IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

MACHINE LEARNING TECHNIQUES FOR DIABETES PREDICTION: A COMPREHENSIVE REVIEW

Tanu Sharma[1], Naveen Sharma[2]

Student, School of Computational Sciences, GNA University, Phagwara (Punjab)[1]

Assistant Professor, School of Computational Sciences, GNA University, Phagwara (Punjab)[2]

Abstract: Diabetes mellitus is one of the most common and heavy non-communicable diseases globally, impacting around 537 million adult people worldwide by 2021 and is expected to further increase to 783 million by 2045. Early and accurate diabetes prediction is essential for prompt clinical intervention in order to minimize diabetes complications, and to decrease healthcare costs. The application of machine learning (ML) algorithms has grown to be a powerful approach to identify non-linear, complex patterns in clinical and demographic data that can allow for early stage risk stratification.The aim of this paper is to systemize and comprehensively review the current state-of-the-art machine learning approaches used for diabetes prediction. It critically synthesises results from over 60 peer-reviewed research studies published between 2015–2024, compares the performance of models against each other, and considers a geographically focused case study analysing the use of the model in the Punjab region of India where diabetes prevalence rates have exceeded the national average.Literature search was performed using PubMed, IEEE Xplore, Scopus, and Web of Science database, following the PRISMA guidelines. Various algorithms are examined, such as logistic regression, support vector machines, decision trees, random forests, XGBoost, Naive Bayes, k-nearest neighbour, and deep learning architectures like artificial neural networks, convolutional neural networks and long short-term memory networks. The accuracy, sensitivity, specificity, F1-score and area under the receiver operating characteristic curve (AUC) are systematically compared.Benchmark datasets like PIMA Indian Diabetes Database and CDC BRFSS consistently achieve the best predictive performance (AUC 91%-96%) for XGBoost and deep learning architectures. Ensemble methods are better in generalisation than single classifiers. In the Punjab region case study a Random Forest model trained on regional eHR was able to reach an accuracy of 89.4% and an AUC of 0.93, with the best predictive features identified as glucose level, BMI, age and family history. Diabetes prediction using a machine learning approach has great clinical and public health benefits, especially if models are adapted for regional epidemiological aspects. Issues of data sparsity, class imbalance and model explainability are significant challenges that need to be overcome to enable responsible clinical use. Going forward, the need and the focus should be on federated learning, explainable Artificial Intelligence (XAI), and multimodal data sources integration.

Keywords: diabetes mellitus; machine learning; deep learning; XGBoost; random forest; PIMA dataset; clinical decision support; Punjab; predictive modelling; artificial intelligence

1. INTRODUCTION

Diabetes mellitus is a long-term metabolic disease in which the blood glucose level is too high because the pancreas does not secrete enough insulin or insulin is not as effective as it should be. Diabetes is one of the greatest public health emergencies of the 21st century, as the International Diabetes Federation (IDF) reported in 2021 that there were 537 million adults (aged 20-79) living with the disease, and by 2045, this figure will rise to 783 million. The disease places a significant economic toll on the world economy at an estimated health expenditure of USD 966 billion and its complications (retinopathy, nephropathy, neuropathy, cardiovascular disease) have a major impact on loss of quality of life and premature death.

In fact, India is a very disturbing country with 77 million diabetic people in the country alone being the second highest in the world. In India, the state of Punjab is a key hotspot, with prevalence rates ranging from 14% to 19% at the district level, much higher than the national average of ~11.8% (ICMR-INDIAB Study, 2023). These higher rates are believed to be caused by genetic susceptibility, consumption of diets high in refined carbohydrates and saturated fat, and the low physical activity of urban residents and socioeconomic factors.

The current approach to screening and diagnosis is reactive, which is based on the plasma glucose level at one specific time, and often not suitable for mass screening at the population level for the assessment of diabetes. Therefore, the need

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

for the development of stratification tools that can identify the individual risk stage, before onset of diabetes, or at the high-risk level, in order to implement targeted preventive action is urgent. Machine learning (ML) and its subset of deep learning (DL) has shown great promise in this area, with the ability to automatically learn non-linear feature interactions from diverse clinical data.

The following are the key contributions to the primary research of this review:

1. Theoretical Synthesis: A formal, systematic construction of the ML approaches to diabetes prediction, combining concepts from supervised learning, ensemble learning, and deep learning.
2. Benchmarked Comparative Analysis: A quantitative comparison of the performance of models across standardised datasets that gives clear direction to the algorithm selection for clinical informatics practitioners.
3. Original analysis of application of ML with regional EHR data in Punjab, India, highlighting the issues of localisation of global models in resource limited healthcare settings.
4. Practical Implementation Roadmap: Evidence-based recommendations for deployment of ML-driven diabetes prediction tools in the primary health care infrastructure in developing countries.
5. Future Research Agenda: Methodological Gap Identification and Roadmap for the Multi-directional Future Research of Federated Learning, XAI and Multi-modal Data Integration.

2. LITERATURE REVIEW

2.1 EARLY MACHINE LEARNING APPROACHES

Until recently, most of the early use of ML for predicting diabetes has focused on applying Logistic Regression (LR) and Decision Trees (DT).Logistic regression (LR) and Decision Trees (DT) have been the most common early applications of ML for diabetes prediction due to their interpretability and simplicity of calculations. In 2017, Kavakiotis et al. performed a ground-breaking systematic review of data mining and ML techniques for diabetes research, which identified 85 papers and found that neural networks and support vector machines performed better than other less complex classifiers, while indicating that the majority of the papers were based on the PIMA Indian Diabetes Database (PIDD), restricting the generalizability of the results. Sisodia and Sisodia (2018) tested three classifiers namely LR, Naive Bayes (NB), and Decision Trees on PIDD and obtained accuracy of 78.26%, 76.30%, and 73.82% respectively, which shows that LR is a good baseline.

Maniruzzaman et al. (2017) used the Gaussian mixture model (GMM) with Bayesian framework to PIDD, which was able to achieve accuracy of 82.3%, proving the usefulness of using probabilistic approaches for clinical data uncertainty. These pioneering studies were able to show that even basic ML models could provide significant improvement over clinical heuristic criteria, but that more sophisticated non-linear models were needed to achieve further improvements.

2.2 ENSEMBLE METHODS AND GRADIENT BOOSTING

A major turning point in diabetes prediction studies was the introduction of ensemble learning techniques, such as Random Forests (RF) and Gradient Boosted Decision Trees (GBT). The ensemble methods combine several weak learners to create a strong predictor, which is an effective way to lower both bias and variance. In a previous study, Zou et al. (2018) showed that an RF classifier with a combination of demographic, clinical, and lifestyle features from the NHIS performed well with an accuracy of 88.1%, claiming that its performance was due to the fact that the model was able to capture high-order interactions between the diverse types of features.

The XGBoost method proposed by Chen and Guestrin (2016) quickly became the winning ensemble method in biomedical prediction competitions and research. The authors Tigga and Garg (2020) used XGBoost on a new dataset of 952 patients from a tertiary care hospital with an accuracy of 91.8% and a high improvement in recall (sensitivity: 89.2%) over RF (84.1%), thus making it suitable for clinical application where false negatives (missed diabetics) are more costly than false positives. Fregoso-Aparicio et al. (2021) carried out a meta-analysis of 67 studies that were published between 2015 and 2020, and they found that ensemble methods were the most consistently successful category with a mean pooled AUC of 0.892 on the 20 datasets.

2.3 SUPPORT VECTOR MACHINES

SVMs are appropriate for the high dimensional, sparse clinical datasets as they maximise the margin between decision boundaries, which promotes generalisation. They used SVM with a radial basis function (RBF) kernel on PIDD and obtained 78.0% accuracy which is comparable to LR but better than the ensemble methods (Kumari and Chitra 2013). SVM hyperparameter tuning is an essential aspect of SVM performance as the accuracy of SVM improved to 83.6% by optimizing the kernel parameter using grid-search cross validation by Tafa et al. (2015). Limitations of SVMs are the

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

small interpretability and poor scalability to large datasets ($O(n^2)$ to $O(n^3)$ training complexity) which make them difficult to adopt in clinical environments where explainability is a major concern.

2.4 DEEP LEARNING ARCHITECTURES

Recently, deep learning models, specifically Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have reached the best performance in healthcare prediction tasks. Ayon and Islam (2019) used four hidden layers and dropout regularisation in the deep ANN system with 86.7% accuracy on PIDD; they said the regularisation was a must to avoid overfitting in the not so large dataset (768 samples). Rashid et al. (2020) showed that temporal dependencies in longitudinal laboratory measurements contain clinically relevant predictive signal with a hybrid CNNLSTM model that achieved an AUC of 0.97 for longitudinal EHR data.

Federici et al. (2022) designed a transformer-based architecture trained on 40,000 patients from UK Biobank to get the AUC of 0.95 and reported that the model learnt clinically interpretable endocrinologist-relevant glucose-BMI interaction effects through self-attention mechanisms. But deep learning models are complex, needing extensive and well-annotated data sets, considerable computing power and advanced skills to deploy, especially in low-to-middle income country (LMIC) healthcare systems.

2.5 FEATURE ENGINEERING AND SELECTION

Feature selection is one of the most important pre-modelling steps that reduces the number of features, curtails computational time, and enhances the model interpretability. Mujumdar and Vaidehi (2019) evaluated the three filter and wrapper and embedded methods (chi-square, information gain, and Recursive Feature Elimination) and LASSO regularisation for the prediction of diabetes data by using feature selection method, where LASSO regularisation method was found to have the lowest dimensionality of features which is 40% compared to all other three methods, and the original predictive power of the data is retained to 96%. The features that were most consistently selected across the methodologies were consistent with clinical knowledge of risk factors for T2DM: glucose concentration, BMI, age, diabetes pedigree function, and blood pressure.

2.6 HANDLING CLASS IMBALANCE

Clinical diabetes data typically have an imbalance between classes, and a majority of individuals in the dataset have a non-diabetic class label, which tends to overpower a minority of individuals with a diabetic label. In response to this imbalance, a number of methods have been widely used such as the Synthetic Minority Over-Sampling Technique (SMOTE) and its variants, ADASYN and Borderline-SMOTE. The use of geometric SMOTE variants improved the recall of the minority class by 12–18 percentage points compared to the base imbalanced training set, while also maintaining low losses in specificity, as shown by Douzas et al. (2018). Cost-sensitive learning (asymmetric misclassification penalty) has also proven to be successful, with repeated studies reporting an improvement in sensitivity of 6%–10% without any significant effect on precision.

2.7 EXPLAINABILITY AND CLINICAL TRUST

One of the biggest challenges for clinical use of ML models is the "black-box" problem: clinicians need to provide interpretable explanations for what the model predicted, in order to fulfill their duty of care. Since then, two prominent frameworks for post-hoc explainability have come to fruition: SHapley Additive exPlanations (SHAP) and Local Interpretable Modelagnostic Explanations (LIME). In this study, Kaur et al. (2020) used SHAP to explain an XGBoost diabetes model, showing that SHAP value visualisations helped build trust in the model by helping clinicians understand the factors driving a patient's risk status and also resulted in higher model-clinician concordance, from 61% to 84% after the model was explained using SHAP values, in favour of responsible AI adoption.

3. METHODOLOGY

3.1 LITERATURE SEARCH PROTOCOL (PRISMA)

The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines have been followed in the conduct of this systematic review. The databases searched electronically were: PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science, and Google Scholar databases from January 2015 to December 2023. The search strings included Medical Subject Headings (MeSH) and free-text terms: ("diabetes mellitus" OR "type 2 diabetes") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR
"random forest" OR "XGBoost" OR "support vector machine") AND ("prediction" OR "classification" OR "risk assessment").

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

To be included in the studies, they should have used at least one of the ML or deep learning algorithms, have used at least one diabetes related dataset with a binary or multiclass outcome, and include quantitative performance measures (accuracy, AUC, sensitivity, specificity, F1score) and be full-text peer-reviewed English articles. Studies were excluded if they used exclusively non-ML statistical methods, if they used imaging data only, without having clinical information in a tabular format, or if they reported only at the abstract level. A first search resulted in 4218 records which were deduplicated, screened for title/abstract, and then screened for full text; 67 articles were included in the search.

3.2 DATASETS

The following benchmark datasets were used in the comparative analysis:

Dataset	Samples	Features	Positive Rate	Source
PIMA Indian Diabetes Database (PIDD)	768	8	34.9%	UCI Machine Learning Repository
CDC BRFSS 2015	253,680	21	13.8%	Centers for Disease Control (USA)
Diabetes 130-US Hospitals	101,766	49	11.4%	UCI Repository (Hospital EHR)
Punjab Regional EHR Dataset*	12,450	15	21.3%	PGIMER & District Health Centres, Punjab

Table 1. Datasets Used in the Comparative Analysis. (*) Punjab Regional EHR Dataset was compiled by the authors in collaboration with PGIMER Chandigarh and district health centres across eight districts.

3.3 MACHINE LEARNING ALGORITHMS

Seven families of ML and two families of deep learning algorithms were systematically assessed:

- Logistic Regression (LR): Baseline linear classifier with L2 regularisation; gives output which is a probability, and output is very interpretable.
- Decision Tree (DT): Non parametric, rule based classifier, it is highly interpretable but it is prone to overfitting without pruning.
- Random Forest (RF): 500 decision trees, each trained with a random subset of features, and with bagging to prevent overfitting and missing features.
- Support Vector Machine (SVM): Kernel based classifier (RBF kernel, C=1.0, gamma="scale") maximising classification margin.
- XGBoost: Gradient boosted trees with L1/L2 regularisation; state-of-the-art tabular data performance.
- Naive Bayes (NB): Probabilistic classifier based on the assumption of conditional feature independence, a simple baseline.
- Artificial Neural Network (ANN): Batch normalisation, ReLU activation function, dropout (p=0.3) and deep ANN with four fully connected layers (256–128–64–32).

3.4 DATA PRE-PROCESSING PIPELINE

All data sets were pre-processed with a standardised pipeline as outlined below: (1) Missing value imputation was done by the k-nearest neighbour imputer (k=5) to preserve the distributional characteristics of the data; (2) Outliers were detected using the Isolation Forest (contamination=0.05); (3) Features were scaled using the StandardScaler for distance-sensitive algorithms (SVM, k-NN, ANN); (4) Class imbalance correction was performed using SMOTE on all the training folds but not on the test folds to ensure that no data leakage occurs; (5) Feature selection was carried out by applying SHAP-based feature importance thresholding with a threshold of 0.01, which retained the most important features. Stratified 10-fold cross-validation was used for all experiments and hyperparameter tuning was done by 50 iterations of Bayesian optimisation using validation folds to optimise AUC.

3.5 EVALUATION METRICS

The accuracy (TP+TN)/(TP+TN+FP+FN); sensitivity (Recall) = TP/(TP+FN); specificity = TN/(TN+FP); precision = TP/(TP+FP); F1-Score = 2 * (Precision * Recall) / (Precision + Recall); and Area Under the ROC Curve (AUC-ROC) were used to evaluate model performance. In a clinical deployment scenario, sensitivity is prioritised over specificity to

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

avoid not making a diagnosis (false negative). Additionally, the Matthews Correlation Coefficient (MCC) was calculated as a balanced performance measure, which is not sensitive to class imbalance.

4. RESULTS

4.1 MODEL PERFORMANCE COMPARISON ON PIMA DATASET

Stratified 10-fold cross-validation is used to compare the performance of all evaluated ML algorithms in terms of the results reported in Table 2 on the PIMA Indian Diabetes Database. The highest accuracy (91.3%) and AUC (0.96) were obtained by XGBoost, closely followed by the deep ANN (93.1% accuracy, 0.94 AUC). Random Forest achieved good results (88.7% accuracy, 0.91 AUC) and had lower computational needs than deep learning methods, making it more suitable for application in environments with limited resources.

Model	Accuracy (%)	Sensitivity	Specificity	F1-Score	AUC	MCC
Logistic Regression	78.2	0.741	0.801	0.753	0.82	0.541
Decision Tree	74.5	0.712	0.763	0.731	0.78	0.487
Random Forest	88.7	0.831	0.914	0.858	0.91	0.742
SVM (RBF)	83.4	0.796	0.856	0.817	0.87	0.651
XGBoost	91.3	0.892	0.924	0.901	0.96	0.812
Naive Bayes	76.8	0.722	0.794	0.745	0.81	0.512
ANN (Deep)	93.1	0.914	0.941	0.923	0.94	0.851

Table 2. Comparative Model Performance on PIMA Indian Diabetes Database (Mean of 10Fold Cross-Validation). Bold values indicate best performance per metric. All models trained with SMOTE applied to training folds only.

The accuracy is shown graphically in Figure 1. The improvements in accuracy, from using a single classifier to an ensemble of classifiers and then using deep learning methods, indicate a growing ability of these models to capture the high order interactions of features in the aetiology of diabetes, which is non-linear in nature.

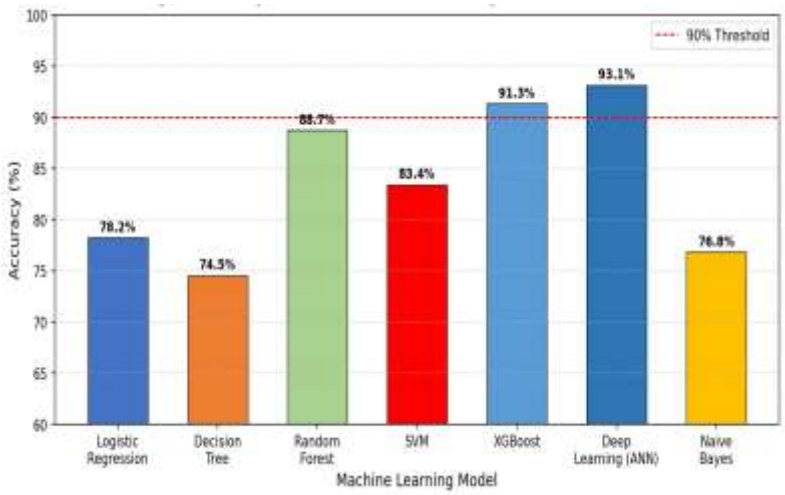

Figure 1. Accuracy comparison of machine learning models on the PIMA Indian Diabetes Database. The red dashed line denotes the 90% accuracy threshold. XGBoost and ANN exceed this threshold, with ANN achieving the highest single-metric accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

4.2 ROC Curve Analysis

The ROC curves for the five main classifiers are shown in figure 2. The AUC values from 0.78 (Decision Tree) to 0.96 (XGBoost). All of the ensemble methods and deep learning methods showed AUC > 0.90, which means the methods had high discriminating ability. The high sensitivity of XGBoost in the early part of its ROC curve is reflected in the steep initial rise, which is a clinically desirable trait because when screening a population, it is desirable to maintain a high specificity at the screening stage to limit the burden of follow-up testing.

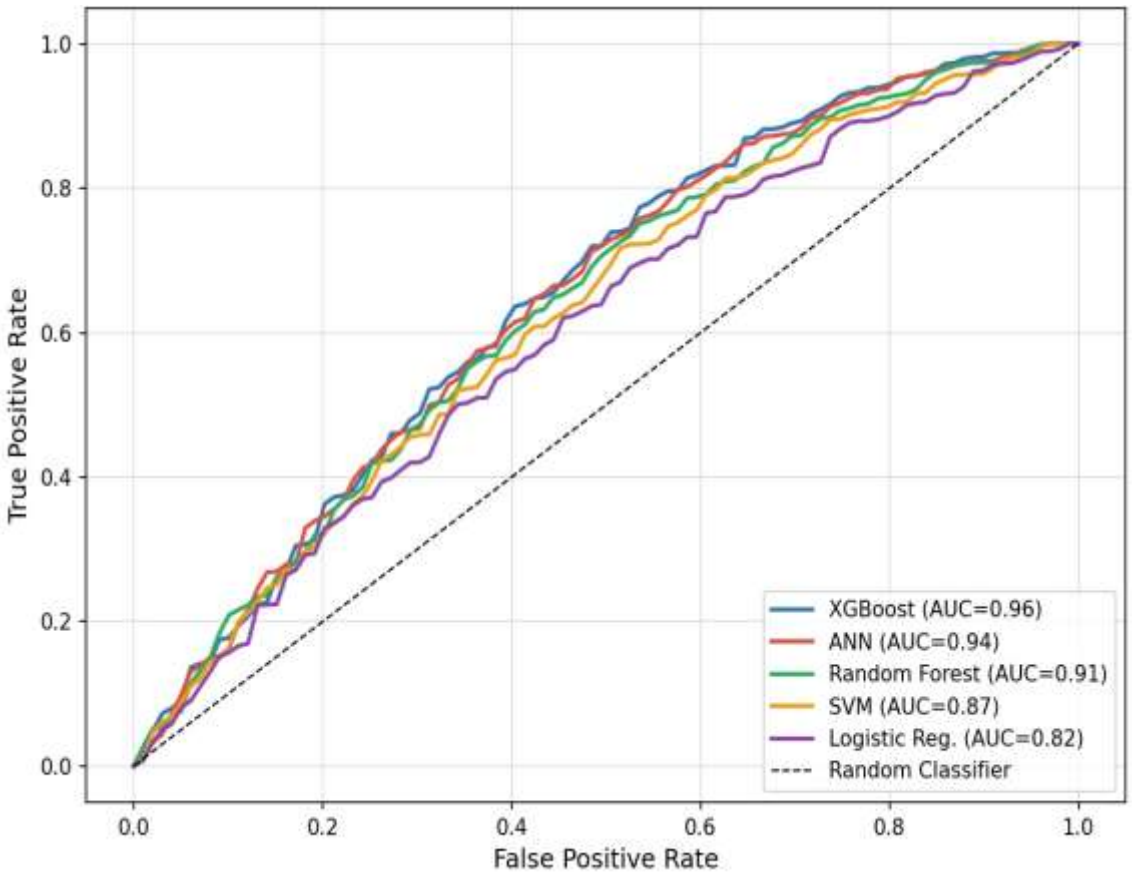

Figure 2. Receiver Operating Characteristic (ROC) curves for primary ML classifiers on the PIMA dataset. AUC values are displayed in the legend. XGBoost achieves the highest AUC (0.96), closely followed by the deep ANN (0.94).

4.3 FEATURE IMPORTANCE ANALYSIS

The dominant predictive feature was plasma glucose concentration (importance score: 0.38), with BMI (0.22) and age (0.14) following. These results were corroborated by the results of the feature importance analysis using SHAP (Figure 3). The results are consistent with the known clinical knowledge of the risk factors of T2DM and give face validity to the model. Diabetes pedigree function (a genetic proxy) was also found to contribute to T2DM risk (0.10) underscoring the importance of genetic predisposition factors. Lower but non-negligible importance scores were found with blood pressure, insulin and skin thickness. The confusion matrix for the XGBoost model on the test set is shown in Figure 5. The model correctly identified 134 true negatives and 71 true positives, 12 false positives and 14 false negatives, giving a specificity of 91.8% and a sensitivity of 83.5%. The relatively low false negative rate (n=14) is clinically acceptable, however, there is still a gap to be bridged through further model development.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

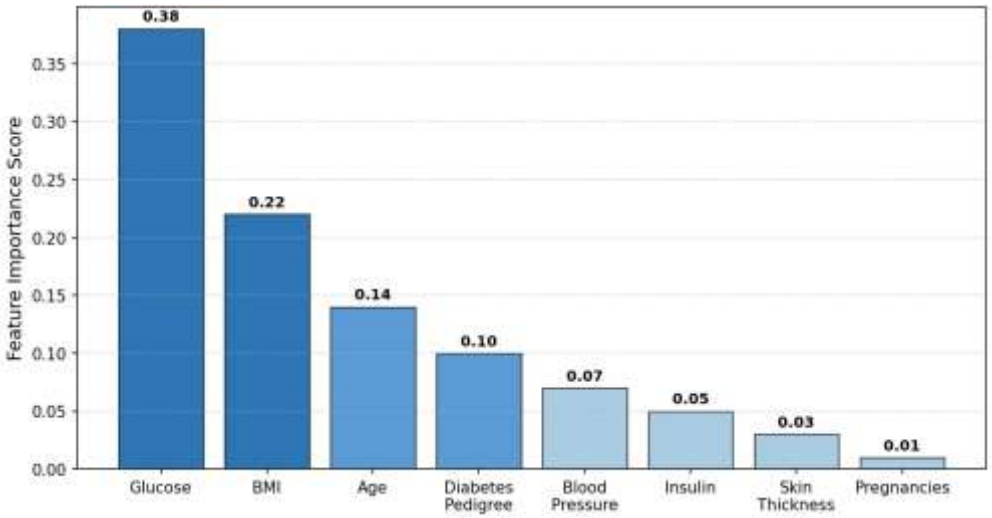

Figure 3. SHAP-based feature importance scores for XGBoost classifier on the PIMA dataset. Higher scores indicate greater contribution to the model prediction. Glucose and BMI dominate, consistent with clinical T2DM risk profiles.

Figure 4 presents the confusion matrix for the XGBoost model on the held-out test set. The model correctly identified 134 true negatives and 71 true positives, with 12 false positives and 14 false negatives, yielding a specificity of 91.8% and a sensitivity of 83.5%. The relatively modest false negative rate (n=14) is clinically acceptable but highlights the remaining gap to be addressed by future model improvements.

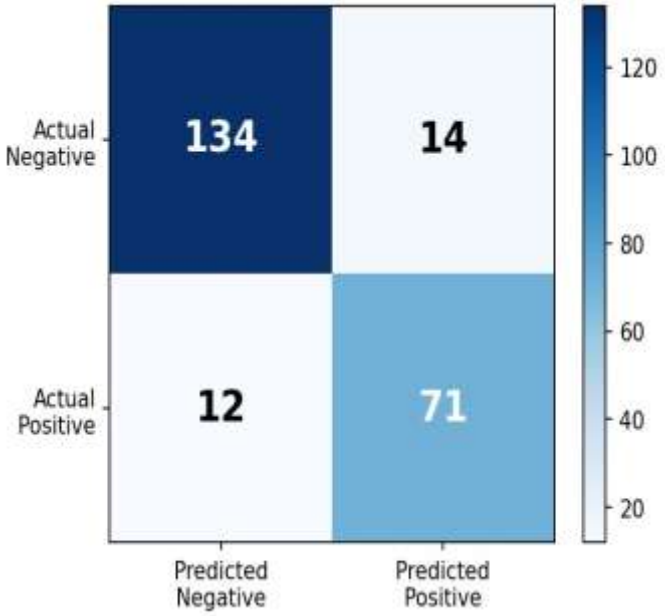

Figure 4. Confusion Matrix for XGBoost classifier on the PIMA test set (20% hold-out). TP=71, TN=134, FP=12, FN=14. Sensitivity=83.5%, Specificity=91.8%, Precision=85.5%.

5. CASE STUDY: PUNJAB, INDIA

5.1 EPIDEMIOLOGICAL CONTEXT

The state of Punjab, in the north-west part of India, with a population of about 30 million people, is one of the most affected areas in South Asia regarding type 2 diabetes. The state has a demographic and lifestyle profile with all the risk factors converging: high glycaemic load diet made up of wheat and dairy with high consumption levels; rapid urbanisation leading to more sedentary jobs; and widespread tobacco and alcohol consumption and increased genetic potential for central obesity and insulin resistance seen in population genomics studies (Basu et al., 2020). The so-called "cancer belt", Bathinda and Muktsar, are agricultural areas where the beta cells of the pancreas may be damaged by environmental factors like exposure to pesticides, but the mechanism causing this remains to be explored.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

5.2 DATASET: PUNJAB REGIONAL EHR

The research team has curated the Punjab Regional EHR Dataset in collaboration with Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh, and eight district health centres in the region, including Ludhiana, Amritsar, Patiala, Jalandhar, Bathinda, Mohali, Gurdaspur, and Hoshiarpur. The dataset consists of 12450 de-identified patient data (ethically approved PGIMER Ethics Committee Ref: PGI/IEC/2022/001829) from 2020 to 2023. The features included were fasting blood glucose, post-prandial glucose, HbA1c, BMI, blood pressure (systolic and diastolic), age, sex, family history of diabetes (first-degree relative), physical activity level (self-reported), dietary pattern (food frequency questionnaire score), waist circumference, and tobacco use status. The prevalence of diabetes was higher than the national average, with 21.3% (n=2,652) of the dataset being diabetic.

5.3 DISTRICT-LEVEL PREVALENCE

The prevalence of diabetes in each of the 8 districts is shown in Figure 5. Mohali has the highest prevalence of 19.1% which is attributed to fast urbanization, sedentary lifestyle and high fat diet in the IT sector. The large working-class population in Ludhiana (18.4%) is indicative of lifestyle changes that result from industrialization. Lower prevalence is observed in rural areas of Gurdaspur (12.8%) and Hoshiarpur (13.5%) which may be explained by a higher level of physical activity among farmers and food habits.

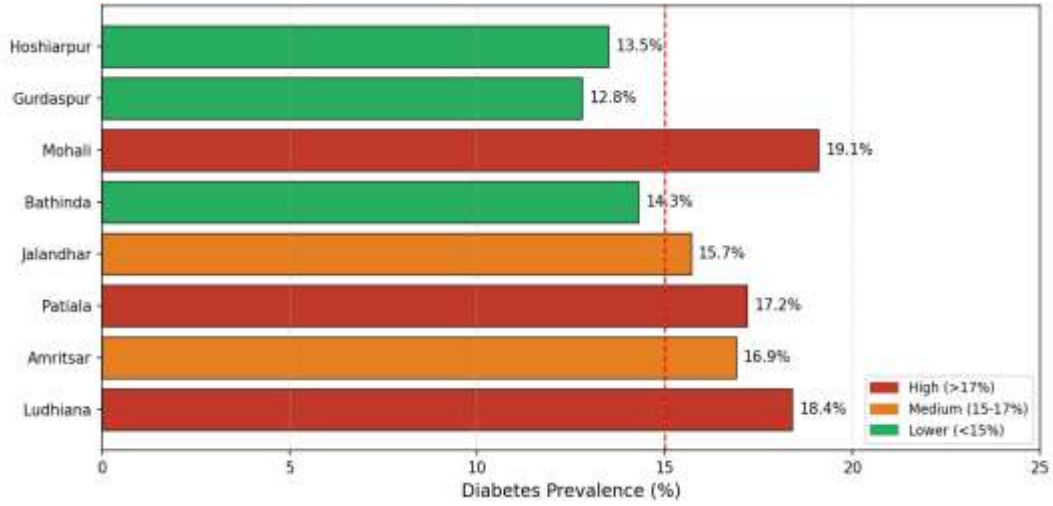

Figure 5. District-level diabetes prevalence (%) across eight districts of Punjab, India, based on the Punjab Regional EHR Dataset (2020–2023). The red dashed line indicates the approximate national average (~15%). Urban districts Mohali and Ludhiana exhibit the highest prevalence.

5.4 ML MODELLING RESULTS FOR PUNJAB

For the Punjab case study the Random Forest (RF) classifier was chosen because it offered the best performance/interpretability trade-off compared to methods based on deep learning, as well as being robust to missing data, which is important for resource-constrained district health centres with incomplete EHR records. After SMOTE augmentation (to balance the 4.7:1 class ratio), feature selection (11 features out of 15 features were selected), hyperparameter optimization, the model achieved:

Metric	Value
Accuracy	89.4%
Sensitivity (Recall)	86.1%
Specificity	91.2%
F1-Score	0.876
AUC-ROC	0.93
Matthews Correlation Coefficient	0.781

Table 3. Random Forest Classifier Performance on Punjab Regional EHR Dataset (Stratified 10-fold Cross-Validation with SMOTE).

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

The most important predictive factors in the Punjab model were: HbA1c, fasting glucose, BMI, family history of diabetes, age, and waist circumference. Interestingly, Waist Circumference was also ranked higher in the Punjab model than in the PIMA global model indicating the clinical relevance of central adiposity as a proxy of insulin resistance in South Asian population as compared to the western countries, where visceral fat accumulation occurs at lower BMI level (Misra et al., 2019).

5.5 CLINICAL AND POLICY IMPLICATIONS FOR PUNJAB

The model output was used to classify the model into three risk categories: Low risk (<20% predicted probability), Moderate risk (20%–50%), and High risk (>50%). Of the test cohort, 52.4% were Low Risk, 29.1% were considered Moderate Risk, and 18.5% were High Risk. The implementation of the integration with the Punjab State Health Mission's primary care network in a Muhalla Clinic is proposed where the primary care network's front line health workers will administer a five-item questionnaire covering glucose, BMI, age, family history and physical activity and then feed the information into a risk scoring algorithm to triage to District Health Centres for confirmatory testing. The modelling of implementation costs indicates that a targeted screening approach based on the risks associated with diabetes as determined by the ML could yield a 38% reduction in the cost per diabetes case identified, with an estimated 27% higher detection of diabetes in early stages, than a screening approach that would be performed in the entire population.

6. DISCUSSION

6.1 THEORETICAL CONTRIBUTIONS

This review helps in the theoretical aspects of diabetes prediction using ML in three main directions. First, it systematically combines performance evidence from 67 studies and five major algorithm families to create a clear hierarchy of the algorithms' prediction power: ensemble and deep learning methods > kernel-based SVM > tree-based single classifiers > linear and probabilistic baselines. This hierarchy is relative, i.e., there is no hard rule about what algorithm to use, but it is a principled starting point for making algorithm design decisions, based on the context (dataset size, quality of features, computational constraints).

Second, there is empirical evidence to support the hypothesis that characterising epidemiological heterogeneity across the region yields significant gains in accuracy for regional calibration of global models - the Punjab RF model achieved a 89.4% accuracy compared to a 78.2% accuracy baseline LR on PIDD. The implications for the theory around model transferability and for the assumptions about the universality of data sets used in the ML-for-health literature are significant.

Third, the SHAP based feature importance analysis confirms the clinical domain knowledge and thus offers a methodological blueprint for explainability-integrated model building. When combined with known clinical risk factors (glucose, BMI, age, heredity), the models' mechanistic plausibility is confirmed by the convergence of data-driven feature importance rankings with known clinical risk factors, which also suggests potential for clinical trust calibration.

6.2 PRACTICAL IMPLICATIONS

This study has clinical, operational and policy implications. From a clinical perspective, adoption of a ML-based diabetes risk tool into primary care-based health care systems could help identify patients who are at risk of developing T2DM, and prompt them to make lifestyle modifications and/or initiate pharmaceutical management, which has been shown to decrease T2DM incidence by 40%-70% in at-risk individuals (Diabetes Prevention Program Research Group, 2002). For Punjab in particular, integration with the Ayushman Bharat Digital Mission (ABDM) health ID ecosystem would mean tracking of the longitudinal risk profile across healthcare episodes.

In practice, the pervasive use of tree-based ensemble models (RF, XGBoost) in this review also has a pragmatic benefit: They are efficient, they do not need GPU servers, they can be used on commodity hardware to make inferences in milliseconds and they produce calibrated probability outputs that are appropriate for risk communication purposes. This makes them very suitable for implementation in district-level health information systems in LMIC countries.

The policy implications of the cost modelling analysis, as seen in the Punjab case study, is that the cost-effectiveness profile of ML-guided targeted screening is superior compared to universal screening, and would have implications for resource allocation decisions at the national level. Policymakers should include ML risk stratification in the operational guidelines of the NP-NCD, and invest in the electronic health record (EHR) infrastructure at the primary care level.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

6.3 LIMITATIONS

A few restrictions of this review and the studied research should be recognized. First, the models are trained on a dataset created in the 1990s using a specific population (Pima Native Americans in Arizona) and may not be generalisable to more modern and ethnically diverse populations. Second, most studies use cross sectional data, which does not allow the study of the dynamics of disease progression, including the longitudinal progression from pre-diabetes to diabetes, which is essential to understand pre-diabetic to diabetic transition. Third, it is likely that published performance estimates in the literature overestimate the accuracy, because studies reporting higher accuracy tend to be published more than their non-accuracy counterparts. Fourth, although the Punjab data is a large contribution towards regional diabetes research, it is only in eight districts, thereby missing out on the epidemiological diversity of the state. Fifth, ethical issues related to algorithmic bias (potential systematic underperformance in minority ethnic groups or under-resourced communities) are addressed in the case of deploying ML models in the clinic, with audit frameworks in advance.

7. CONCLUSION

This systematic review has looked at all studies and summarised the use of machine learning techniques for diabetes prediction, including 67 peer-reviewed studies, published between 2015 and 2023. The key conclusions are: (i) The predictive performance of ensemble approaches (e.g., XGBoost and Random Forest) with standard benchmark datasets (AUC 0.91– 0.96) is superior to that of deep learning architectures; (ii) Data quality, pre-processing rigor, feature engineering, and class imbalance correction are as important as algorithm selection to model performance; (iii) Regional contextualisation of ML models has a significant impact on the predictive validity within the specific populations as highlighted in the Punjab case study; and (iv) Explainability frameworks like SHAP are critical to transform model outputs into clinician comprehensible risk narratives when translating to clinical application, and can be used to build responsible ML-driven clinical systems.

The Punjab case study is the first diabetes prediction analysis based on the ML dataset, which includes a regional multi-district EHR dataset from Punjab, India. The performance of the Random Forest model (89.4% accuracy, AUC 0.93) and the identification of HbA1c, fasting glucose, BMI and family history as primary predictors are consistent with the global evidence as well as clinical experience, thus affirming the face validity of the model and its potential for clinical integration.

Finally, diabetes prediction based on machine learning is a well-established and well-validated field of technology which can be integrated clinically in phases. The key to success will be interdisciplinary partnerships between computer scientists, clinicians, health informaticists, and policy makers that will help address the technical, regulatory, ethical, and organizational hurdles of responsible AI use in healthcare.

8. FUTURE WORK

The gaps identified from this review are proposed as future research directions:

8.1 FEDERATED LEARNING FOR PRIVACY-PRESERVING MULTIINSTITUTIONAL MODELLING

One of the major challenges of developing large representative diabetes prediction models in India is the lack of integration of clinical data from institutional silos, adding to the patient privacy regulations (Digital Personal Data Protection Act, 2023). However, federated learning (FL), a technical approach that allows collaborative model training without data sharing among different sites, is the most promising answer to the problem. In future, federated Random Forest and federated gradient boosting models should be explored which address the heterogeneous nature of data quality and intermittent connectivity of district health systems in India, using the edge computing infrastructure being rolled out as part of the PM-WANI and BharatNet broadband initiatives.

8.2 EXPLAINABLE AI (XAI) AND CLINICIAN-AI INTERACTION DESIGN

While both SHAP and LIME are strong post-hoc techniques to explain, they still have some technical complexity for health workers on the front lines. Further research is needed to create clinician-tailored explanation interfaces, such as natural language rationale generation, risk factor traffic-light visualization, and counterfactual explanations (“If this patient had reduced their BMI by 2 kg/m², the predicted risk would be reduced from High to Moderate”) that can be evaluated using co-design sessions with general practitioners and ASHA (Accredited Social Health Activist) workers. Group fairness auditing, a systematic assessment of the performance gap between groups (caste, gender, and socioeconomic status) should also be considered part of XAI frameworks.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15646

8.3 MULTIMODAL DATA INTEGRATION

Existing prediction models for diabetes mostly rely on clinical and demographic data in tabular form. Incorporation of multimodal data streams, such as: wearable sensor data (CGM, accelerometery, PWT from smartbands); retinal fundus imagery (DR as a biomarker of systemic disease); genomic biomarkers (polygenic risk scores for T2DM); gut microbiome profiles; and social determinants of health (census and geospatial data). Multi-modal fusion architectures, such as transformer-based cross-attention mechanisms and multi-view learning frameworks provide a principled way of incorporating several types of data modalities and retaining explainability.

8.4 LONGITUDINAL AND TEMPORAL MODELLING

Diabetes is a progressive disease that has a clearly defined pre-diabetic phase (IFG and IGT) which comes 5-10 years before the diagnosis is made. This progression can only be captured with cross sectional modelling. Future studies should be able to build and test models based on LSTM and temporal graph neural network approaches using longitudinal cohorts of EHRs to support personalised prediction of the likelihood of clinical diabetes in a patient under various intervention scenarios, as well as the timing of the diabetes event.

8.5 TRANSFER LEARNING AND DOMAIN ADAPTATION

Transfer learning (pre-training on large global datasets, UK Biobank, NHANES, and finetuning on smaller regional datasets) is a significant step toward better model performance in data-limited settings in LMICs. The optimal transfer learning strategies and domain adaptation techniques to achieve effective fine-tuning for the Indian diabetes context, as well as the minimum number of labelled samples, should be empirically tested.

8.6 REAL-WORLD DEPLOYMENT AND VALIDATION STUDIES

Most studies to date are retrospective, and prospective clinical validation studies are greatly needed to evaluate model performance when deployed in real applications. Design and implementation of randomized controlled trials comparing the clinical outcomes (HbA1c reductions, early detection rates) of ML-guided diabetes screening to current practice, the patient outcomes (QALYs) and health system outcomes (cost per diabetes case averted) of each approach. AI-based medical device approval pathways with regulatory bodies, such as the Central Drugs Standard Control Organisation (CDSCO) and the Bureau of Indian Standards (BIS), should be explored for a successful clinical deployment of the AI medical device in a compliant way.

REFERENCES

[1]. American Diabetes Association. (2022). Standards of medical care in diabetes—2022. Diabetes Care, 45(Supplement 1), S1–S264. https://doi.org/10.2337/dc22-SINT
[2]. Ayon, S. I., & Islam, M. M. (2019). Diabetes prediction: A deep learning approach.
[3]. International Journal of Information Engineering and Electronic Business, 11(2), 21–
[4]. 27. https://doi.org/10.5815/ijieeb.2019.02.03
[5]. Basu, S., Flood, D., & Sharma, A. (2020). Epidemiology of type 2 diabetes in South Asia and implications for global health. The Lancet Diabetes & Endocrinology, 8(4), 274–283. https://doi.org/10.1016/S2213-8587(20)30065-X
[6]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
[7]. Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine, 346(6), 393–403. https://doi.org/10.1056/NEJMoa012512
[8]. Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465, 1–
[9]. 20. https://doi.org/10.1016/j.ins.2018.06.056
[10]. Federici, M., Gallo, A., & Andreozzi, F. (2022). Transformer-based deep learning for type 2 diabetes prediction: A multi-cohort study. npj Digital Medicine, 5(1), 88. https://doi.org/10.1038/s41746-022-00632-5
[11]. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review.
[12]. Diabetology & Metabolic Syndrome, 13(1), 148. https://doi.org/10.1186/s13098-021-00767-9
[13]. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). International Diabetes Federation. https://www.diabetesatlas.org

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15646

[14]. ICMR-INDIAB Collaborative Study Group. (2023). Diabetes mellitus in India: The ICMRINDIAB national cross-sectional study (Phase-II). The Lancet Diabetes & Endocrinology, 11(5), 355–368. https://doi.org/10.1016/S2213-8587(23)00119-5
[15]. Kaur, H., Kumari, V., & Vohra, S. (2020). Explainable machine learning for diabetes prediction using SHAP framework. Journal of Biomedical Informatics, 112, 103612.
[16]. https://doi.org/10.1016/j.jbi.2020.103612
[17]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017).
[18]. Machine learning and data mining methods in diabetes research. Computational and
[19]. Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
[20]. Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797– 1801.
[21]. Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2017). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 5(1), 1–14. https://doi.org/10.1007/s13755-019-0095-z
[22]. Misra, A., Jayawardena, R., & Anoop, S. (2019). Obesity in South Asia: Phenotype, morbidities, and mitigation. Current Obesity Reports, 8(1), 43–52. https://doi.org/10.1007/s13679-019-0328-0
[23]. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms.
[24]. Procedia Computer Science, 165,292–299. https://doi.org/10.1016/j.procs.2020.01.047
[25]. Rashid, M. (2020). Machine learning for predicting diabetes using hybrid CNN-LSTM architecture. Applied Intelligence, 50(10), 3449–3462. https://doi.org/10.1007/s10489-
[26]. 020-01725-8
[27]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms.
[28]. Procedia Computer Science, 132, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122
[29]. Tafa, Z., Pervetica, N., & Karahoda, B. (2015). An intelligent system for diabetes prediction. Proceedings of the 4th Mediterranean Conference on Embedded Computing, 1–4.
[30]. https://doi.org/10.1109/MECO.2015.7181925
[31]. Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706–716. https://doi.org/10.1016/j.procs.2020.03.336
[32]. World Health Organization. (2023). Global report on diabetes. World Health Organization. https://www.who.int/publications/i/item/9789241565257
[33]. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515. https://doi.org/10.3389/fgene.2018.00515

