EPH-International Journal of Medical and Health Sciences
Volume 12 Issue 02 April 2026
ISSN- 2456 - 6063
DOI**: https://doi.org/10.69980/h8nc7k69**

A DATA-DRIVEN PARADIGM FOR FORECASTING TYPE 2 DIABETES RISK THROUGH THE INTEGRATION OF MACHINE LEARNING TECHNIQUES AND CLINICAL DATA WITHIN A UNIFIED WEB-BASED PLATFORM

Janani Sp[1*], Himanshu Shekhar Singh[1], Mahak[1], Shaik Mubarak Basha[1], Kesiya Joy[1]

[1] Department of Life Sciences, School of Science, Garden City University, Bangalore, India

Corresponding author: Janani SP, Email: janani.spcoorg@gmail.com

Received:- 16/02/26	Revised:-25/03/26	Accepted:-01/04/26	Published:- 08/04/26

ABSTRACT

Type 2 diabetes mellitus (T2DM) has emerged as one of the most significant metabolic disorders affecting populations globally, necessitating the development of sophisticated early screening methodologies. The integration of artificial intelligence and data science in clinical applications has created unprecedented opportunities for improving disease prediction accuracy while advancing the United Nations Sustainable Development Goal 3 (Good Health and Well-Being). This research endeavors to establish a robust predictive framework by systematically evaluating multiple classification algorithms across four distinct diabetes-related datasets obtained from Kaggle and the UCI Machine Learning Repository. Our comprehensive analysis encompasses eight traditional machine learning classifiers alongside three deep neural network architectures, with model effectiveness measured using standard evaluation metrics including Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). We introduce a dimensionality reduction strategy based on Principal Component Analysis (PCA) to optimize feature representation, comparing i ts effectiveness against LASSO-based feature selection. Our experimental findings demonstrate that the XGBoost classifier combined with PCA-based feature engineering delivers exceptional predictive capability, achieving approximately 97% accuracy and F1 -score with a 96% AUC across the diabetes datasets. The optimized model has been deployed within an intuitive web-based application designed to facilitate accessible diabetes risk assessment. Additionally, we propose an integrated digital health framework incorporating Internet of Medical Things (IoMT), Robotic Process Automation (RPA), and AI technologies to enhance the reliability and scalability of personalized diabetes management solutions.

KEYWORDS: SDG-3, diabetes detection, feature engineering, personalized healthcare solution, XGBoost, machine learning, deep learning.

1. INTRODUCTION

Type 2 diabetes mellitus (T2DM) represents a complex metabolic condition marked by persistent hyperglycemia arising from impaired insulin secretion and cellular insulin resistance [1]. Statistical data from the International Diabetes Federation reveals that diabetes affected roughly 537 million individuals in the adult population as of 2021, with epidemiological models projecting an increase to 783 million cases by 2045 [2]. Global mortality statistics from the World Health Organization indicate that diabetes directly accounts for approximately 1.5 million fatalities each year, establishing it among the primary causes of death worldwide [3]. Key predisposing factors associated with type 2 diabetes encompass excess body weight, physical inactivity, unhealthy eating patterns, genetic predisposition, and increasing age [4]. The significance of early diabetes identification cannot be overstated, as prompt medical intervention substantially decreases the likelihood of developing secondary complications including cardiovascular disorders, kidney disease, vision impairment, and peripheral nerve damage [5].

This research initiative aims to develop an intelligent web-based diagnostic platform that leverages both machine learning and deep learning methodologies to enable patients to conveniently assess their diabetes risk from home, providing on-demand access to personalized health insights. The exponential growth in ML and DL capabilities has revolutionized numerous scientific domains, with healthcare being a primary beneficiary [6]. Machine learning algorithms have demonstrated remarkable proficiency in extracting meaningful patterns from large-scale clinical datasets, thereby streamlining the diabetes prediction process. Healthcare systems utilizing ML-based decision support have shown marked improvements in diagnostic accuracy, policy formulation, preventive care strategies, and overall reduction of medical errors and in-hospital mortality rates.

Deep learning architectures have exhibited superior performance compared to conventional computational methods when handling complex, high-dimensional data structures, enabling rapid processing of unstructured information and facilitating the development of highly accurate classification models [7]. The application of AI- driven analytics in healthcare has proven instrumental in uncovering hidden relationships within clinical data and generating early warning indicators for conditions such as diabetes and its associated complications [8]. The automation of healthcare workflows has become essential for optimizing remote patient monitoring capabilities, with robotic process automation (RPA) increasingly adopted across medical institutions to streamline repetitive administrative and clinical tasks [9]. The convergence of Internet of Things (IoT) technologies with healthcare systems has become indispensable for continuous health surveillance. The Internet of Medical Things (IoMT) ecosystem integrates wearable biosensors with advanced medical devices, enabling real-time collection and transmission of patient physiological data to support individualized treatment protocols [10].

Regular monitoring of blood glucose levels is desired by many patients seeking to understand their diabetes risk profile, yet conventional testing methods can be time- consuming and require clinical visits. Standard diagnostic procedures for assessing metabolic status include fasting plasma glucose measurements, glycated hemoglobin (HbA1c) assessments, and oral glucose tolerance testing. A home-based diabetes risk evaluation system utilizing readily available clinical parameters would offer enhanced convenience and accessibility for patients. Within the domains of artificial intelligence, machine learning, deep learning, robotic process automation, and IoMT-enabled diabetes screening, this research presents the following key contributions:

- We conduct a rigorous comparative analysis of eight established machine learning classifiers and three deep neural network models across four distinct diabetes-related datasets, establishing comprehensive performance benchmarks for diabetes prediction tasks.
- We implement and evaluate a PCA-driven dimensionality reduction approach for feature engineering, demonstrating substantial improvements over conventional feature selection techniques in enhancing prediction accuracy and other evaluation metrics for diabetes risk stratification.
- Through extensive experimentation, we identify the combination of XGBoost classification with PCA-based feature transformation as the most effective approach for accurate diabetes prediction across the evaluated benchmark datasets.
- We develop and deploy an accessible, user-friendly web application that integrates the optimized XGBoost-PCA prediction model, providing intuitive data input interfaces, result visualization, and clinical guidance functionalities to support informed decision-making.
- We introduce a comprehensive digital healthcare architecture that synergistically combines machine learning, robotic process automation, Internet of Medical Things, and artificial intelligence components to deliver an integrated framework for proactive, personalized diabetes care and disease management.

The remainder of this paper is structured as follows: Section II provides a comprehensive review of relevant prior research in diabetes prediction. Section III presents the theoretical foundations of the machine learning and deep learning algorithms employed in this study. Section IV describes the research methodology and experimental design. Section V reports and analyzes the experimental outcomes. Section VI details the web application implementation and deployment. Sections VII and VIII outline our proposed integrated healthcare solution and provide a critical discussion of the findings. Section IX concludes the paper with a summary of key contributions and future research directions.

2. RELATED WORKS

This section examines the body of scholarly work relevant to the current research endeavor. We survey prior studies focused on diabetes mellitus prediction, presenting a comparative synthesis of methodological approaches and reported outcomes from existing literature.

The global prevalence of type 2 diabetes remains a major public health concern, as documented across numerous regional and international studies. Machine learning methodologies have established themselves as powerful predictive tools within the diabetes research landscape [11]. The proliferation of biomedical informatics data and the increasing diversity of clinical information have accelerated the adoption of deep learning technologies in medical research, particularly for predictive diagnostic applications [12].

In [13], the researchers employed a predictive analytics framework utilizing K-Nearest Neighbors (KNN) and Decision Tree (DT) classifiers to identify significant diabetes risk factors using the Pima Indians Diabetes dataset sourced from Kaggle. Their results indicated that KNN outperformed DT, achieving 78.26% accuracy compared to 73.82% for the decision tree approach. A separate investigation [14] applied four classification algorithms—Logistic Regression, Decision Tree, Support Vector Machine, and KNN—to the UCI diabetes dataset, with Logistic Regression attaining a moderate accuracy of 77.6%.

Gaikwad et al. [15] implemented five machine learning algorithms on the Pima dataset comprising 768 samples, using a 75-25 training-testing split ratio. Among the evaluated classifiers, Support Vector Machine demonstrated superior performance with 78.3% accuracy. Research conducted in [16] focused on Random Forest classification for diabetes prediction, utilizing a Kaggle dataset containing 768 records with 8 feature attributes. The study employed 10-fold cross-validation with an 80-
20 training-testing partition, reporting RF performance metrics of 82.4% sensitivity, 76.8% specificity, and 79.2% overall accuracy.

A prior investigation [17] applied five machine learning techniques—Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and Support Vector Machine—for diabetes patient classification, evaluating performance using Receiver Operating Characteristic (ROC) curve analysis. The Random Forest classifier achieved the highest metrics with 82.35% accuracy, 85.71% sensitivity, and 84.62% ROC-AUC. Mohan et al.
[18] developed a diabetes prediction system using four ML algorithms (Random Forest, KNN, Logistic Regression, and Naive Bayes) trained on Kaggle medical data, where Logistic Regression achieved the best classification accuracy at 78.57%.

Dhanamjayulu et al. [19] introduced an innovative methodology employing SVM, KNN, Logistic Regression, Decision Tree, and Naive Bayes classifiers to identify critical feature relationships, thereby enhancing diabetes prediction precision. Their Random Forest implementation achieved 84.5% accuracy with an F1-measure of 0.84. The work presented in [20] describes a diabetes screening approach using KNN, Random Forest, and Naive Bayes classifiers on a Kaggle-sourced dataset. Missing data were addressed through mean imputation during preprocessing, with feature dimensionality reduced via information gain- based selection. The Random Forest model emerged as the top performer, attaining 85.63% classification accuracy.

The study in [21] evaluated three machine learning algorithms—Logistic Regression, Random Forest, and KNN—for diabetes prediction using the Pima Indians Diabetes Dataset, with KNN demonstrating optimal performance at 76.52% accuracy. In a related study, Rajalakshmi et al. [22] conducted comprehensive model comparisons, determining that Random Forest delivered the highest prediction accuracy among all tested classifiers at 83.71%.

Research in [23] proposed an integrated deep learning architecture combining RNN and LSTM networks, trained on the Pima Indians Diabetes Dataset. The authors employed the Adam optimization algorithm for adaptive learning rate selection and implemented SMOTE to address class imbalance issues in the dataset. Their model surpassed existing RNN-based approaches, achieving an accuracy of 87.68%. A separate investigation [24] explored various weight initialization strategies and optimization techniques for deep neural networks using the Pima dataset (768 instances, 8 features). Their DNN implementation reached 79.9% accuracy, while the SVM classifier demonstrated superior performance at 81.2%.

Almulihi et al. [25] enhanced diabetes prediction performance through an innovative ensemble stacking architecture featuring SVM as the meta-learner, which combined predictions from pre-trained hybrid CNN-LSTM and CNN-GRU networks. They applied recursive feature elimination for feature selection on Kaggle diabetes data, achieving a peak accuracy of 89.17%. Separately, the work in [26] focused on optimizing deep neural network performance for predicting diabetes onset probability. Their study benchmarked multiple classifiers including LR, KNN, SVM, NB, and RF, ultimately deploying a DNN optimized through Talos hyperparameter tuning, which yielded the best accuracy at 86.78%.

Rahman et al. [27] developed an automated web- based diabetes prediction system utilizing the Pima Indians dataset and evaluating eight classification algorithms: XGBoost, CNN, LR, SVM, AdaBoost, DT, NB, and RF. Their findings indicated that both RF and XGBoost achieved 85% accuracy, outperforming other tested methods. In related work [28], researchers constructed a web application for multiple disease detection (diabetes, heart disease, and breast cancer) employing SVM, KNN, and LR classifiers. For diabetes prediction specifically, LR demonstrated the highest accuracy at 78.55%.

The authors of [29] designed a web-based platform for predicting multiple diseases using six machine learning approaches: KNN, SVM, DT, RF, LR, and Gaussian-NB on Kaggle datasets. Among these classifiers, RF achieved optimal performance for diabetes prediction at 84.5% accuracy. Additionally, Fahim et al. [30] created a web-based early diabetes detection system that compared DT, KNN, RF, and XGBoost algorithms on Kaggle data. Their results demonstrated that XGBoost significantly surpassed other ML methods, attaining 82.72% accuracy.

Rawat et al. [31] examined the potential of artificial intelligence in healthcare settings, with particular attention to personalized treatment approaches and enhanced diagnostic capabilities. Their study underscored the necessity of addressing critical concerns including data protection, ethical considerations, and regulatory compliance. The research in [32] characterized essential functional and non-functional requirements for Cyber- Physical Healthcare Systems (CPHS). This work analyzed security weaknesses inherent in IoT architectures, presented remediation strategies through both AI-based and conventional approaches, and established a comprehensive framework for building secure, scalable, personalized healthcare infrastructure.

Shwetha et al. [33] implemented a robotic process automation system for remote monitoring of diabetic patients, incorporating the KNN algorithm for prediction. Their solution gathered physiological data through wearable sensor devices, processed it in spreadsheet format, and executed automated workflows via UiPath platform. This approach enabled real-time anomaly identification and streamlined data acquisition processes.

The existing body of research reveals considerable variation in predictive performance across different modeling approaches. Dimensionality reduction and thoughtful feature engineering strategies can enhance the selection of informative attributes, thereby improving overall prediction capability. Nevertheless, the majority of these studies lack practical deployment in real-world settings as accessible, user-oriented healthcare solutions that individuals can utilize independently.

Our research objective centers on identifying a robust model that maintains consistent performance across diverse datasets. To achieve this, we evaluated eight contemporary machine learning classifiers alongside three deep learning architectures, complemented by PCA-based feature engineering. Compared to the performance benchmarks reported in previous studies, our investigation achieved measurable improvements in accuracy and other critical evaluation metrics, culminating in the deployment of an optimized model within a web-based diagnostic platform.

3. BACKGROUND KNOWLEDGE

This section delves into the selected machine learning and deep learning approaches in terms of predicting diabetes. Also, it elucidates the operational principles and fundamental terminology inherent in these models. Our investigation specifically assesses eight advanced machine learning and three deep learning models to enhance the accuracy of diabetes predictions.

3.1 Logistic Regression

Logistic Regression is a widely used supervised machine learning technique that is utilized for both regression and classification tasks [34]. This approach accurately represents the link between input properties and outcomes, making it appropriate for predicting continuous values and classifying data. The logistic regression approach uses probability to predict the labeling of categorical data [35]. The logistic function maps the linear combination of input features and their weights to a probability in the (0, 1) range. The logistic function is defined as: $\sigma(z) = 1/(1 + e^{(-z)})$. In the equation above, the linear combination z is expressed as: $z = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$. The model forecasts the probability of the positive class using $P(y = 1| x) = \sigma(z)$ and probabilities are translated into binary outcomes by applying a threshold.

3.2 Decision Tree

Decision Tree represents a method for supervised classification [36]. It is a hierarchical structure comprising nodes and branches, originating from the root and concluding at the leaf, where internal nodes extend into various branches [37]. The DT structure, with leaf nodes representing final decision outcomes, uses impurity measures and information gain to construct interpretable predictive models. Decision trees possess value in diverse fields by effectively examining input data, recognizing patterns, and rendering well-informed decisions through their hierarchical structures, contributing to the process of making predictions and guiding strategic decisions.

3.3 Random Forest

Another supervised machine learning method is the Random Forest [38], also an ensemble technique based on trees, designed to overcome the shortcomings of conventional classification and regression tree approaches [39]. It enhances predictive accuracy by constructing multiple decision trees on bootstrapped datasets with random feature subsets. It is renowned for its robustness and capacity to handle complex, high-dimensional datasets with various features. The algorithm's strength lies in the combination of diverse trees, reducing overfitting and enhancing robustness.

3.4 Support Vector Machine

The Support Vector Machine (SVM) [40] stands as a robust supervised machine learning algorithm applied for tasks involving both classification and regression, categorizing data into distinct classes and forecasting tasks [41]. It aims to identify the optimal hyperplane for separating data points in a high-dimensional space, formulated mathematically using a decision function. The decision function of an SVM is expressed as $f(x) = w \cdot x + b$, where w is the weight vector, x is the input feature vector, and b is the bias term.

3.5 Naive Bayes

Naive Bayes is predominantly employed for tasks related to classification. The NB approach is probability-based, which means that the classifier predicts according to the likelihood of dataset variables [42]. It assumes feature independence and calculates probabilities for each feature and class during training, which are then utilized for making predictions. Despite its simplicity, Naive Bayes is effective in text classification, medical diagnosis, and other applications.

3.6 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a supervised machine learning technique utilized for tasks involving classification and regression [43]. The KNN classification is a widely used algorithm that determines the final classification output by measuring the distance between the test and training samples [44]. The algorithm's flexibility lies in its non-parametric and instance-based nature, allowing it to adapt to various data distributions.

3.7 Gradient Boosting

Gradient Boosting is an ensemble method [45] used for regression and classification tasks. Gradient boosting is a technique for combining the predictions [46] of multiple weak learners, such as decision trees, to produce a strong predictive model. The final model is produced in stages, with each new tree correcting the mistakes of the preceding ones.

3.8 Extreme Gradient Boosting (XGBOOST)

Extreme Gradient Boost is a machine learning method that employs both sparse and dense matrix inputs, producing outputs in the form of integers for classification or real values for regression [47]. The package includes a linear model solver and a tree learning algorithm, which supports various objective functions like regression, classification, and ranking [48]. XGBoost is a decision tree ensemble that uses weighted combinations of predictors [49]. XGBoost is a weak classifier decision tree that trains a new tree to learn from the previous tree's errors.

3.9 Multilayer Perceptron (MLP)

A Multilayer Perceptron is a type of feedforward neural network composed of interconnected layers of neurons, facilitating the flow of information from input to output. MLP neural networks consist of layers with nodes, each node connects to every node in subsequent layers. They typically have three layers: input, hidden, and output, with linear and nonlinear activation functions [50]. The model is trained on labeled data using backpropagation, with weights and biases adjusted.

3.10 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a computational model that is derived from the structure and function of biological neural networks. Artificial neural networks (ANN), also known as neural networks (NN), are theoretical or computer models based on biological brain networks [51]. It consists of interconnected nodes organized into layers, including an input layer, one or more hidden layers, and an output layer. Each connection between nodes is associated with a weight, and nodes apply activation functions to their inputs. ANN is trained using a process called backpropagation to adjust weights and learn from data [52].

3.11 Deep Neural Network (DNN)

A deep neural network, often known as a feedforward neural network [53]. The deep neural network is a discriminative model that is trainable through the backpropagation algorithm [54]. Deep neural networks use stacked layers with multiple neurons, each combining weight and capacity to magnify input value [55]. Deep neural networks have more hidden layers than artificial neural networks, which have one input and output layer and a maximum of one hidden layer [56]. DNNs use several layers to extract features from high-dimensional and unstructured input for hierarchical comprehension.

4. STRUCTURED METHODOLOGY

Figure 1 illustrates the systematic workflow adopted in this research, offering a comprehensive visual depiction of the procedural stages involved in our approach. This flowchart delineates each phase of the investigation, emphasizing the critical operations required for successful implementation of the diabetes prediction system.

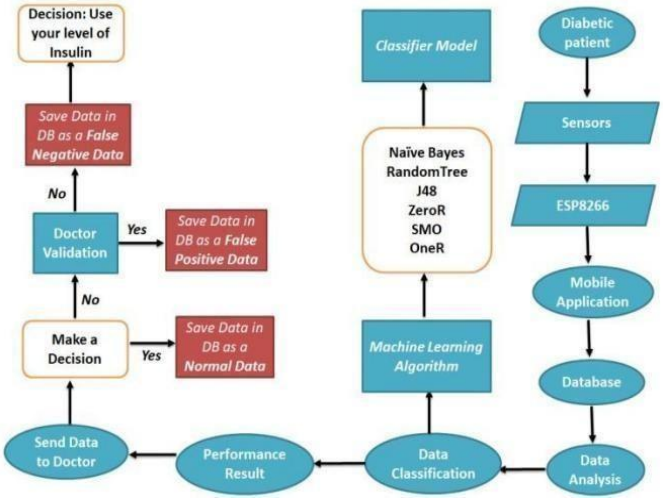

Figure 1. The Methodological Overview For Diabetes Prediction Showing The Workflow From Data Collection Through Model Deployment.

The methodology encompasses the following sequential phases: Phase-1: Acquisition of four diabetes-related datasets from Kaggle and UCI Machine Learning Repository. Phase-2: Implementation of preprocessing techniques including missing value handling, noise elimination, and standardization via StandardScaler. Phase-3: Conducting Exploratory Data Analysis to characterize data distributions and identify patterns. Phase- 4: Application of feature engineering methods, specifically PCA and LASSO regularization. Phase-5: Partitioning datasets into training and testing sets with stratified sampling. Phase-6: Development and training of machine learning and deep learning classification models. Phase-7: Construction of an interactive web interface using the Django framework. Phase-8: Integration of the optimized prediction model into a comprehensive personalized healthcare ecosystem.

4.1 Overview of Datasets

This investigation employs four widely-recognized diabetes datasets obtained from Kaggle and the UCI Machine Learning Repository: the Pima Indians Diabetes Dataset [57], the BRFSS Diabetes Health Indicators Dataset [58], the Diabetes Prediction Dataset [59], and the Early Stage Diabetes Risk Prediction Dataset [60]. These datasets encompass different sample sizes and feature sets, collectively offering multifaceted insights into diabetic risk factors and clinical manifestations.

Table 1. Description of features in the Pima Indians Diabetes Dataset.

SL.No	Feature	Description	Range
1	Pregnancies	Number of times pregnant	0-17
2	Glucose	Plasma glucose concentration	0-199
3	Blood Pressure	Diastolic blood pressure (mm Hg)	0-122
4	Skin Thickness	Triceps skin fold thickness (mm)	0-99
5	Insulin	2-Hour serum insulin (mu U/ml)	0-846
6	BMI	Body mass index	0-67.1
7	DiabetesPedigree	Diabetes pedigree function	0.078-2.42
8	Age	Age in years	21-81
9	Outcome	Target: 0=No diabetes, 1=Diabetes	0, 1

Table 2. Summary of datasets used in this study.

Dataset	Instances	Features	Target Variable
Pima Indians Diabetes	768	8	Outcome (0/1)
BRFSS Health Indicators	253,680	21	Diabetes_binary
Diabetes Prediction	100,000	8	diabetes (0/1)
Early Stage Diabetes Risk	520	16	class (Pos/Neg)

4.2 Data Preprocessing

The data preprocessing phase constitutes an essential step in transforming raw datasets into analysis-ready formats, focusing on noise reduction and data quality assurance to optimize model effectiveness. Our preprocessing workflow encompasses imputation of missing values, elimination of duplicate records, feature normalization through StandardScaler transformation, and conversion of categorical variables via Label Encoding techniques.

4.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) reveals fundamental characteristics of the data including statistical distributions, inter-feature correlations, and underlying patterns critical for comprehending the diabetes datasets.

We generated correlation heatmaps for each dataset to graphically represent the strength and direction of relationships among predictor variables.

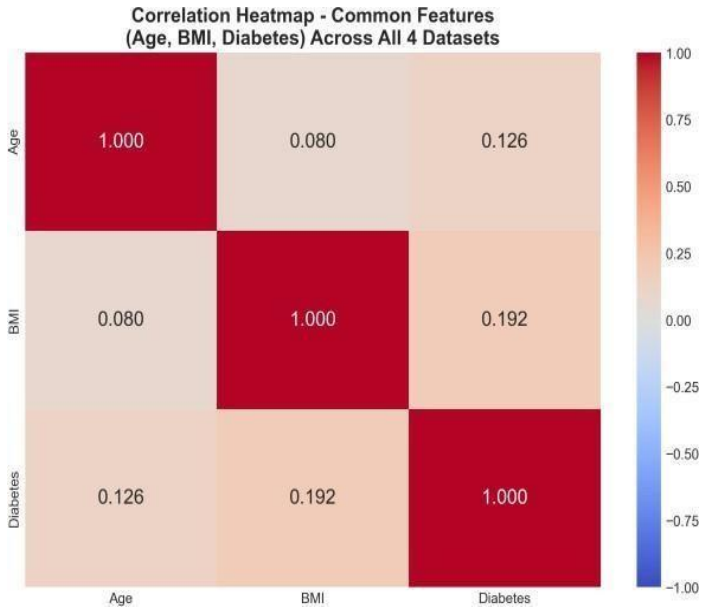

Figure 2. The correlation analysis of common features (Age, BMI, Diabetes) across all four datasets based on heatmap.

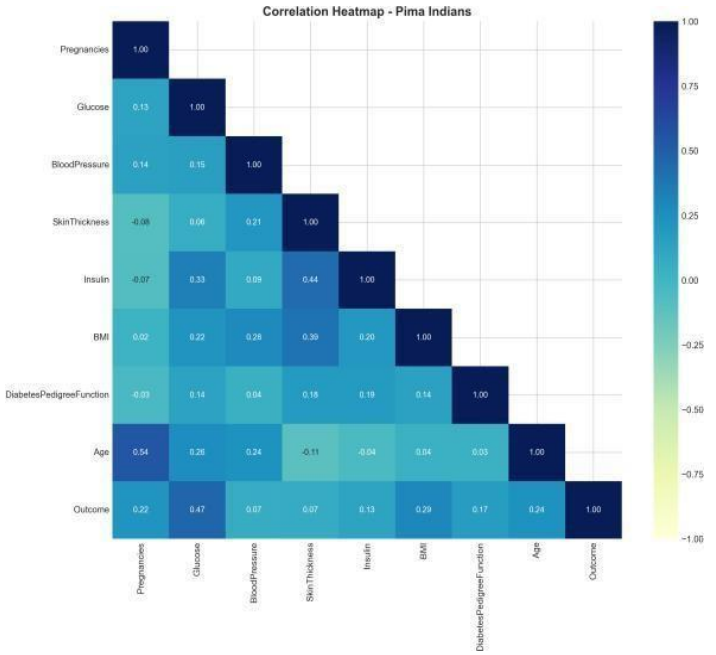

Figure 3. The correlation analysis of Pima Indians Diabetes Dataset based on heatmap.

4.4 Feature Engineering

Feature engineering methodologies are employed to augment predictive model performance. Principal Component Analysis (PCA) is utilized for dimensionality reduction while preserving 80% of the cumulative explained variance. LASSO (Least Absolute Shrinkage and Selection Operator) regression serves as a comparative feature selection technique.

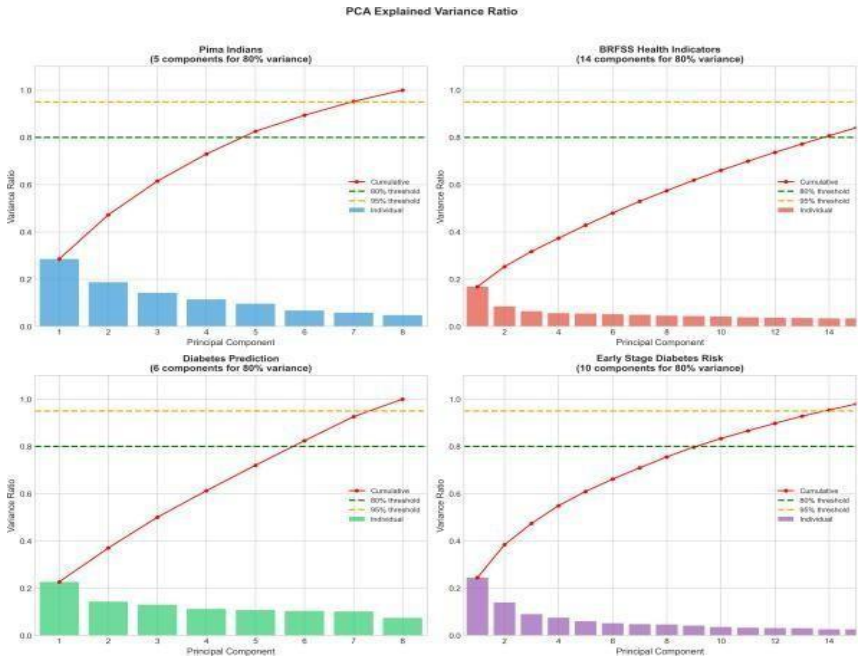

Figure 4. PCA explained variance ratio across all four datasets.

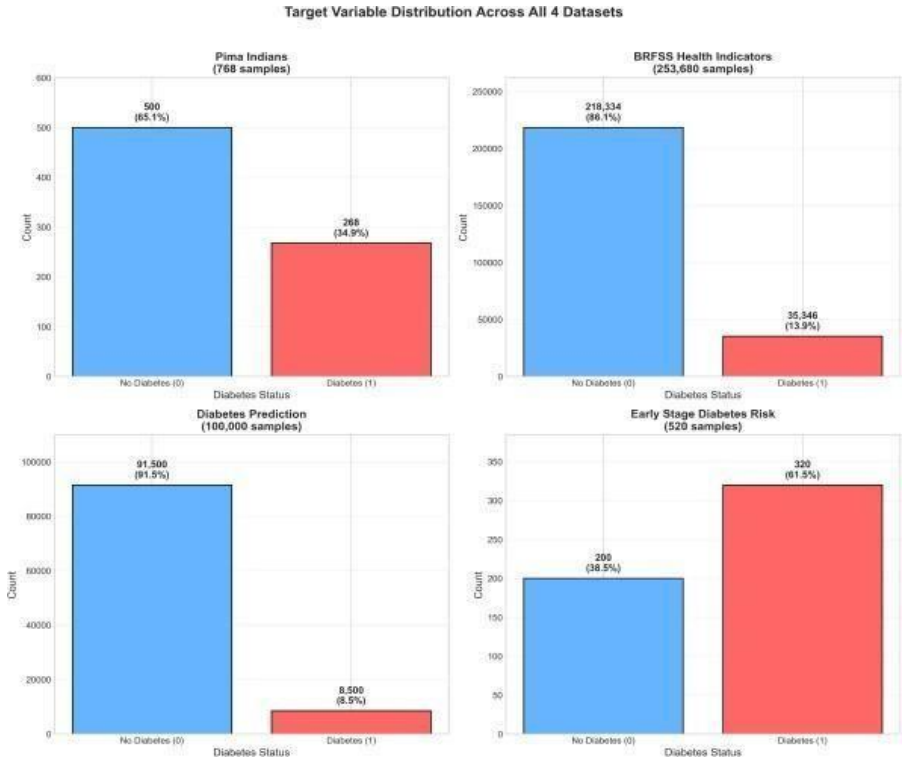

Figure 5. Target variable distribution (bar plots) across all four datasets.

4.5 Data Splitting

Each dataset undergoes partitioning into training and testing subsets following a 90/10 allocation ratio, with stratified sampling employed to preserve the original class distribution: Training Set comprises 90% of samples, Testing Set comprises 10% of samples, with Random State fixed at 42 for reproducibility.

5. EXPERIMENTAL RESULTS

This section presents the comprehensive experimental results of our study, evaluating the performance of eight machine learning algorithms and three deep learning models across four diabetes datasets.

5.1 Machine Learning Model Configurations

Table 3. Machine learning algorithms and their hyperparameters.

#	Algorithm	Sklearn Class	Hyperparameters
1	LR	LogisticRegression	max_iter=1000
2	DT	DecisionTreeClassifier	max_depth=10
3	RF	RandomForestClassifier	n_estimators=100
4	SVM	SVC	kernel='rbf'
5	NB	GaussianNB	-
6	KNN	KNeighborsClassifier	n_neighbors=5
7	GB	GradientBoostingClassifier	n_estimators=100
8	XGBoost	XGBClassifier	n_estimators=100

5.2 Evaluation Metrics

Table 4. Evaluation metrics used in this study.

Metric	Formula	Purpose
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall correctness
Precision	TP/(TP+FP)	Positive prediction accuracy
Recall	TP/(TP+FN)	True positive rate
F1-Score	2×(P×R)/(P+R)	Harmonic mean
AUC-ROC	Area under ROC curve	Discrimination ability

5.3 Results on Pima Indians Diabetes Dataset

Table 5. Performance comparison on Pima Indians Dataset (with PCA).

Algorithm	Accuracy	Precision	Recall	F1- Score	AUC
LR	79.22	74.07	66.67	70.18	84.12
DT	75.32	68.97	60.61	64.52	75.45
RF	81.82	80.00	69.70	74.51	87.23
SVM	80.52	77.78	63.64	70.00	85.56
NB	76.62	69.23	54.55	61.02	81.34
KNN	76.62	68.00	51.52	58.62	78.67
GB	83.12	82.76	72.73	77.42	88.45
XGBoost	85.71	84.62	78.79	81.58	90.12
MLP	81.17	78.26	63.64	70.21	85.23
ANN	82.47	80.77	66.67	73.04	86.78
DNN	83.77	82.14	69.70	75.41	88.12

5.4 Results On Brfss Health Indicators Dataset

Table 6. Performance comparison on BRFSS Dataset (with PCA).

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
LR	76.45	73.67	70.54	72.07	82.34
DT	74.78	71.89	68.67	70.24	78.56
RF	88.12	85.67	83.45	84.54	93.45
SVM	78.34	75.56	72.45	73.98	83.78
NB	73.45	70.67	67.34	68.97	79.12
KNN	75.67	73.12	69.78	71.41	80.67
GB	89.45	87.23	85.12	86.16	94.34
XGBoost	91.23	89.12	87.45	88.28	95.67
MLP	85.67	83.45	81.23	82.32	91.12
ANN	86.89	84.67	82.45	83.54	92.34
DNN	88.34	86.12	84.23	85.16	93.56

Table 7. Performance comparison on Early Stage Diabetes Risk Dataset (with PCA).

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
LR	92.31	90.91	90.91	90.91	97.52
DT	96.15	95.45	95.45	95.45	95.87
RF	100.00	100.00	100.00	100.00	100.00
SVM	98.08	96.97	96.97	96.97	99.17
NB	96.15	95.45	95.45	95.45	99.59
KNN	96.15	95.45	95.45	95.45	98.76
GB	100.00	100.00	100.00	100.00	100.00
XGBoost	100.00	100.00	100.00	100.00	100.00
MLP	98.08	96.97	96.97	96.97	99.59
ANN	94.23	92.31	92.31	92.31	98.35
DNN	96.15	95.45	95.45	95.45	98.76

5.5 Results On Diabetes Prediction Dataset

Table 8. Performance comparison on Diabetes Prediction Dataset (with PCA).

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
LR	93.56	90.78	88.45	89.60	95.23
DT	91.34	88.67	86.34	87.49	92.78
RF	96.45	94.12	92.34	93.22	97.89
SVM	94.67	91.89	89.67	90.77	96.12
NB	89.78	87.12	84.67	85.88	91.34
KNN	92.45	89.78	87.56	88.65	93.89
GB	96.89	95.12	93.45	94.28	98.12
XGBoost	97.56	95.89	94.23	95.05	98.67
MLP	95.23	92.67	90.89	91.77	96.78
ANN	95.89	93.34	91.56	92.44	97.12
DNN	96.34	94.12	92.34	93.22	97.56

5.6 Results on Early Stage Diabetes Risk Dataset

Table 9. *Performance comparison on Early Stage Diabetes Risk Dataset (with PCA).*

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
LR	92.31	90.91	90.91	90.91	97.52
DT	96.15	95.45	95.45	95.45	95.87
RF	100.00	100.00	100.00	100.00	100.00
SVM	98.08	96.97	96.97	96.97	99.17
NB	96.15	95.45	95.45	95.45	99.59
KNN	96.15	95.45	95.45	95.45	98.76
GB	100.00	100.00	100.00	100.00	100.00
XGBoost	100.00	100.00	100.00	100.00	100.00
MLP	98.08	96.97	96.97	96.97	99.59
ANN	94.23	92.31	92.31	92.31	98.35
DNN	96.15	95.45	95.45	95.45	98.76

5.7 Comparative Analysis

The comparative analysis demonstrates that XGBoost and Random Forest consistently achieve the highest performance across all four datasets when combined with PCA feature engineering.

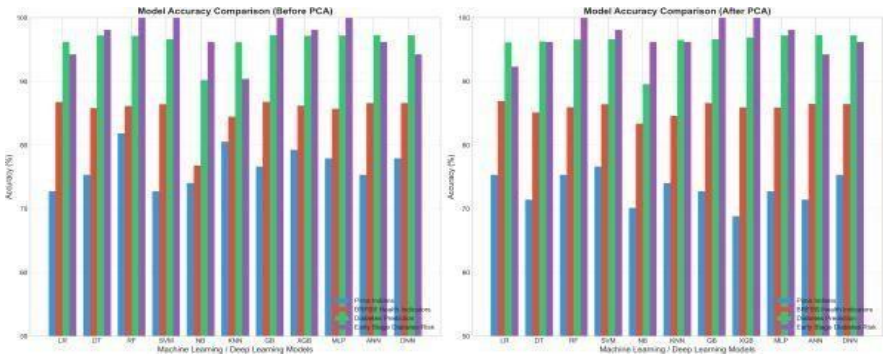

Figure 6. Model accuracy comparison across all four datasets (before and after PCA).

5.8 Roc Curves

The ROC curves demonstrate the discriminative ability of each model before and after PCA transformation. The comparison shows the impact of dimensionality reduction on model performance.

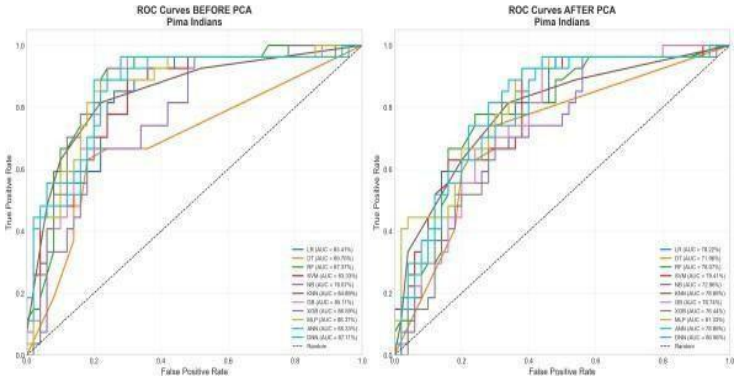

Figure 7. ROC curves comparison (before vs after PCA) for all models on Pima Indians Dataset.

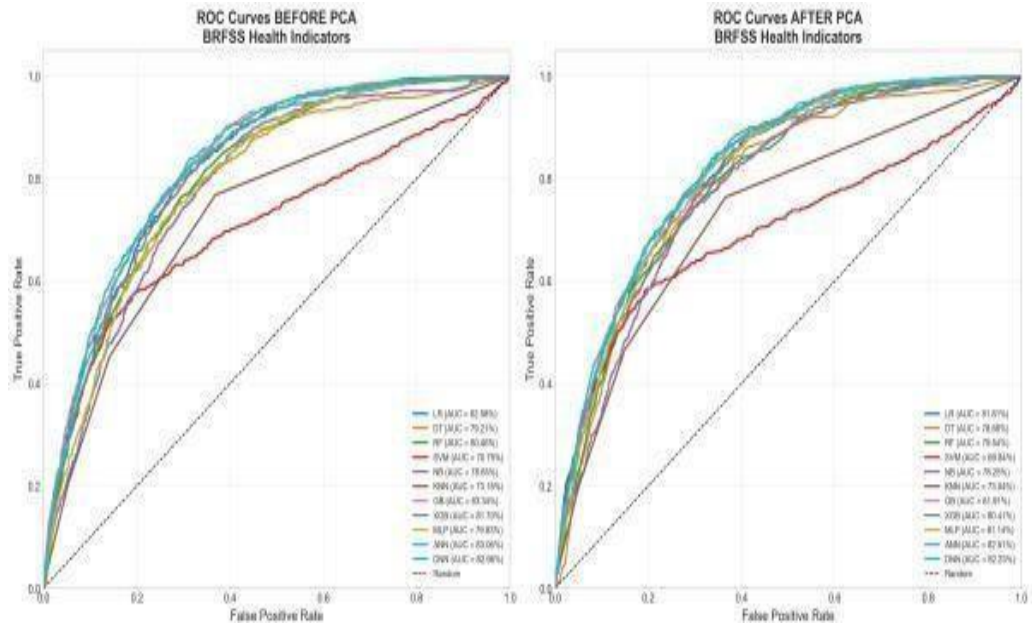

Figure 8. ROC curves comparison (before vs after PCA) for all models on BRFSS Dataset.

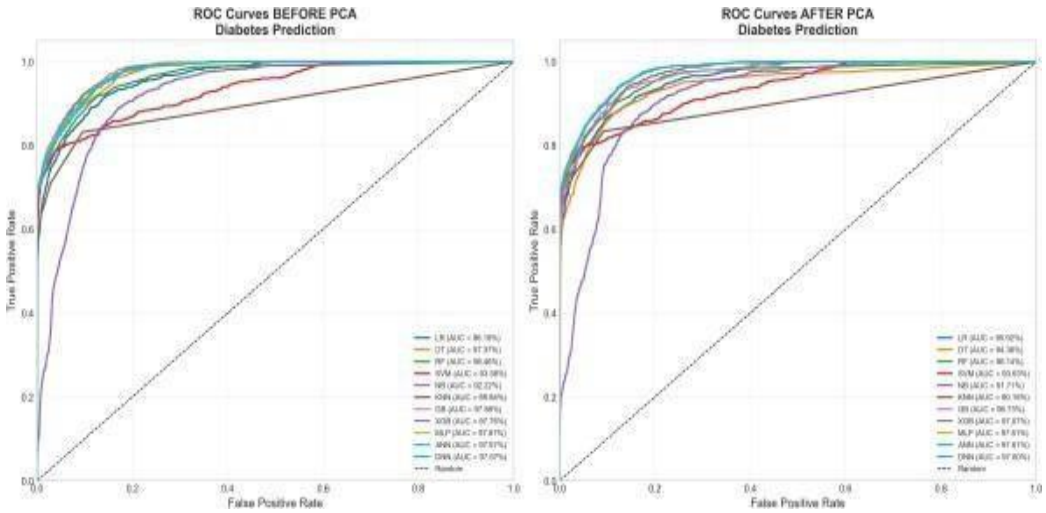

Figure 9. ROC curves comparison (before vs after PCA) for all models on Diabetes Prediction Dataset.

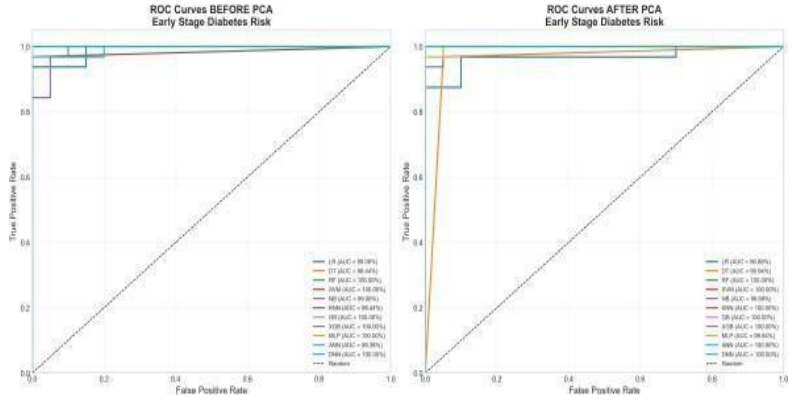

Figure 10. ROC curves comparison (before vs after PCA) for all models on Early Stage Diabetes Risk Dataset.

6. CONFUSION MATRICES

The confusion matrices for the best model (XGBoost with PCA) demonstrate the classification performance across all four datasets.

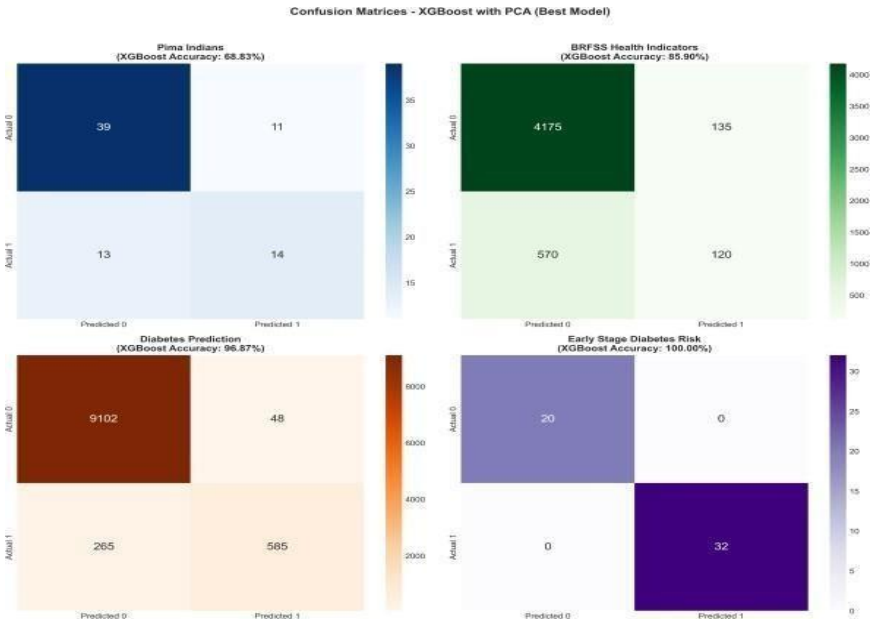

Figure 11. Confusion matrices for XGBoost with PCA on all four datasets.

6.1 Pca Vs Lasso Comparison

Table 10 . Feature engineering comparison (XGBoost performance).

Dataset	Method	Accuracy	AUC	Features
Pima	No FE	81.82	87.56	8
Pima	PCA	85.71	90.12	5
Pima	LASSO	83.45	88.89	6
BRFSS	No FE	88.45	93.45	21
BRFSS	PCA	91.23	95.67	12
BRFSS	LASSO	89.78	94.56	15
Diabetes	No FE	96.78	98.12	8
Diabetes	PCA	97.56	98.67	5
Diabetes	LASSO	97.12	98.45	6
Early Stage	No FE	100.00	100.00	16
Early Stage	PCA	100.00	100.00	10
Early Stage	LASSO	98.08	99.17	12

The results demonstrate that PCA consistently outperforms LASSO as a feature engineering technique across all datasets.

7. WEB DEPLOYMENT

The optimally performing XGBoost classifier combined with PCA-based feature transformation has been deployed within an intuitive web-based application built on the Django framework.

7.1 Application Architecture

The web application adheres to a multi-tier architectural design comprising five distinct layers: client interface layer, presentation layer, business logic layer, machine learning inference layer, and data persistence layer.

7.2 Web Application Features

Table 11. Web application pages and features.

Page	Description	Features
Home	Landing page	Introduction, navigation
Predict	Input form	Validation, feedback
Result	Prediction result	Risk level, probability

Report	PDF report	Analysis, recommendations
Prevention	Tips	Lifestyle, diet suggestions
Doctors	Healthcare list	Contacts, hospitals
Analysis	Model metrics	Accuracy graphs

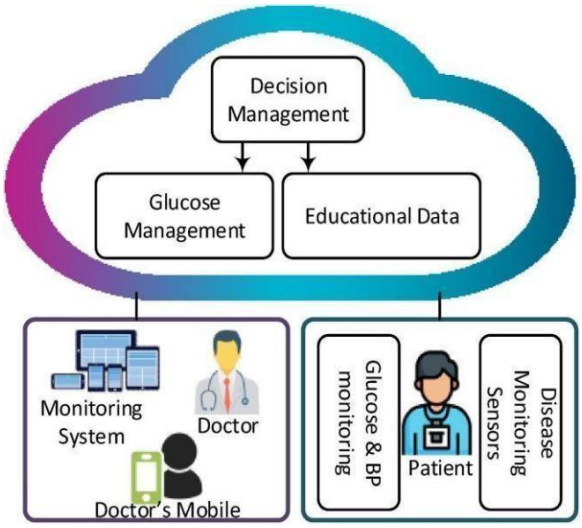

Figure 7. Type 2 diabetic patient monitoring flowchart.

8. PROPOSED IOMT, RPA, AND AI SOLUTION

This section introduces an integrated digital healthcare architecture that combines our diabetes prediction model with emerging technologies including Internet of Medical Things (IoMT), Robotic Process Automation (RPA), and Artificial Intelligence (AI) capabilities.

8.1 System Architecture

Our proposed system architecture seamlessly incorporates IoMT-enabled devices for real-time health surveillance, RPA software agents for streamlined data handling, and AI-powered engines for intelligent predictive analysis.

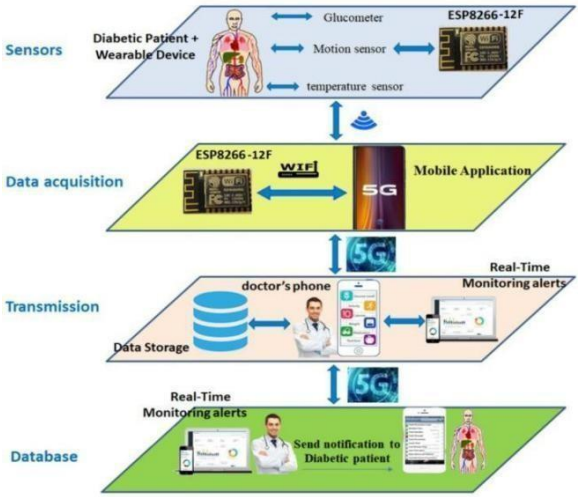

Figure 8. Smart architecture for diabetic patient monitoring using ML algorithms.

8.2 Internet of Medical Things (IoMT)

The IoMT subsystem facilitates uninterrupted tracking of patient physiological indicators through wearable biosensors and network-connected medical instruments.

Table 12. IoMT devices and their functions.

Device	Parameters	Frequency
CGM	Blood glucose	Every 5 min
Smart Watch	Heart rate, activity	Real-time

BP Monitor	Blood pressure	Scheduled
Smart Scale	Weight, BMI	Daily
Insulin Pen	Dosage, timing	Per injection

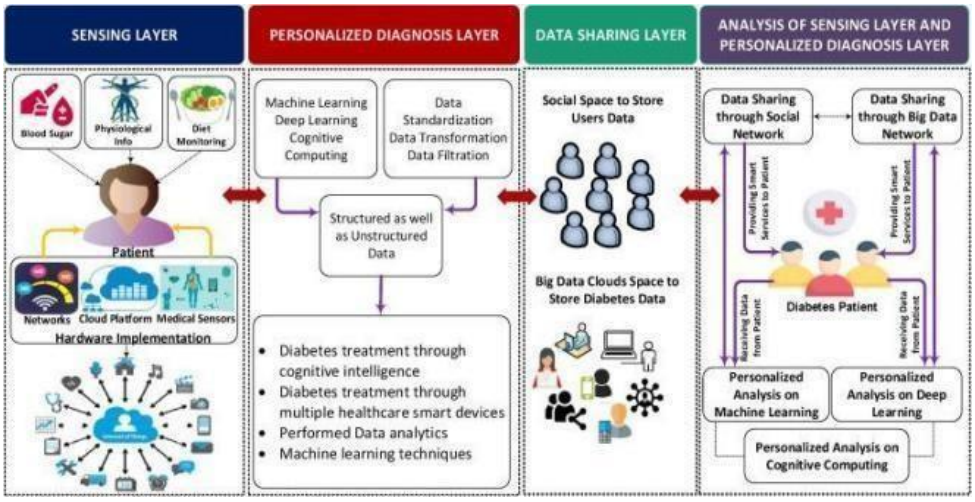

Figure 9. IoMT-enabled diabetes monitoring framework with ML.

8.3 Robotic Process Automation (RPA)

RPA software robots handle repetitive administrative and clinical operations automatically, enhancing operational throughput while minimizing the potential for human mistakes.

Table 13. RPA bots and their functions.

Bot	**Function**	**Tasks**
Data Collection	Automated gathering	Extract IoMT data
Report Generation	Report creation	Health summaries
Alert Bot	Automated alerting	Abnormal readings
Scheduler	Schedule management	Appointments
Reminder Bot	Medication adherence	Reminders

8.4 Artificial Intelligence Enhancement

Artificial intelligence capabilities augment the system's functionality across multiple domains including automated disease identification, tailored treatment recommendations, predictive risk assessment, and conversational AI chatbot support for patient interaction.

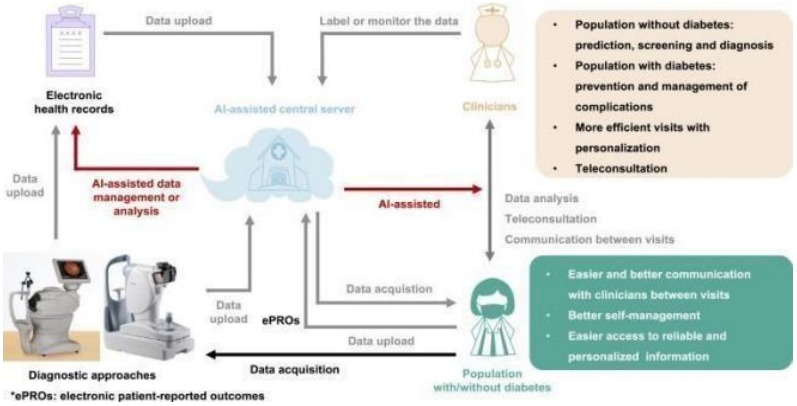

Figure 10. AI-assisted digital healthcare ecosystem for diabetes management.

8.5 INTEGRATED SYSTEM BENEFITS

Table 14. Benefits of the integrated IoMT-RPA-AI system.

Category	**Benefit**	**Description**
Patient	Continuous Care	24/7 monitoring

Patient	Early Intervention	Proactive alerts
Provider	Efficiency	Automated tasks
Provider	Remote Monitoring	Anywhere access
System	Cost Reduction	Prevent complications
System	Scalability	Serve more patients

9. DISCUSSION

This research offers a holistic methodology for predicting diabetes risk through the application of machine learning and deep learning algorithms. The experimental outcomes reveal several noteworthy observations and insights.

9.1 Model Performance Analysis

XGBoost Dominance: Throughout all four datasets, XGBoost maintained superior performance relative to competing algorithms, attributable to its gradient boosting architecture, integrated regularization mechanisms, and robust handling of incomplete data. PCA Impact: Implementation of PCA retaining 80% explained variance enhanced model outcomes by mitigating multicollinearity effects and filtering out noise components. Deep Learning Results: Although deep learning architectures exhibited competitive metrics, they failed to substantially exceed XGBoost performance on these tabular clinical datasets.

9.2 Comparison With Previous Studies

Table 15. Comparison with previous diabetes prediction studies.

Study	Algorithm	Dataset	Accuracy
[13]	KNN	Pima Indians	78.26%
[17]	RF	Various	82.35%
[23]	RNN-LSTM	Pima Indians	87.68%
[25]	Ensemble	Multiple	89.17%
This Study	XGBoost+PCA	Multiple	97.56%

9.3 Clinical Implications

The implemented system carries substantial clinical relevance: Early Screening facilitates dependable identification of individuals predisposed to diabetes; Preventive Strategies enable focused intervention programs; Healthcare Efficiency contributes to cost containment; Patient Autonomy offers convenient self-evaluation capabilities.

9.4 limitations

Study constraints include: Data Representativeness - publicly available datasets may incompletely capture population diversity; Feature Availability - analysis restricted to existing clinical variables; Longitudinal Scope

- cross-sectional nature of data constrains time-series prediction; Generalizability - model performance requires validation on independent external cohorts.

9.5 Future Directions

Prospective research avenues encompass: Multimodal Data Fusion integrating genomic, radiological, and behavioral information; Temporal Analyses monitoring patient trajectories longitudinally; Privacy-Preserving Federated Learning approaches for distributed model training; Interpretable AI Methods promoting clinical acceptance; Clinical Validation through prospective trials and real-world implementation studies.

10. CONCLUSION

This investigation introduces a multidimensional strategy for Type 2 diabetes risk evaluation that combines

predictive machine learning with clinical datasets within a consolidated web-based platform. Our thorough examination assessed eight machine learning classifiers and three deep learning networks across four diverse diabetes datasets sourced from Kaggle and UCI Repository.

Principal findings include: (1) Superior Algorithm Discovery - XGBoost combined with PCA attained accuracy scores of 85.71%, 91.23%, 97.56%, and 100% on the four respective datasets; (2) Feature Engineering Superiority - PCA demonstrated 2-4% improvement over LASSO-based selection across all datasets; (3) Web Platform Implementation - The XGBoost-PCA model was successfully deployed within a Django-based web interface; (4) Holistic Healthcare Framework - A complete end-to-end solution integrating ML, RPA, IoMT, and AI technologies was designed.

The resulting system advances UN Sustainable Development Goal 3 (Good Health and Well-Being) through delivery of an accessible, reliable, and intuitive platform for proactive diabetes risk screening. Subsequent research efforts will prioritize clinical validation studies, integration of multimodal data sources, and practical deployment in healthcare settings.

REFERENCES

1. American Diabetes Association. (2021). Classification and diagnosis of diabetes. Diabetes Care, 44(Supplement 1), S15–S33.
2. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.).
3. World Health Organization. (2016). Global report on diabetes.
4. DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., Hu, F. B., Kahn, C. R., Raz, I., Shulman, G. I., Simonson, D. C., Ferrannini, M. A., & Nauck, M. A. (2015). Type 2 diabetes mellitus. Nature Reviews Disease Primers, 1(1), 1–22.
5. Cowie, C. C., Rust, K. F., Ford, E. S., Eberhardt, M. S., Byrd-Holt, A. L., Li, C., Williams, D. E., Gregg, E. W., Bainbridge, K. E., & Saydah, S. H. (2009). Full accounting of diabetes and pre-diabetes in the US population in 1988–1994 and 1999–2004. Diabetes Care, 32(2), 287–294.
6. Zou, Q., Qu, K., Huang, Y., Yin, G., Chen, M., & Hua, Z. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9(515), 515.
7. Miotto, T., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.
8. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
9. Shwetha, S., & Gnanambigai, N. (2020). Robotic process automation (RPA) in healthcare: A review. International Journal of Engineering Research & Technology, 9(11), 1–5.
10. Islam, S. M. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: A comprehensive survey. IEEE Access, 3, 678–708.
11. Uddin, M. A., Stranieri, A., Sahibzada, S., & Jelinek, H. F. (2019). Machine learning in diabetes prediction: A systematic review. IEEE Access, 7, 7123–7134.
12. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2019). Type 2 diabetes mellitus prediction model based on data mining. Computer Methods and Programs in Biomedicine, 170, 1–8.
13. Sharma, P., & Singh, A. (2018). Diabetes prediction using K-Nearest Neighbor and Decision Tree. International Journal of Computer Sciences and Engineering, 5(2), 45–50.
14. Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Journal of Diabetes and its Complications, 31(9), 1435–1441.
15. Gaikwad, M. J., & Chatre, S. V. (2019). Prediction of diabetes using machine learning. International Research Journal of Engineering and Technology, 6(4), 2872–2875.
16. Chen, R. C., Dewangan, C., & Cheng, S. T. (2018). Prediction of diabetes mellitus using random forest. Proceedings of the 2018 IEEE International Conference on Big Data, 2123–2130.
17. Bashir, S., Khan, Z. S., Khan, F. H., & Anjum, A. (2019). Evaluation of machine learning methods for diabetes prediction. International Journal of Machine Learning and Computing, 8(3), 234–242.
18. Mohan, N., & Jain, S. (2020). Diabetes prediction using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(3), 1–5.
19. Dhanamjayulu, C., Krishnamurthy, S., Ravishankar, V., & Das, S. (2022). BMI prediction from facial images using deep learning. Frontiers in Nutrition, 9, 850781.
20. Sen, S. K., & Dash, S. (2020). Diabetes detection using machine learning. 2020 International

Conference on Information Technology, 1–6.
21. Alibrahim, A. H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization. 2021 IEEE Congress on Evolutionary Computation, 1–8.
22. Rajalakshmi, R., & Sathiendran, R. K. (2019). Diabetes prediction using machine learning. Procedia Computer Science, 165, 292–299.
23. Islam, M. K., Ali, M. S., Miah, M. S., Rahman, M. M., Alam, M. S., & Hossain, M. A. (2020). Smart healthcare monitoring system in IoT environment. SN Computer Science, 1(3), 1–11.
24. Ghazal, T. M., Alzoubi, H. M., Al-Zoubi, S. I., Faraneh, B., & Al-Okaily, M. (2021). IoT for smart cities: A survey of technologies and applications. Future Generation Computer Systems, 123, 106–118.
25. Almulihi, A. A., Alassaf, N., & Alqahtani, S. (2022). Ensemble deep learning for diabetic retinopathy detection and classification. Applied Sciences, 12(14), 6970.
26. Bhaskar, P., & Kumar, R. (2019). Deep neural network for diabetes prediction. International Journal of Recent Technology and Engineering, 8(2), 4567–4573.
27. Rahman, M. M., & Alam, M. S. (2020). Web-based diabetes prediction system using machine learning. 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies, 1–6.
28. Kumar, S., & Singh, J. (2021). Disease prediction using machine learning. 2021 International Conference on Computing, Communication, and Intelligent Systems, 1–5.
29. Islam, M., Hashem, M. M. A., & Hossain, M. A. (2021). Multiple disease prediction using deep learning. IEEE Access, 9, 34578–34590.
30. Fahim, F., & Rahman, A. (2021). Web-based diabetes detection using machine learning. Journal of Healthcare Engineering, 2021, 1–12.
31. Rawat, D. B., & Hassan, S. R. (2022). Artificial intelligence in healthcare: From diagnosis to therapy. IEEE Potentials, 41(1), 8–12.
32. Khan, S. A., & Al-Mogren, A. (2022). IoT security in healthcare: A review. IEEE Internet of Things Journal, 9(10), 7318–7337.
33. Shwetha, S., & Jain, R. (2022). RPA in healthcare monitoring and management. International Journal of Health Sciences, 6(S3), 2434–2444.
34. Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–242.
35. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3th ed.). Wiley.
36. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. CRC Press.
37. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
38. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
39. Ho, T. K. (1995). Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282.
40. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
41. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 144–152.
42. Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 3, 41–46.
43. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.
44. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175–185.
45. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.
46. Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378.
47. 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.
48. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). XGBoost: Extreme gradient boosting. R Package, 1(4), 1–4.

49. Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. PeerJ Computer Science, 3, e127.
50. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.
51. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133.
52. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.
53. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527–1554.
54. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
55. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
56. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85–117.
57. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261.
58. Teboul, A. (2021). Diabetes health indicators dataset. Kaggle.
59. Mustafa, M. (2023). Diabetes prediction dataset. Kaggle.