A Comprehensive Survey on Machine Learning and Deep Learning Techniques for Diabetes Prediction

[1]Vandana C, [2]Hema B, [3]Monika S, [4]Sinchana T, [5]Dr. Manjunath T K

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor & HOD
[1]Department of AI&DS,
K. S. School of Engineering and Management, Bengaluru, India
[1]vandanac1709@gmail.com, [2]19sc21hema@gmail.com, [3]gowdamonika026@gmail.com ,
[4]tsinchana04@gmail.com, [5]manjunathtk@kssem.edu.in

***Abstract*— Diabetes is among the most common chronic conditions globally, having a significant influence on global healthcare systems. Early and precise prediction of diabetes can help in early intervention and management. This paper is a detailed survey of current methods employed for diabetes prediction, with emphasis on a broad spectrum of machine learning (ML) and deep learning (DL) methods. The literature reviewed contains recent research studies that used algorithms like Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayes, XGBoost, Artificial Neural Networks, Convolutional Neural Networks, ensemble learning models, and optimization-based methods. Metrics such as accuracy, AUC, RMSE, and MAE were taken into account for comparative assessment. The research shows that ensemble and hybrid models tend to perform better than single classifiers, with some models having a prediction accuracy of up to 100%. The paper also identifies new trends, including the use of sensor data, reinforcement learning, and explainable AI for predicting diabetes. In summary, this survey gives a systematic reference for researchers and practitioners seeking to develop precise, trustworthy, and scalable solutions for diabetes monitoring and diagnosis.**

***Index Terms*— Diabetes Prediction, Machine Learning, Deep Learning, Artificial Intelligence, Medical Diagnosis, Healthcare Analytics, Classification Algorithms, Disease Prediction, Ensemble Models, Neural Networks.**

I. INTRODUCTION

Diabetes is a chronic disease that disables the body to control blood glucose levels, frequently resulting in severe complications like cardiovascular disease, kidney failure, vision loss, and nerve damage. With an important and increasing number of affected people worldwide, early detection and control of diabetes have become ever more crucial to alleviate the load on healthcare systems. As per the World Health Organization (WHO), the worldwide prevalence of diabetes is increasing at a very fast rate, especially in developing nations because of lifestyle changes, eating patterns, and genetic susceptibility. Early diagnosis can decrease the risk of complications and enhance the quality of life for people suffering from the disease.

Historically, diabetes diagnosis has been based on blood tests and clinical assessments, including the fasting blood sugar test, HbA1c test, and oral glucose tolerance test. Although useful, these tests can be time-consuming, costly, and inaccessible to people in remote or underserved communities. Consequently, there is a growing need for smart, automated systems that can predict diabetes from patient health data.

In this regard, Machine Learning (ML) and Deep Learning (DL) are becoming increasingly robust tools for processing large and complicated health data sets. ML is a branch of artificial intelligence that allows systems to learn from data and make decisions or predictions without being explicitly coded. It includes a variety of algorithms, such as Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forests (RF), and ensembling methods.

Deep Learning, a more sophisticated branch of ML, relies on artificial neural networks that replicate the structure and operation of the human brain. DL models—like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks—are especially suited to dealing with high-dimensional data and extracting hidden patterns from it. These models have been highly successful in medical image analysis, time-series prediction, and electronic health record processing.

This paper provides an extensive review of multiple machine learning and deep learning methods for predicting diabetes. The review centers on algorithms employed, datasets utilized, and the measures of performance presented in different studies. Through the evaluation of the merits and demerits of each method, the paper seeks to establish potential areas for research and facilitate the creation of smart diagnostic tools for diabetes.

II. BACKGROUND & DEFINITIONS

Diabetes: A long-term metabolic disease associated with hyperglycemia, resulting in long-term complications unless controlled.

Machine Learning (ML): A branch of artificial intelligence that allows systems to learn from experience and make better decisions without being programmed.

Deep Learning (DL): A type of ML that employs artificial neural networks with deep layers to handle complex data patterns.

Ensemble Learning: A method that uses an ensemble of different ML models to enhance prediction accuracy and stability.

Data Preprocessing: The operation of cleaning and converting raw data into a format appropriate for ML algorithms.

Feature Selection: The operation of choosing a subset of useful features from the original data to enhance model performance and simplify complexity.

Class Imbalance: A scenario in which the number of instances in various classes of a classification problem is significantly disparate.

Pima Indians Diabetes Database (PIDD): One of the most popular datasets used for diabetes prediction, comprising medical records of Pima Indian females.

Electronic Health Records (EHRs): Electronic copies of patients' paper charts that have medical and treatment history.

IoT (Internet of Things): An internet of physical objects embedded with sensors, software, and other technologies to enable connection and exchange of data.

Accuracy: The ratio of accurately predicted instances out of the total instances.

Precision: The ratio of accurately predicted positive instances out of the total predicted positive instances.

Recall (Sensitivity): The ratio of accurately predicted positive instances out of the total actual positive instances.

F1-score: Harmonic mean of recall and precision.

AUC (Area Under the ROC Curve): A metric for how well the model can discriminate between classes.

ROC (Receiver Operating Characteristic) Curve: A visual plot of the model's performance across various classification thresholds.

RMSE (Root Mean Squared Error): A metric for differences between predicted and true values in regression tasks.

III. Related Work

K. V. Daliya & T. K. Ramesh [1] used an optimized ensemble model based on LightGBM and KNN for the prediction of diabetes progression with an accuracy of 75% and improved scalability for cloud-based systems. The dataset of 442 patient records, containing attributes like age, sex, BMI, blood pressure, and six blood serum measurements (LDL, HDL, TC, TG, Glucose, and LTG levels). The outcome variable, diabetic disease development, was transformed into a two-class classification problem with low and high-risk groups. Their data are publicly available and originally employed in research on Least Angle Regression.

G. Parimala et al. [2] studied various ML algorithms like SVM, Decision Trees, Random Forest and KNN for classification of diabetes, underlining the significance of data preprocessing in improving prediction accuracy, achieving an accuracy of 98% through Random Forest.

Khaled Alnowaiser [3] presented a Tri-Ensemble Model with KNN-imputed features to manage missing values and attained a remarkable accuracy of 97.49%. This research utilized the Pima Indians Diabetes Database on Kaggle, which had 768 patient records, with clinical features like Age, Diabetes, Insulin, Mass, Pedigree, Pregnant, Triceps, blood pressure, and glucose level.

Anuj Mangal & Vinod Jain [4] conducted a comparative analysis of ML models for the prediction of diabetes and identified that the Random Forest model provided the highest accuracy of 99.03%. This dataset of 520 patient records, which were retrieved from Sylhet Diabetes Hospital, Bangladesh. It contained 17 diabetic risk factors like age, obesity, and weakness, with 320 diabetic and 200 non-diabetic cases.

G. Ravi Kumar et al. [5] deployed a web application with ML classifiers like Naïve Bayes, Logistic Regression, KNN, SVM, Decision Trees, and Random Forest, enhancing access to diabetes prediction tools and identified that Random Forest and SVM provided highest accuracy of 80%. They compiled data from **two independent sources**, integrating diverse global statistics on diabetes and health conditions.

Arwatki Chen Lyngdoh et al. [6] analysed diabetes prediction by applying machine learning and deep learning algorithms. Pima Indians Diabetes Database (PIDD) by the National Institute of Diabetes and Digestive and Kidney Diseases is popular for such a study that comprises 768 cases with 8 diagnostic features like glucose level, BMI, and insulin. Several models such as KNN, SVM, Random Forest, and Deep Learning-based hybrid models have been used to enhance the accuracy of predictions. Feature selection methods and K-Fold Cross Validation are used to improve model performance and generalization.

Ifra Shaheen et al. [7] introduced ensemble learning approaches, Hi-Le and HiTCLe, for diabetes prediction with deep learning and explainable AI. The Diabetes Prediction Dataset (DPD) was employed, obtained from different Electronic Health Records (EHRs) and publicly available on Kaggle, consisting of 100,000 instances with 9 features: age, gender, blood sugar, BMI, smoking history, heart disease, and hypertension. Because of class imbalance (91.5% non-diabetic, 8.5% diabetic), ProWSyn was used for balancing. Hi-Le employs Highway and LeNet models, whereas HiTCLe combines TCN, LeNet, and Highway for prediction. Preprocessing of data included encoding categorical features and imbalances to enhance classification performance.

Ahmed Ali Linkon et al. [8] performed a thorough research on feature transformation and machine learning models for diagnosing early diabetes. They applied a publicly available dataset from Sylhet Diabetes Hospital, which has 16 features and 1 target variable, gathered through direct patient questionnaires. The research included data preprocessing, such as missing value handling, Label Encoding of categorical features, and correlation analysis of features with Pearson's correlation coefficient. Feature importance was calculated using Mutual Information (MI) scores, and feature scaling methods like Standard Scaler, Min-Max Scaling, and L2 Normalization were employed to improve model performance. Comparison study was conducted in order to measure the effect of various scaling procedures on machine learning model accuracy and convergence, underlining the relevance of feature preprocessing in predictive models.

Thavavel Vaiyapuri et al. [9] suggested an IoT-based EDCD-DLDR model for early diabetes detection, including data acquisition, normalization, feature selection, diabetes detection, and hyperparameter tuning. The system uses Arduino-based IoT sensors to track patient vitals such as blood pressure, temperature, oxygen saturation, and glucose. RFID authentication provides

secure access, with real-time monitoring and automated medical record update. Z-score normalization normalizes the data, enhancing comparability and machine learning performance with effective management of outliers and varying scales.

Usama Ahmed et al. [10] suggested a Fused Model for Diabetes Prediction (FMDP) with two stages: Training and Testing. The Training Layer includes data collection, preprocessing, classification (SVMs and ANNs), performance analysis, and machine-learning fusion. The dataset is collected from the UCI Machine Learning Repository and cleaned, normalized, and divided into training and test sets. Fuzzy rules combine SVM and ANN outputs for making final predictions. The Testing Layer takes the trained model from the cloud and makes predictions of diabetes cases in a highly accurate manner using measures such as precision, sensitivity, specificity, and F1 score.

Santosh Kumar Sharma et al. [11] suggested a cloud-based diabetes diagnosis model combining feature selection and classification methods through Principal Component Analysis (PCA) and Extreme Learning Machine (ELM). The PIMA Indian Diabetes Dataset was used to validate the model, showing enhanced accuracy in classification.

Aditi Site et al. [12] proposed a multi-sensor data-driven machine learning model for diabetes prediction. The D1NAMO dataset, containing ECG, breathing, and glucose sensor data, was used in the research to investigate glycemic events under non-clinical conditions. Multimodal data significantly enhanced predictive accuracy.

Nor Nisha Nadhira Nazirun et al. [13] also carried out a systematic review of prediction models for Type 2 diabetes progression. The research compared 41 datasets, ranging from Electronic Health Records (EHRs) to randomized clinical trials and publicly available databases. The review noted the PIMA dataset to be one of the most widely used datasets in diabetes prediction but also noted its limitations in real-time usage.

Mana Saleh Al Reshan et al. [14] introduced an ensemble deep learning-based Clinical Decision Support System (CDSS) for predication of diabetes. The experiment was conducted with three datasets namely PIMA-IDD-I, DDFH-G, and IDPD-I, which correspond to both binary and multi-class classification problems. The model that was proposed presented high accuracy when it came to classifying individuals as diabetic, prediabetic, and non-diabetic.

Rishi Jain et al. [15] examined gender and age heterogeneity in diabetes prediction through a multi-model ensemble learning method. The dataset was collected from a diabetologist and maintained by Dr. Reddy's Lab, consisting of 13 independent variables and one dependent variable (outcome). The independent variables include gender, age, HbA1c, BMI, fasting glucose (GTT0), LDL, HDL, non-HDL, triglycerides (TG), uric acid, systolic BP, and diastolic BP. The outcome variable is categorical, classifying individuals as non-diabetic, pre-diabetic, or diabetic, based on Glucose Tolerance Test (GTT) values. The dataset comprises 311 individuals aged 30-40 and 297 millennials, with a gender distribution of 461 females and 211 males, allowing for demographic-based diabetes analysis.

Radwa Marzouk et al. [16] suggested analytical predictive models based on Synthetic dataset and the PIMA Diabetes Dataset (PIDD). Their research highlighted the significance of BMI as a key factor in predicting diabetes risk and used correlation analysis to determine the important attributes.

Mohammad Zubair Khan et al. [17] proposed a bio-inspired Particle Swarm Optimization (PSO) strategy to improve diabetes prediction using neural networks. The PIDD dataset was utilized for training and validation of the developed model, illustrating improvements in the accuracy of classification.

Shamim Ahmed et al. [18] carried out a comparative study of LIME and SHAP interpreters for explainable machine learning-based diabetes prediction using the BRFSS 2015 dataset. They examined feature importance and explainability to enhance decision making in healthcare contexts.

Nada Y. Philip et al. [19] created a sophisticated data analytics platform for Type 2 Diabetes (T2D) data analysis, combining exploratory, predictive, and visual analytics. The research used patient datasets from Croydon/Prowellness and Diamond databases to construct risk models for complications of diabetes.

Hakim El Massari et al. [20] used ontology-based machine learning methods for predicting diabetes from the PIDD dataset. Their method used structured knowledge representation to enhance interpretability and prediction accuracy.

Liya Jia et al. [21] used the Pima Indian Diabetes Dataset (PIDD) and two additional diabetes datasets, RSMH and Tabriz. The authors included a correlation coefficient matrix heatmap for the PIDD. However, the specific size and feature details of the RSMH and Tabriz datasets are not provided.

Pierluigi Francesco De Paola et al. [22] take a different approach and do not use a traditional dataset. Instead, they propose a mathematical model that incorporates interleukin-6, glucose, and insulin levels to simulate the long-term effects of physical activity on diabetes progression.

Hamdi A. Al-Jamimi [23] utilized the Sylhet diabetes dataset. While the exact size of this dataset is not specified, it includes demographic (age, gender), lifestyle (smoking, hypertension), and medical indicators (chest pain, family history of heart disease, electrocardiographic pattern) information. The dataset contains 20 features, which the study refines to 12 relevant features through feature engineering.

Nur Ghaniaviyanto Ramadhan et al. [24] conducted a systematic literature review, examining various studies and datasets related to chronic disease prediction. The datasets reviewed contain data from medical examinations, laboratory results, doctor consultations, and general check-ups. Given the review nature of this paper, the dataset sizes vary.

Kok-Lim Alvin Yau et al. [25] present a survey of reinforcement learning (RL) models and algorithms for diabetes management. As a review, it discusses data types used in RL for diabetes, including demographic data (age, gender) and physical examination data (BMI, temperature, blood pressure), medical history, and lab test results (glucose levels, insulin dosage).

Muhammad Nauman et al. [26] used the CDC's Behavioral Risk Factor Surveillance System (BRFSS) dataset. The exact size of the dataset is not specified in the text. The features include variables related to diabetes, high blood pressure, high cholesterol, cholesterol checks, BMI, smoking status, stroke, heart disease, physical activity, fruit and vegetable consumption, alcohol consumption, and healthcare access.

Virginie Felizardo et al. [27] present a review of hypoglycemia prediction models. As a review, they discuss various datasets used in the studies they analyzed. The key data types in these datasets include blood glucose levels and information related to hypoglycemia events.

Serena Zanelli et al. [28] used a dataset from the University Hospital of Nice (Cohort Register No. BS-004). The dataset is described as small and contains raw Photoplethysmography (PPG) waves.

The paper by Hafeez Ur Rehman Siddiqui et al. [29] focuses on improving automated Peripheral Sensory Neuropathy (PSN) assessment in type 2 diabetes patients through plantar lesion recognition and probe avoidance techniques. The study evaluates different feature sets—Local Binary Pattern (LBP), Mel Frequency Cepstral Coefficients (MFCC), and Scale-Invariant Feature Transform (SIFT)—across various classifiers, including Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Random Forest (RF), Naïve Bayes (NB), and XGBoost, to classify lesions on the plantar surface.

Mohammad Z. Atwany et al. [30] present a survey on deep learning techniques for Diabetic Retinopathy (DR) classification. The paper reviews state-of-the-art deep learning methods in supervised, self-supervised, and Vision Transformer setups for retinal fundus image classification and detection. It also discusses available retinal fundus datasets used for DR detection, classification, and segmentation, and addresses research gaps and challenges in the field.

Avishek Anishkar Ram et al. [31] propose a guided neural network approach to predict early readmission of diabetic patients. The study applies a new method to Artificial Neural Networks (ANNs) to guide gradient descent optimizers by distinguishing between consistent and inconsistent data in each batch. The results demonstrate improved classification accuracies and better error convergence compared to standard ANNs.

Mayank Jichkar et al. [32] utilized a Kaggle diabetes dataset consisting of 768 adult participants with 8 input features. They performed exploratory data analysis and applied preprocessing steps such as SMOTE for class balancing and StandardScaler for feature normalization.

Niels F. Cleymans et al. [33] analyze Random Forest's predictive capability for Type 1 Diabetes progression. The study constructs random forest survival models to predict the time to clinical onset of T1D using genetic and immune biomarkers. The results show that random forest models outperform traditional Cox regression methods.

Mr. Khaja Mannanuddin et al. [34] worked on the Messidor dataset with 1150 samples and 19 features focused on diabetic retinopathy detection using image-based attributes like microaneurysms and exudates.

Sadia Afrin Shampa et al. [35] present a study on machine learning-based diabetes prediction from a cross-country perspective. The research examines diabetes data from Bangladesh, India, and Germany using various machine learning models. The findings indicate that boosting algorithms like AdaBoost, CatBoost, Gradient Boost, and XGBoost perform well, especially with the Bangladesh dataset.

Gresha Bhatia et al. [36] leveraged data from the Centers for Disease Control and Prevention (CDC) to predict infectious diseases, integrating climatic parameters like temperature, precipitation, and humidity. Their preprocessing involved null value handling, forward-filling, standardization, and weekly aggregation

Navaneeth Bhaskar et al. [37] propose an automated medical system for detecting type 2 diabetes from exhaled human breath using a deep hybrid architecture. The system analyzes the concentration of acetone in exhaled breath with a sensing module and employs Convolutional Neural Networks (CNNs) for prediction.

Praveen Tumuluru et al. [38] also sourced data from Kaggle, derived from health questionnaires annotated by specialists. Their dataset included lifestyle-based parameters such as food consumption habits, physical activity, sleep duration, and genetic predisposition. Their preprocessing pipeline addressed real-world data challenges like non-linearity, missing values, and outliers, utilizing imputation, irrelevant feature removal, and oversampling techniques.

Farrukh Aslam Khan et al. [39] presents an extensive survey of data mining approaches applied to diabetes prediction and detection. The authors explore a wide array of machine learning models, algorithms, and frameworks that have been used in the

medical field to support clinical decision-making. It also discusses the significance of early diagnosis, common challenges, and future research directions in diabetes-related data mining.

Asif Hassan Syed and Tabrej Khan [40] conducted a retrospective cross-sectional study using a dataset of 4896 participants from different regions in Saudi Arabia. The study identified 10 diabetes risk factors including demographic details, BMI, waist size, diet, exercise, and family history. The classification models used categorical variables and aimed to distinguish between high and low-risk diabetic individuals. The dataset is available on IEEE DataPort.

Evgenii A. Pustozervov et al. [41] and collaborators used data from a clinical trial in Russia involving 235 pregnant women, including those with gestational diabetes mellitus (GDM). Using CGM (Continuous Glucose Monitoring) data and meal logs, they predicted postprandial blood glucose levels. The study employed Python and scikit-learn for machine learning implementation and focused on the impact of glycemic goals on GDM management.

Shirina Samreen [42] proposed a novel machine learning pipeline using feature selection with Crow Search Optimization and a stacking ensemble model. The dataset used was sourced from the UCI repository and contained 520 instances with 16 attributes. It focused on early diagnosis of Type 2 diabetes using binary symptoms and categorical inputs, transformed through one-hot encoding. Performance was evaluated using accuracy, F-measure, and AUC.

Bum Ju Lee et al. [43] analyzed data from 4870 Korean individuals as part of the KHGES database. Anthropometric data such as height, weight, and body circumferences from various sites were collected and used to predict fasting plasma glucose status. The study emphasized the effectiveness of body measurement ratios in non-invasive diabetes prediction using regression analysis and statistical modeling.

Sajida Perveen et al. [44] used data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), comprising 667,907 records. After filtering for complete information on metabolic syndrome (MetS) features, a final dataset of 4,403 records was used. They applied machine learning techniques to predict diabetes mellitus based on clinical variables like BMI, FBS, HDL, triglycerides, blood pressure, and sex.

Md. Shafiqul Islam et al. [45] utilized the San Antonio Heart Study (SAHS) dataset, which included 1,368 randomly selected subjects. The study focused on predicting the future progression of Type 2 Diabetes using machine learning. The data were imbalanced (only 11% diabetic cases), so they created seven balanced groups to enhance classification performance.

Qian Wang et al. [46] worked on the Pima Indians Diabetes Dataset (PIDD) from the UCI repository, which contains 768 samples, of which 268 are diabetic. Due to missing values and class imbalance, they proposed the DMP_MI algorithm, designed to handle both challenges effectively.

Nikos Fazakis et al. [47] developed long-term risk prediction tools for Type 2 Diabetes using machine learning and validated them using the ELSA dataset. Their approach included socio-demographic, clinical, and lifestyle factors and was compared with traditional diabetes risk scores like FINDRISC and Leicester.

Anastasios Alexiadis et al. [48] created a next-day hypoglycemia prediction model using data from the for Diabetes mobile app. Their dataset included 317549 glucose records and 33287 blood pressure records from 998 participants, and the study aimed to enhance self-management in diabetes patients through timely predictions.

Norma Latif Fitriyani et al. [49] used four datasets from Dr. John Schorling, Golino et al., and Dr. P. Soundarapandian to develop a Disease Prediction Model (DPM) for type 2 diabetes and hypertension. These included data on 403 subjects for diabetes, 175 males for hypertension, 224 females for prehypertension, and 400 subjects for CKD, with varying features related to risk factors and blood pressure.

Mehrbakhsh Nilashi et al. [50] employed the Pima Indians Diabetes Dataset (768 instances, 8 features) from the UCI Repository to enhance diabetes classification using advanced ensemble learning techniques.

Tuan Minh Le et al [51] used a dataset from Sylhet Diabetes Hospital, Bangladesh (520 instances, 16 attributes), to apply a wrapper-based feature selection method for early diabetes prediction.

Giulia Noaro et al. [52] simulated data from a virtual cohort of 100 adults with type 1 diabetes over 1200 and 180 days, using features like carbohydrate ratio and meal timing to develop a personalized insulin bolus calculator via double deep Q-learning.

Md. Kamrul Hasan et al. [53] and H. Roopa and T. Asha [55] used the PIMA Indians Diabetes Dataset (768 instances, 8 attributes) to train machine learning models for diabetes classification, with Roopa et al. [55] focusing on PCA-based linear modeling.

Saul Langarica et al. [54] gathered real-time ambulatory data from 20 participants in Chile over seven days, including glucose levels, heart rate, insulin administration, and dietary habits, to support T1DM monitoring and prediction

IV. CLASSIFICATION OF EXISTING APPROACHES

Traditional Machine Learning Methods

Older machine learning models like Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees have been used very widely because of their interpretability and simplicity of use. Researchers such as Ifra Shaheen et al. [7], Ahmed Ali Linkon et al. [8], and Sadia Afrin Shampa et al. [35] employed these models with varying degrees of success. While easier to use than newer models, classical algorithms are still in common use for baseline tests.

Ensemble Learning Methods

A prevalent trend among recent studies is the adoption of ensemble models including Random Forest, XGBoost, LightGBM, AdaBoost, and Soft Voting Classifiers. These models improve performance by combining multiple learners. G. Parimala et al. [2], Anuj Mangal and Vinod Jain [4], and Hafeez Ur Rehman Siddiqui et al. [29] obtained accuracy rates of more than 98% using such methods. Ensemble models were also proved useful while being combined with feature selection and sampling methods [44], [53].

Deep Learning Models

Deep learning-based models, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Extreme Learning Machines (ELM), have shown superior performance in learning complicated patterns in high-dimensional medical data. Mana Saleh Al Reshan et al. [14] and Mohammad Zubair Khan et al. [17] obtained accuracy up to 99.5%, which proves them to be apt for accurate diagnosis.

Hybrid and Stacking Models

Some researchers utilized hybrid models or stacking ensembles of the predictions of several algorithms to improve robustness. For instance, Khaled Alnowaiser [3] suggested a tri-ensemble model with KNN-imputed features with 97.49% accuracy. Rishi Jain et al. [15] utilized a soft voting classifier to attain 99.4%, whereas Shirina Samreen [42] combined Crow Search with stacking ensembles to achieve 98.46% accuracy. Figure 1 depicts the process of the tri-ensemble technique.

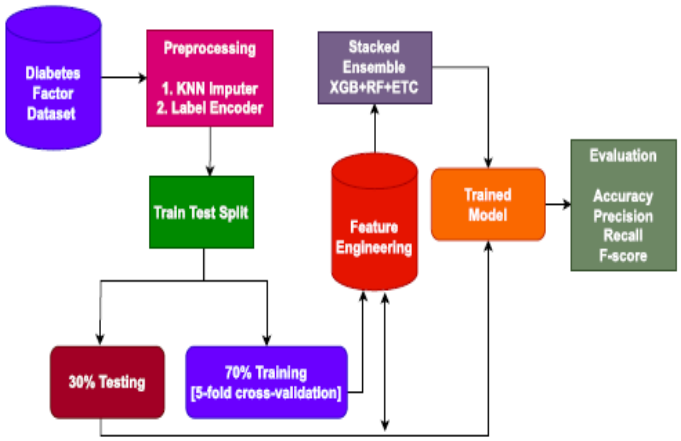

Fig.1 Proposed methodology for diabetes detection [3]

Optimization-based Methods

Optimization algorithms like Bayesian Optimization and Particle Swarm Optimization have been investigated for the hyperparameter tuning for enhanced generalization of the model. Hamdi A. Al-Jamimi et al. [23] obtained 100% accuracy with Bayesian Optimization, and Mohammad Zubair Khan et al. [17] used PSO for neural network-based prediction. Figure 2 depicts the architecture of Particle Swarm Optimization Neural-Network Diabetes Prediction.

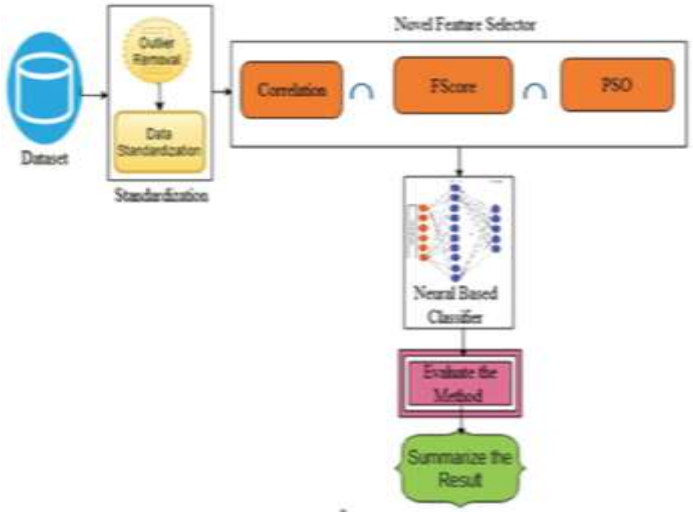

Fig. 2 The architecture of PSO-NNDP [17]

Feature Engineering and Sampling Techniques

Most of the studies handled data imbalance and feature redundancy through methods like Principal Component Analysis (PCA), SMOTE, ADASYN, and LIME. These preprocessing operations greatly enhanced model performance in research by H. Roopa & T. Asha [55], Sajida Perveen et al. [44], and Qian Wang et al. [46].

Sensor-based and Real-Time Systems

Recent advancements also involve multisensor-based systems in which real-time sensor data from sensors such as glucose, ECG, and accelerometers are combined with ML models. Aditi Site et al. [12] integrated such sensor data with XGBoost with 98.2% accuracy.

Reinforcement Learning and Interpretability

New methods like Double Deep Q-Learning (DDQN) [52] are designed to tailor diabetes management systems to learn best treatment policies. Some research also targets interpretability and clinical trust, employing techniques like LIME [18] and ontology classifiers [20] to explain model decisions.

V. RESULTS AND ANALYSIS

The models were tested against accuracy, precision, recall, and F1-score. The accuracy of various ML algorithms used for predicting diabetes is summarized below in a table:

Table 1 Comparison of Accuracy

Author	Algorithm Used	Accuracy (%)
V. K. Daliya, T. K. Ramesh [1]	LightGBM with KNN Imputer	75.0
G. Parimala et al. [2]	Random Forest, SVM, Decision Trees, KNN	98.0 (Random Forest)
Khaled Alnowaiser [3]	Tri-Ensemble Model with KNN-imputed features	97.49
Anuj Mangal & Vinod Jain [4]	Random Forest	99.03
G. Ravi Kumar et al. [5]	Naïve Bayes, Logistic Regression, KNN, SVM, Decision Trees, Random Forest	80.0 (SVM)
Arwatki Chen Lyngdoh et al.[6]	K-Nearest Neighbors (KNN) Naïve Bayes Decision Tree Classifier Random Forest Support Vector Machine	76 (KNN)
Ifra Shaheen et al.[7]	Hi-Le and HiTCLe model	94
Ahmed Ali Linkon et al. [8]	Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, KNN, Random Forest, Gradient Boosting, AdaBoost, Extra Trees, Bagging Classifier, Light Gradient Boosting Machine (LGBM), XGBoost	82.91% (LBGM)
Thavavel Vaiyapuri et al. [9]	EDCD-DLDR technique, ARO-FS model	97.14
Usama Ahmed et al. [10]	Fused ML Decision	94.87
Santosh Kumar Sharma et al. [11]	Extreme Learning Machine	90.57
Aditi Site et al. [12]	Multisensor (glucose, ECG, ACC sensors) with the XGBoost	98.2
Mana Saleh Al Reshan et al. [14]	Stacked ANN, Stacked LSTM, Stacked CNN	99.5(Stacked ANN)
Rishi Jain et al. [15]	Soft Voting Classifier	99.4
Radwa Marzouk et al. [16]	Logistic Regression, Naive Bayes, Support Vector Clustering, Decision Tree, KNN, Random Forest, Gradient Boosting, ANN	81.69(ANN)
Mohammad Zubair Khan et al. [17]	PSO NNDP	99.5
Shamim Ahmed et al. [18]	Logistic Regression(LIME and SHARP)	86
Nada Y. Philip et al. [19]	SVM	65.05
Hakim El Massari et al. [20]	Ontology Classifier	79.7
Liyan Jia Et Al. [21]	Probabilistic Ensemble: Lmegnb, K-Means SMOTE, Xgboost, RF, WKNN	94.53%
Hamdi A. Al-Jamimi Et Al [23]	Bayesian Optimization	100%
Muhammad Nauman et al. [26]	Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, Ann, Random Forest, Gradient Boosted Trees, Xgboost	75%(Lr)
Serena Zanelli et al. [28]	Lightweight CNN	AUC 75.5
Hafeez Ur Rehman Siddiqui et al. [29]	SVM, MLP, RF, NB, XGBoost	100% (NB with LBP)
Avishek Anishkar Ram et al. [31]	RF, SVM, Guided ANN	89.16 (Guided ANN)
Mayank Jichkar et al. [32]	Logistic Regression, KNN, Naïve Bayes, Decision Tree	81% (LR & KNN)
Niels F. Cleymans et al. [33]	Random Forest	Outperformed the available Cox regression

		models
Mr. Khaja Mannanuddin et al. [34]	Decision Tree, SVM, Naive Bayes, Random Forest	77.48% (SVM)
Sadia Afrin Shampa et al. [35]	Decision Tree, SVM, Naive Bayes, Random Forest, ANN, KNN, AdaBoost, CatBoost, Gradient Boosting, XGBoost.	100% (AdaBoost, CatBoost, Gradient Boost, and XGBoost on Bangladesh dataset)
Gresha Bhatia et al. [36]	Deep RNN	Mean Absolute Error = 13.05
Navaneeth Bhaskar et al. [37]	CORNN hybrid model	98.02%
Praveen Tumuluru et al. [38]	Logistic Regression, KNN, Decision Tree, SVM, Random Forest.	77.73 (SVM)
Asif Hassan Syed and Tabrej Khan [40]	Descision Forest	82%
Evgenii A. Pustozervov et al. [41]	Boosted Decision Trees	R=0.644, MAE= 0.377 mmol/L*h
Shirina Samreen [42]	Crow Search + Stacking Ensemble	98.46%
Bum Ju Lee et al. [43]	Logistic Regression, Naïve Bayes	AUC= 74.1(LR)
Sajida Perveen et al. [44]	Naïve Bayes, Decision Tree	ROC= 88.3% Naïve Bayes with K-medoids under-sampling technique
Md. Shafiqul Islam et al. [45]	Polynomial, SVM, Ensemble Learning	95.94% (Ensemble Model)
Qian Wang et al. [46]	DMP_ML model(Naïve Bayes, ADASYN, Random Forest)	87.10%
Nikos Fazakis et al. [47]	Logistic Regression, Naïve Bayes, Decision Tree, Random Forest	AUC=88.4% (Ensemble Weighted Voting LR, RFs ML model)
Anastasios Alexiadis et al. [48]	Random Forests, Support Vector Machines, Adaptive Boosting and Feed-Forward Artificial Neural Networks	81.4% (Random Forest)
Norma Latif Fitriyani et al. [49]	Disease Prediction Model(iForest, SMOTE-Tomek, Ensemble Learning)	100%
Mehrbakhsh Nilashi et al. [50]	PCA-SOM-NN	92.28%
Tuan Minh Le et al [51]	SVM, DT, RFC, NBC, LR, KNN, GWO-MLP, APGWO-MLP	97% (APGWO-MLP)
Giulia Noaro et al. [52]	Double Deep Q-Learning (DDQN) algorithm	The method improved the time in target range from 68.35% to 70.08% and significantly reduced the time in hypoglycemia
Md. Kamrul Hasan et al. [53]	KNN, DT, AdaBoost, Random Forest, Naïve Bayes, XGBoost	AUC= 95%(Ensemble of AdaBoost & XGBoost)
Saul Langarica et al. [54]	Long Short-Term Memory, Encoder-Decoder, Bidirectional Encoder-Decoder, Encoder-Decoder With Double Attention	RMSE ≈ 22.188 mg/dL (ENC-DEC DATTN)
H. Roopa & T.Asha [55]	PCA-LRM	82.1%

From the comparative study of research papers on diabetes prediction with machine learning and deep learning algorithms, the following conclusions were made:

- **Highest Accuracy Achieved:** Hamdi A. Al-Jamimi et al. [23], Sadia Afrin Shampa et al. [35], and Hafeez Ur Rehman Siddiqui et al. [29] achieved 100% accuracy, using ensemble algorithms like Bayesian Optimization, XGBoost, AdaBoost, and Naïve Bayes with LBP.
- **Popular and Robust Algorithms:** Random Forest and XGBoost were the top-used and produced uniform performance on datasets, varying between 80% and 99%.
- **Deep Learning Models** such as Stacked ANN, Guided ANN, and CNNs registered up to 99.5% accuracy, reflecting their capacity to process sophisticated data patterns.
- **Hybrid and Stacked Models** such as the Tri-Ensemble Model [3], Soft Voting Classifier [15], and Crow Search with Stacking Ensemble [42] attained high accuracy (over 97%), demonstrating the efficacy of model fusion.
- **Optimization-Based Methods** enhanced model tuning and generalization, with PSO and Bayesian Optimization achieving near-perfect performance [17], [23].
- **Reinforcement Learning and Sensor-Based Methods** demonstrated promise for real-time monitoring and adaptive decision-making, though additional clinical testing is needed.

VI. RESEARCH GAPS & CHALLENGES

Research gap and challenges are:

- Lack of real-time prediction systems for clinical use.
- Limited diversity in datasets; most use PIMA dataset.
- Insufficient studies incorporating continuous glucose monitoring.
- Need for interpretable models to aid healthcare professionals.
- Inadequate handling of class imbalance and noisy data.

VII. CONCLUSION & FUTURE DIRECTION

The research concludes that no single approach or strategy best for every data or dataset that predicts diabetes, rather ensemble by a group of robust classifiers, feature design, and optimization techniques is best. Ensemble methodologies involving Random Forest, XGBoost, AdaBoost, and stacked models proved to be superior in terms of accuracy and reliability over single stand-alone classifiers. Deep learning models similarly give the most accurate prediction performance if they are trained using extensive and varied data sets. Integration of sensor-based data and reinforcement learning is an emerging trend toward developing smart, adaptive healthcare systems. Generally speaking, the future of diabetes prediction will lie in bringing together accurate, interpretable, and real-time decision-support systems integrating data-driven methodologies with domain knowledge.

VIII. ACKNOWLEDGMENT

"Authors acknowledge the support from department of Al&DS, K S School of Engineering and Management for the facilities and support provided to carry out project work."

REFERENCES

[1] V. K. Daliya, T. K. Ramesh, "A Cloud-Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression," IEEE Access, vol. 13, pp. 11560-11563, 2025. DOI: 10.1109/ACCESS.2025.3528033.

[2] G. Parimala, R. Kayalvizhi, S. Nithiya, "Diabetes Prediction using Machine Learning," Proc. ICCCI 2023, pp. 1-5. DOI: 10.1109/ICCCI56745.2023.10128216.

[3] Khaled Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," IEEE Access, vol. 12, pp. 16783-16785, 2024. DOI: 10.1109/ACCESS.2024.3359760

[4] Anuj Mangal, Vinod Jain, "Performance Analysis of Machine Learning Models for Prediction of Diabetes," IEEE CISCIT, 2022, DOI: 10.1109/CISCIT55310.2022.10046630R.

[5] G Ravi Kumar, Reddyvari Venkateswara Reddy, Jayarathna M, N Pughazendi, S Vidyullatha and Pundru Chandra Shaker Reddy, "Web Application-Based Diabetes Prediction Using Machine Learning," Proc. ACCAI 2023, DOI: 10.1109/ACCAI58221.2023.10200323.

[6] Arwatki Chen Lyngdoh, Nurul Amin Choudhury, and Soumen Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithms," in *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2020. DOI: 10.1109/IECBES48179.2021.9398759.

[7] Ifra Shaheen, Nadeem Javaid, Nabil Alrajeh, Yousra Asim, and Sheraz Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3398198.

[8] Ahmed Ali Linkon, Inshad Rahman Noman, Md Rashedul Islam, Joy Chakra Bortty, Kanchon Kumar Bishnu, Araf Islam, Rakibul Hasan, and Masuk Abdullah, "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3488743.

[9] Thavavel Vaiyapuri, Ghada Alharbi, Santhi Muttipoll Dharmarajlu, Yassine Bouteraa, Sanket Misra, Janjhyam Venkata Naga Ramesh And Sachi Nandan Mohanty, "IoT-Enabled Early Detection of Diabetes Diseases Using Deep Learning and Dimensionality Reduction Techniques," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3455751.

[10] Usama Ahmed, Ghassan F. Issa, Muhammad Adnan Khan, Shabib. Aftab, Muhammad Farhan Khan, Raed A. T. Said, Taher M. Ghazal, and Munir Ahmad , "Prediction of Diabetes Empowered With Fused Machine Learning," *IEEE Access*, 2022. DOI: 10.1109/ACCESS.2022.3142097.

[11] Santosh Kumar Sharma, Abu Taha Zamani, Ahmed Abdelsalam, Debendra Muduli, Amerah A. Alabrah, Nikhat Parveen, and Sultan M. Alanazi, "A diabetes monitoring system and health-medical service composition model in cloud environment," *IEEE Access*, 2023. DOI: *10.1109/ACCESS.2023.3258549*

[12] Aditi Site, Jari Nurmi, and Elena Simona Lohan, "Machine-learning-based diabetes prediction using multisensor data," *IEEE Access*, vol. 23, NO. 22, pp. 15-11, 2023.

[13] Nor Nisha Nadhira Nazirun, Asnida Abdul Wahab, Ali Selamat, Hamido Fujita, Ondrej Krejcar, Kamil Kuca, and Gan Hong Seng, "Prediction models for Type 2 diabetes progression: A systematic review," *IEEE Access*, 2024. DOI: 1*0.1109/ACCESS.2024.3432118*

[14] Mana Saleh Al Reshan, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Hani Alshahrani, Asadullah Shaikh, and Mohamed A. Elmagzoub, "An innovative ensemble deep learning clinical decision support system for diabetes prediction," *IEEE Access*, 2024. DOI: *10.1109/ACCESS.2024.3436641*

[15] Rishi Jain, Nitin Kumar Tripathi, Millie Pant, Chutiporn Anutariya, and Chaklam Silpasuwanchai, "Investigating gender and age variability in diabetes prediction: A multi-model ensemble learning approach," *IEEE Access*, 2024. DOI: *10.1109/ACCESS.2024.3402350*

[16] Radwa Marzouk, Ala Saleh Alluhaidan, and Sahar A. El Rahman, "An analytical predictive model and secure web-based personalized diabetes monitoring system," *IEEE Access*, 2022. DOI: *10.1109/ACCESS.2022.3211264*

[17] Mohammad Zubair Khan, R. Mangayarkarasi, C. Vanmathi, and M. Angulakshmi, "Bio-inspired PSO for improving neural-based diabetes prediction system," *pp. 06-05*-2020. DOI: 0.13052/jicts2245-800X.1025

[18] Shamim Ahmed, M. Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson, "A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions," *IEEE Access*, 2024. DOI: *10.1109/ACCESS.2024.3422319*

[19] Nada Y. Philip, Manzoor Razaak, John Chang, Suchetha M., Maurice O'Kane, and Barbara K. Pierscionek, "A Data Analytics suite for exploratory, predictive, and visual analysis of Type 2 diabetes," *IEEE Access*, 2022. DOI: *10.1109/ACCESS.2022.3146884*

[20] Hakim El Massari, Zineb Sabouri, Sajida Mhammedi, and Noreddine Gherabi, "Diabetes prediction using machine learning algorithms and ontology," pp. 11-05, 2022. doi: 10.13052/jicts2245-800X.10212

[21] Liyan Jia, Zhiping Wang, Siqi Lv, and Zhaohui Xu, "PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance and Missing Values," *IEEE Access*, vol. 10, pp. 109422–109432, 2022, doi: 10.1109/ACCESS.2022.3212067.

[22] Pierluigi Francesco De Paola, Alessia Paglialonga, Pasquale Palumbo, Karim Keshavjee, Fabrizio Dabbene, and Alessandro Borri, "The Long-Term Effects of Physical Activity on Blood Glucose Regulation: A Model to Unravel Diabetes Progression," *IEEE Control Systems Letters*, vol. 7, pp. 247–252, 2023.

[23] Hamdi A. Al-Jamimi, "Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction," *IEEE Access*, vol. 12, pp. 54812–54825, 2024, doi: 10.1109/ACCESS.2024.3395512.

[24] Nur Ghaniaviyanto Ramadhan, Adiwijaya, Warih Maharani, and Alfian Akbar Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," *IEEE Access*, vol. 12, pp. 67692–67710, 2024, doi: 10.1109/ACCESS.2024.3406748.

[25] Kok-Lim Alvin Yau, Yung-Wey Chong, Xiumei Fan, Celimuge Wu, Yasir Saleem And Phei-Ching Lim, "Reinforcement Learning Models and Algorithms for Diabetes Management," *IEEE Access*, vol. 11, pp. 123456–123469, 2023, doi: 10.1109/ACCESS.2023.3259425.

[26] Muhammad Nauman, Ahmad S. Almadhor, Mohammed Albekairi, Ali R. Ansari, Muhammad A. B. Fayyaz, and Raheel Nawaz, "The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making," *IEEE Access*, vol. 13, pp. 123456–123470, 2025, doi: 10.1109/ACCESS.2025.3526456.
[27] Virginie Felizardo, Diogo Machado, Nuno M. Garcia, Nuno Pombo, and Pedro Brandão, "Hypoglycaemia Prediction Models With Auto Explanation," *IEEE Access*, vol. 9, pp. 126509–126524, 2021, doi: 10.1109/ACCESS.2021.3117340.
[28] Serena Zanelli, Mounim A. El Yacoubi, Magid Hallab, and Mehdi Ammi, "Type 2 Diabetes Detection With Light CNN From Single Raw PPG Wave," *IEEE Access*, vol. 11, pp. 20323–20335, 2023, doi: 10.1109/ACCESS.2023.3274484.
[29] Hafeez UR Rehman Siddiqui, Riccardo Russo, Adil Ali Saleem, Sandra Dudley, and Furqan Rustam, "Improving Automated PSN Assessment in Type 2 Diabetes: A Study on Plantar Lesion Recognition and Probe Avoidance Techniques," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3430194
[30] Mohammad Z. Atwany, Abdulwahab H. Sahyoun, and Mohammad Yaqub, "Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey," *IEEE Access*, 2022. DOI: 10.1109/ACCESS.2022.3157632
[31] Avishek Anishkar Ram, Zain Ali, Vandana Krishna, Nandita Nishika, and Anuranganand Sharma, "A Guided Neural Network Approach to Predict Early Readmission of Diabetic Patients," *IEEE Access*, 2023. DOI: 10.1109/ACCESS.2023.3275086
[32] Mayank Jichkar, Ritesh Shende, Om Bonde, Poorva Agrawal, Gopal Kumar Gupta, and Ajit K. Singh, "Diabetes Prediction Using Machine Learning," in *2024 IEEE International Conference on SILCON*, 2024. DOI: 10.1109/SILCON63976.2024.10910769
[33] Niels F. Cleymans, Mark Van De Casteele, Julie Vandewalle, Aster K. Desouter, Frans K. Gorus, and Kurt Barbé, "Analyzing Random Forest's Predictive Capability for Type 1 Diabetes Progression," *IEEE Open Journal of Instrumentation and Measurement*, 2025. DOI: 10.1109/OJIM.2025.3551837
[34] Mr. Khaja Mannanuddin, Logeshwari Dhavamani, and V. Selvakumar, "An Ensemble Learning Approach Towards Prediction of Diabetic Retinopathy," in *2023 IEEE International Conference on Intelligent Informatics and Processing (ICIIP)*, 2023. DOI: 10.1109/ICIIP61524.2023.10537626
[35] Sadia Afrin Shampa, Md. Saiful Islam, and Ayatun Nesa, "Machine Learning-based Diabetes Prediction: A Cross-Country Perspective," in *2023 IEEE International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 2023. DOI: 10.1109/NCIM59001.2023.10212596
[36] Gresha Bhatia, Aditya Deopurkar, Shravan Bhat, Sahil Talreja, and Vivek Choudhary, "Disease Prediction using Deep Learning," in *2021 IEEE International Conference on Emerging Trends in Information Technology and Engineering (INCET)*, 2021. DOI: 10.1109/INCET51464.2021.9456172
[37] Navaneeth Bhaskar, Vinayak Bairagi, Ekkarat Boonchieng, and Mousami V. Munot, "Automated Detection of Diabetes From Exhaled Human Breath Using Deep Hybrid Architecture," *IEEE Access*, 2023. DOI: 10.1109/ACCESS.2023.3278278
[38] Praveen Tumuluru, Lakshmi Ramani Burra, Katuku Krishna Sushanth, Shaik Nagoor Vali, CH.M.H. Sai Baba, and Pachipala Yellamma, "DPMLT: Diabetes Prediction Using Machine Learning Techniques," in *2022 IEEE International Conference on Electronics and Renewable Systems (ICEARS)*, 2022. DOI: 10.1109/ICEARS53579.2022.97519
[39] Farrukh Aslam Khan, Khan Zeb, Mabrook Al-Rakhami, Abdelouahid Derhab, and Syed Ahmad Chan Bukhari, "Detection and prediction of diabetes using data mining: A comprehensive review," *IEEE Access*, vol. 9, pp. 50020–50045, 2021, doi: 10.1109/ACCESS.2021.3059343.
[40] Asif Hassan Syed and Tabrej Khan, "Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: A retrospective cross-sectional study," *IEEE Access*, vol. 8, pp. 199444–199456, 2020, doi: 10.1109/ACCESS.2020.3035026.
[41] Evgenii A. Pustozerov, Aleksandra S. Tkachuk, Elena A. Vasukova, Anna D. Anopova, Maria A. Kokina, Inga V. Gorelova, Tatiana M. Pervunina, Elena N. Grineva, and Polina V. Popova, "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," *IEEE Access*, vol. 9, pp. 143780–143789, 2021, doi: 10.1109/ACCESS.2021.3116383.
[42] Shirina Samreen, "Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble," *IEEE Access*, vol. 9, pp. 144193–144208, 2021, doi: 10.1109/ACCESS.2021.3116383.
[43] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 555–561, Mar. 2014, doi: 10.1109/JBHI.2013.2278281.
[44] Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi, "Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques," *IEEE Access*, vol. 6, pp. 49279–49289, 2018, doi: 10.1109/ACCESS.2018.2884249.
[45] Md. Shafiqul Islam, Marwa K. Qaraqe, Samir Brahim Belhaouari, and Muhammad A. Abdul-Ghani, "Advanced techniques for predicting the future progression of type 2 diabetes," *IEEE Access*, vol. 8, pp. 125162–125172, 2020, doi: 10.1109/ACCESS.2020.3005540.
[46] Qian Wang, Weijia Cao, Jiawei Guo, Jiadong Ren, Yongqiang Cheng, and Darryl N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102231–102242, 2019, doi: 10.1109/ACCESS.2019.2929866.
[47] Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 99642–99656, 2021, doi: 10.1109/ACCESS.2021.3098691.
[48] Anastasios Alexiadis, Athanasios Tsanas, Leonard Shtika, Vassilis Efopoulos, Konstantinos Votis, Dimitrios Tzovaras, and Andreas Triantafyllidis, "Next-day prediction of hypoglycaemic episodes based on the use of a mobile app for diabetes self-management," *IEEE Access*, vol. 12, pp. 6404–6415, Jan. 2024, doi: 10.1109/ACCESS.2024.3350201.
[49] Norma Latif Fitriyani, Muhammad Syafruddin, Ganjar Alfian, and Jongtae Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," *IEEE Access*, vol. 7, pp. 144828–144838, 2019. doi: 10.1109/ACCESS.2019.2945129.
[50] Mehrbakhsh Nilashi, Othman Ibrahim, Mohammad Dalvi, Hossein Ahmadi, and Leila Shahmoradi, "Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset," *Future Generation Computer Systems*, vol. 92, pp. 62–74, 2019. doi: 10.1016/j.fiae.2017.09.006.
[51] Tuan Minh Le, Thanh Minh Vo, Tan Nhat Pham, and Son Vu Truong Dao, "A Novel Wrapper Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic," *IEEE Access*, vol. 9, pp. 184846–184859, Jan. 2021. doi: 10.1109/ACCESS.2020.3047942.
[52] Giulia Noaro, Taiyu Zhu, Giacomo Cappon, Andrea Facchinetti, and Pantelis Georgiou, "A Personalized and Adaptive Insulin Bolus Calculator Based on Double Deep Q-Learning to Improve Type 1 Diabetes Management," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2152–2160, May 2023. doi: 10.1109/JBHI.2023.3245803.
[53] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 44–54, Jun. 2024. doi: 10.1109/OJEMB.2024.3365290.
[54] Saúl Langarica, Diego De La Vega, Nawel Cariman, Martín Miranda, David C. Andrade, Felipe Núñez, and Maria Rodriguez-Fernandez, "Deep Learning-Based Glucose Prediction Models: A Guide for Practitioners and a Curated Dataset for Improved Diabetes Management," *IEEE Access*, vol. 7, pp. 103444–103465, 2019. doi: 10.1109/ACCESS.2019.2931956.
[55] H. Roopa and T. Asha, "A Linear Model Based on Principal Component Analysis for Disease Prediction," *IEEE Access*, vol. 7, pp. 149841–149850, 2019. doi: 10.1109/ACCESS.2019.2949210.