 0.98	[1],[35],[37].
XGBoost	0.71 - 0.89	0.615 - 0.905	0.655 - 0.9	0.581 - 0.91	[4],[24],[31],[37],[35],[41]
Artificial Neural Network (ANN)	0.75 - 0.99	0.75 - 0.85	0.75 - 0.86	0.75 - 0.89	[6],[8],[13],[16]
Convolutional Neural Network (CNN)	0.90 - 0.99	0.83	0.86	0.88	[3],[6],[16],[17]
K-means Clustering	0.78 - 0.91	0.74	0.81	0.95	[12],[37]

Table 1: Comparative Performance Analysis of Machine Learning Techniques for Diabetes Prediction

4 Conclusion

The literature review conducts an examination of numerous methods by which AI-ML can be used for detection of diabetes. Machine learning has the capability to reform the diabetes prediction at an early stage with the help of various models and large amount of dataset. The paper highlights the development of use of different models from traditional to ensemble to deep learning techniques in order to detect diabetes. It was found that models like ANN and Ensemble Methods outperformed other models when applied to diabetes dataset. Moreover, other models performed well as well. However, the

development of hybrid models can potentially yield even better results for the models for prediction of diabetes.

5 Future Scope

This comparative analysis along with the advantages and disadvantages of various models may also help researchers to develop an accurate and more advanced tool that will prove to be a great help in the healthcare industry for detection of diabetes. Further works should focus on working on advance techniques as well as concentrate on diversity of the dataset to remove class imbalance in order to ensure better methods to detect diabetes. Research ahead must address the limitations and explore new approaches which will continue to forge our understanding and improve the detection of diabetes.

References

[1] Ali, R., Siddiqi, M. H., Idris, M., Kang, B. H., & Lee, S. (2014). Prediction of diabetes mellitus based on boosting ensemble modeling. In R. Hervás et al. (Eds.), UCAmI 2014: Lecture Notes in Computer Science (Vol. 8867, pp. 25–28). Springer. https://doi.org/10.1007/978-3-319-13102-3_6

[2] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLoS ONE, 12(7), e0179805. https://doi.org/10.1371/journal.pone.0179805

[3] 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

[4] 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

[5] Awoniran, O. M., Oyelami, M. O., Ikono, R. N., Famutimi, R. F., & Famutimi, T. I. (2022). A Machine Learning Technique for Detection of Diabetes Mellitus. Proceedings of the 5th International Conference on Information Technology for Education and Development: Changing the Narratives Through Building a Secure Society with Disruptive Technologies, ITED 2022. https://doi.org/10.1109/ITED56637.2022.10051439

[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] Buhari, B., & Adewara, A. A. (2022). On the analysis of some machine learning algorithms for the prediction of diabetes. International Journal of Advanced Computer Science and Applications, 13(8), 123-130. https://doi.org/10.14569/IJACSA.2022.0130816

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] Janaka Karalliedde, Luigi Gnudi, Diabetes mellitus, a complex and heterogeneous disease, and the role of insulin resistance as a determinant of diabetic kidney disease, Nephrology Dialysis Transplantation, Volume 31, Issue 2, February 2016, Pages 206–213, https://doi.org/10.1093/ndt/gfu405

[15] 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

[16] 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

[17] 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/ICSCSS57650.2023.1016924

[18] Kuriakose, S. M., Basa Pati, P., & Singh, T. (2022). Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record. 2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022. https://doi.org/10.1109/ICCCNT54827.2022.9984264

[19] 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

[20] 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

[21] Mangal, A., & Jain, V. (2022). Performance analysis of machine learning models for prediction of diabetes. Proceedings - 2022 2nd International Conference on Innovative Sustainable Computational Technologies, CISCT 2022. https://doi.org/10.1109/CISCT55310.2022.10046630

[22] Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. AMIA Annual Symposium Proceedings, 2012, 606–615

[23] Mhaskar, H. N., Pereverzyev, S. V., & van der Walt, M. D. (2017). A deep learning approach to diabetic blood glucose prediction. Frontiers in Applied Mathematics and Statistics, 3(14). https://doi.org/10.3389/fams.2017.00014

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] Tabish SA. Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century? Int J Health Sci (Qassim). 2007 Jul;1(2): V-VIII. PMID: 21475425; PMCID: PMC3068646.

[39] Thabit, Q. Q., Fahad, T. O., & Dawood, A. I. (2022). Detecting Diabetes Using Machine Learning Algorithms. 2022 Iraqi International Conference on Communication and Information Technologies, IICCIT 2022, 131–136. https://doi.org/10.1109/IICCIT55816.2022.10010408

[40] 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

[41] 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.

[42] 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

[43] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(16). https://doi.org/10.1186/1472-6947-10-16

[44] 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/IC4ME253898.2021.97685