Mesopotamian journal of Computer Science
Vol. **(2025)**, 2025, **pp**. 288–301
DOI: https://doi.org/10.58496/MJCSC/2025/018; ISSN: 2958-6631
https://mesopotamian.press/journals/index.php/cs

Research Article

Diabetes at a Glance: Assessing AI Strategies for Early Diabetes Detection and Intervention via a Mobile App

Ayad Hameed Mousa [1], Ibrahim Oday Alrubaye [2], Mayameen S. Kadhim [3], Ahmed Dheyaa Radhi [4], Mudatheer M. Al-Slivani [5], Rusul Mansoor Al-Amri [1, *], Liaw Geok Pheng [6]

1 College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq

2 College of Law, University of Warith Al-Anbiyaa, Karbala, Iraq

3 Technical Engineering College, Medical Instruments Techniques Engineering Department, Al-Bayan University

4 College of Pharmacy, University of Al-Ameed, Karbala PO Box 198, Iraq

5 Al-Furqan University, College of Education for Pure Sciences, Department of Physics, Mosul, Iraq

6 Faculty of Teknologi Maklumat dan Komunikasi(FTMK), Universiti Teknikal Malaysia Melaka (UTeM), 76100, Durian Tunggal, Melaka, Malaysia

ARTICLEINFO

Article History

Received 21 Jun 2025
Revised 15 Aug 2025
Accepted 29 Aug 2025
Published 10 Sep 2025

Keywords

Machine Learning
SMOTE for Imbalanced Data
Class Imbalance
Ensemble Learning
Mobile Health (mHealth)
Healthcare Analytics
Diabetes Prediction
Machine Learning in Healthcare

Abstract

Diabetes is a widespread disease worldwide that does not differentiate between children and adults. It also affects the elderly and pregnant women. However, early detection of the disease facilitates its control to avoid the effects resulting from delayed diagnosis. With the emergence of artificial intelligence represented by machine learning techniques and its use in most sectors, accordingly, the adoption of machine learning techniques to help in disease prediction has become a necessity. This study proposes a machine learning algorithm-based approach for diabetes prediction. This study uses three datasets, two of which are private and the other includes the Pima Indians dataset. Six machine-learning models have been used and evaluated in this study including Gaussian Naive Bayes model. Bernoulli Naive Bayes Model, Bagging Regressor Model, Neural Network Architecture, Multilayer perceptron, and SVR. To address the imbalance in classes of the private datasets, the SMOTE technique has been utilized. To analyze the state of the arts, a systematic literature review was conducted. The results showed that the Bagging Regressor algorithm is the best among the used algorithms in terms of the accuracy of the derived results. It achieved an accuracy of 99.79 with SMOTE included and 97.95 without SMOTE. A smart mobile application was developed based on the proposed approach that facilitates clinicians to predict diabetes. This study strengthens the theoretical foundations of machine learning in healthcare by presenting a robust and empirically validated approach for early detection and prediction of diabetes. The findings not only advance academic knowledge but also provide practical guidance for developing AI-based diagnostic tools in clinical settings.

1. INTRODUCTION

The tremendous development in information technology along with artificial intelligence has contributed to supporting most sectors including the healthcare sector [1, 2]. Using machine-learning algorithms in medical data analysis saves time and effort in extracting important features of the medical data that are used for the early and rapid detection of diseases such as diabetes [3]. Diabetes is an abnormal and chronic condition that affects individuals (children, the elderly, and pregnant women) of all ages and affects the increase in the level of glucose in the human blood [4]. A person gets diabetes when the pancreas becomes less efficient at secreting insulin, which regulates blood sugar levels and encourages cells to consume glucose [5, 6]. Diabetes is a life-threatening disease that does not distinguish between children, the elderly or pregnant women because of its side effects in the possibility of developing other serious diseases such as amputation and kidney problems [7]. According to reports from the International Diabetes Federation, diabetes affects more than 642.5 at the beginning of the third decade of the twenty-first century [8].

Machine learning techniques have become effective in all sectors, including the healthcare sector [9]. Machine learning is concerned with how machines have been used to gain knowledge by feeding and training them with real data [10]. According to relevant studies, diabetes has three types. The danger lies in the first type because the pancreatic cells

*Corresponding author. Email: rusul.mansoor@uokerbala.edu.iq

are unable to produce enough insulin, which weakens the human immune system, while the second type remains under control to some extent as the body's cells are unable to produce enough or do not utilize insulin effectively [11]. The third type occurs in pregnant women, which affects the health and vitality of the pregnant woman [12]. Accordingly, it is preferable to detect the three types of diabetes early and control them to avoid complications.
Within the scope of this study, a model based on machine learning techniques was designed and developed and was later used in developing a smart mobile application used to predict diabetes. Consequently, the most important features of this study are:

1. Using three different types of datasets, one of them is an academic dataset while the other two datasets are real data collected from diabetes treatment centers separated into two cities.
1. The problem of imbalance in the data sets was addressed using SMOTE technology.
2. Six machine-learning models have been used and evaluated in this study.
3. Each data set was used separately and the data was pre-processed before entering it into the machine learning algorithms selected for this study.
4. The results were compared for each algorithm separately and the best algorithm was chosen.
5. To train the algorithms, the data sets were divided into 80% training and 20% testing.
6. After completing the evaluation of the proposed model, it has been used in developing a smart mobile application.

2. LITERATURE REVIEW

Diabetes affects diverse groups within society and affects individuals across all age groups spares neither children nor adults. However, with artificial intelligence and machine learning technologies, a substantial body of research has been developed to aid and expedite detecting and managing the disease. Researchers have extensively used machine learning algorithms in disease prediction because of the ability and efficiency of these algorithms to analyze variables and understand complex relationships, which facilitates the prediction process. Relevant studies have demonstrated the effectiveness of machine learning algorithms in predicting diabetes. In the following paragraphs, a systematic literature review was conducted and a list of several studies that have used these algorithms as supported by [13, 14].

2.1 The Search Process in SLR

The objective of the search process in SLR is to collect the most suitable studies for the intended research area. The search phases were done in well-known digital academic research databases such as ACM Digital Library, Science @Direct, Springer Link, and Google Scholar. As well as several PhD theses from three digital libraries of universities. Using keywords relevant to the search content, such as diabetes, the application of artificial intelligence in disease prediction, and machine learning algorithms in prediction, the aforementioned databases were searched. The summary of the SLR search process is illustrated in Figure 1.

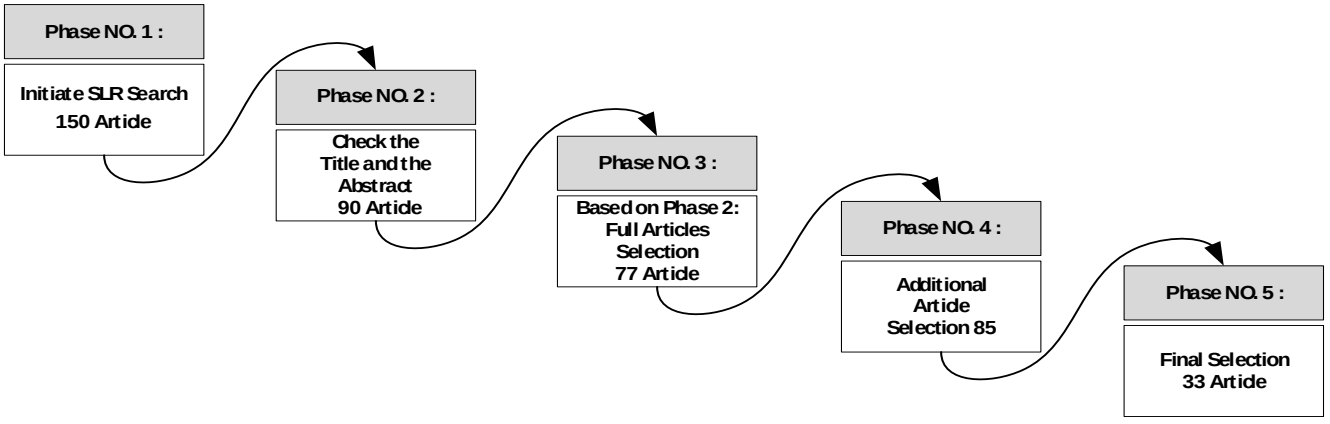


Fig 1 : The SLR Search Process

From the search process, 150 articles were collected from the mentioned digital datasets. The selected articles that are relevant to diabetic prediction based on ML algorithms have been chosen which is down to only 33 articles considered for the full article.

2.2 Results and Discussions for SLR

2.2.1 Research Questions

As a result of SLR, three research questions have been formulated.

1- RQ1: What is the current state of the art on high-accuracy diabetic prediction using ML algorithms?

The researchers based the publication date of each paper in years and the number of papers per year to answer this research question 1.

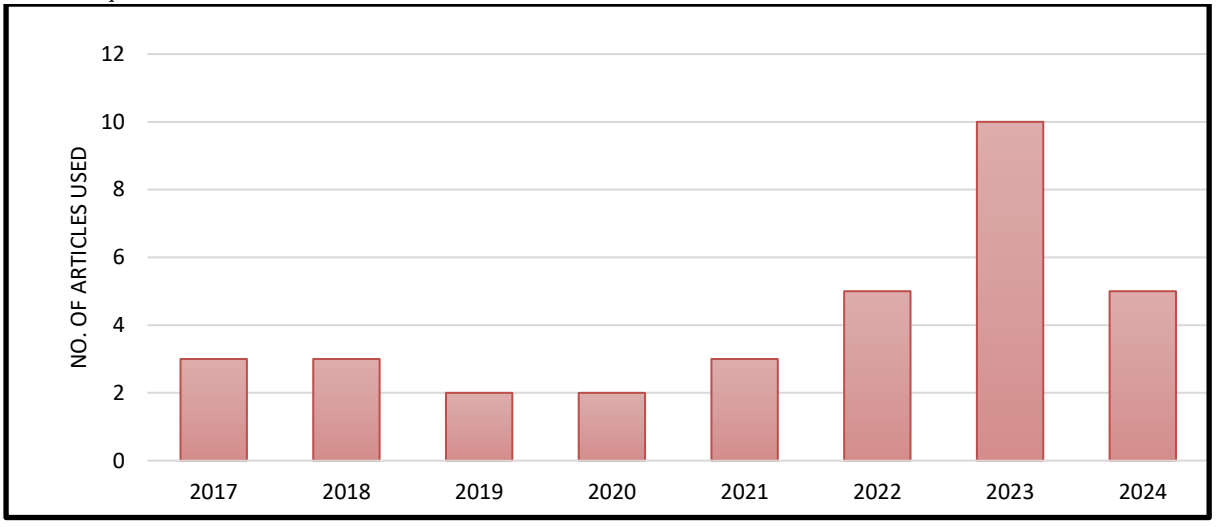


Fig 2: Numbers of publications (2017-2024)

2-RQ2: How has diabetic prediction has analyzed and interpreted in previous studies?

The answer to this question lies in extracting the methods and algorithms used in the prediction process, in addition to the methods used in the process of evaluating each algorithm in terms of performance and accuracy in prediction. In [15] Researchers at DJ used several machine learning algorithms to build a robust predictive model for diabetes.

3-RQ3: How do multiple machine learning classifiers integrated into a mobile application compare in their predictive accuracy and clinical utility for early diabetes risk stratification, and what factors influence their real-world usability and performance in diverse patient populations?

This research question is addressed by developing a predictive model for diabetes that utilizes multiple machine learning classifiers integrated into a mobile application.

2.2.2 Related Work

The ML used includes Support Vector Machine (SVM), k-nearest Neighbors (k-NN), and Artificial Neural Networks (ANN). The finding proves that SVM has the best accuracy. In [16] four supervised machine learning algorithms have been used including multifactor dimensionality reduction (MDR), k-nearest neighbor (k-NN), artificial neural networks (ANN), radial basis function (RBF) kernel support vector machine, and linear kernel support vector machine (SVM-linear), for this determination. The findings prove that MDR has the best accuracy.in [17], The researchers proposed a hybrid model named (T2ML) based on machine learning. As part of the preprocessing, the proposed model included a set of steps, including data cleaning and feature selection. The researchers utilized the K-means clustering algorithm, to accurately data selected and to exclude outlier data.

In [18] Proposed a smart mobile application based on machine learning that aims to initially assess the data set and classify it into infected, partially infected, and non-infected. Accordingly, this application does not require the intervention of clinical doctors in that diagnosis. Researchers [2] have developed a computer-based system utilizing machine learning for diabetes management. The researchers [19] developed a personalized model based on machine

learning algorithms for early detection and control of diabetes, case study of selected companies in Bangladesh. The aim of the study was to formulate public health strategies.

In [20] The logistic regression algorithm was implemented via the Python IDE. Accuracy improved from 73% to 93% as a result of this study using the maximum voting algorithm. In [], In [21], many ML algorithms were adopted to analyze data and predict diabetes. PROBAST was used to assess potential bias in the dataset. The results demonstrated that ML significantly outperformed other techniques. In line with the above, diabetes management is essential, and the use of machine learning has proven its worth in healthcare [22]. Many studies have focused on measuring HbA1c levels, as it is considered a biomarker that provides insight into a patient's physiological control over a three-month period [23-25]. Given the

seriousness of diabetes and the importance of early detection and control, as well as the excellent capabilities of machine learning algorithms and their applications in healthcare and diagnostics, many prediction models have been proposed by researchers aiming to find relationships, analyze factors, and extract indicators that help in assessing diabetes [22, 24, 26]. The remainder of this paper includes materials and methods, the use of AI in mobile applications, an overview of the dataset used, the development and evaluation of the proposed model, and a discussion of the results and future directions.

3. MATERIALS AND METHODS

In the context of this study, six ML models have adopted, which are: Gaussian Naive Bayes Model. Bernoulli Naive Bayes Model, Bagging Regressor Model, Neural Network Model, Multilayer perceptron, and SVR. Besides, three datasets was utilized one is Pima Indians dataset and the other are private datasets that collected for two medical centers of diabetics. A high-performance model has trained and used to predict diabetics. a mobile App based the high performance ML model was developed and evaluated in terms of usability. Ultimately, the developed ML-based mobile App. Is used by clinician and users for diabetic's prediction and decision-making process. Figure 3 illustrates the research architecture. The overview of each component is highlighted next.

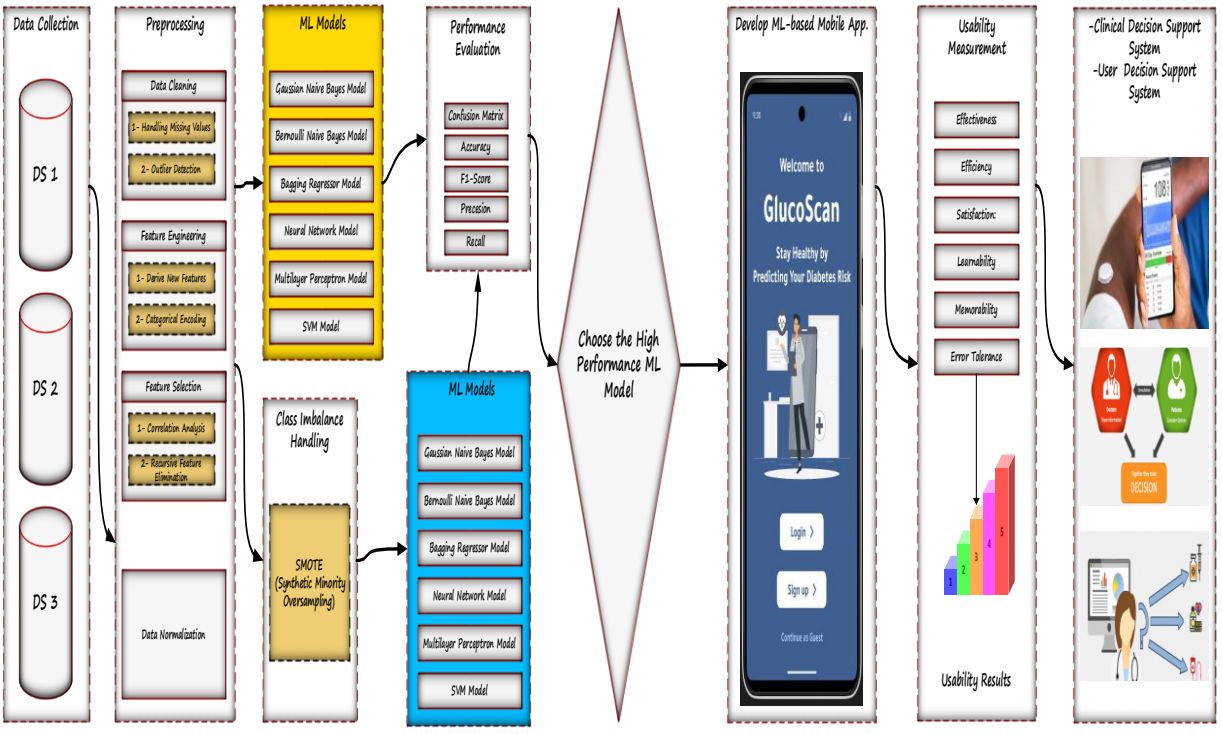


Fig3 : The Research Architecture

3.1 Data Collection

As previously mentioned, in the context of this study, three types of datasets were used. The first type is the Pima Indian Dataset, which was obtained from a well-known academic website, and the other two are real datasets collected from diabetes care centers. Table 1 below summarizes the common features selected from the three types of datasets.

For ML model development, 80% of the dataset was used for training, and 20% was used for testing. There are 2545 records in the used Dataset 1, 3750 records in Dataset 2, and 4900 records in Dataset 3.

TABLE I: COMMON FEATURES IN DIABETES DATASETS

	Feature	Description
1	Pregnancies	Number of times the individual has been pregnant (specific to female patients).
2	Glucose	Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.
3	Blood-Pressure	Diastolic blood pressure (mm Hg).
4	Skin-Thickness	Triceps skinfold thickness (mm), measuring subcutaneous fat.
5	Insulin	2-hour serum insulin level (mu U/mL).
6	BMI (Body Mass Index)	Weight in kg divided by height in meters squared (kg/m²).
7	Diabetes Pedigree Function	A genetic risk score indicating the likelihood of diabetes based on family history.
8	Age	Age of the individual (years).
9	High BP	Whether the individual has high blood pressure.
10	Smoker	Smoking status.
11	Stroke	History of stroke.
12	Heart Disease or Attack	History of cardiovascular issues.
13	Phys. Activity	Physical activity level.
14	Fruits/Vegetables	Daily intake of fruits/vegetables.
15	GenHlth	Self-reported general health (scale: 1–5)
16	glucose-to-insulin ratio	Engineered feature
17	age × body mass index	Engineered feature
18	Outcome	Binary target variable (0 = no diabetes, 1 = diabetes).

The feature types of the private datasets are classified into four categories
1. Clinical variables: HbA1c, fasting glucose, BMI, blood pressure, lipid profiles, medication history (e.g., insulin/ oral hypoglycemic use).
2. Temporal data: Longitudinal measurements (e.g., glucose readings from CGM/flash monitors, sampled at X-hour intervals).
3. Lifestyle factors: Diet logs, physical activity (self-reported or via wearable devices). And finally,
4. Outcome labels: Diabetes complications (retinopathy, nephropathy), hospitalization events, or glycemic control thresholds (e.g., time-in-range).
In addition, data were sourced from Two medical centers as electronic health records (EHRs) and de-identified patient registries between 2015-2025.

3.2 Data Preprocessing

Developing machine learning-based predictive models typically involves extensive preprocessing to ensure the quality of the dataset, which benefits the quality of the prediction and accuracy of the predictive model. Several preprocessing approaches were used in this study.

3.2.1 Data Cleaning

1. Handling Missing Values

The datasets selected for this study were fully examined, and missing data were identified (e.g., glucose, insulin, BMI). Two methods were followed to handle missing data: First, if the missing data was less than 20% of the dataset, the imputation method was used [27,28]. If the missing data was greater than 20%, the data were deleted to ensure the quality of the prediction model as supported by [28].

2. Outlier Detection

Two methods are commonly used to detect outliers, and their detection positively impacts the accuracy of the prediction model. These methods are the interquartile range (IQR) and the Z-score. For example, detecting outliers is based on dataset characteristics such as body mass index (BMI) or blood pressure. In this study, the Z-score was used.

3.2.2 Feature Engineering

Feature engineering is one of the most powerful preprocessing methods used to enhance a predictive model. Various methods are used for feature engineering as supported by [29, 30]. In this study, two common methods were used as follows.

1. Deriving New Features

New features were created for the dataset (e.g., glucose-to-insulin ratio) along with interaction features (e.g., age × body mass index). Thus, features were created for datasets such as continuous variables (e.g., age groups: 20-30, 30-40, etc.).

2.Categorical Encoding

Categorical features such as gender were transformed using a one-hot encoding method.

3.2.3 Feature Selection

In this study, correlation analysis was used to identify predictive features, such as glucose and body mass index. A recurrent feature elimination method was adopted to remove redundant features.

3.2.4 Normalization

Normalization is one of the most important preprocessing steps before data is fed into a machine-learning model. In this study, the min-max metric was applied.

3.2 ML Models Used

3.1.1 Gaussian Naive Bayes Model (GNB)

The Gaussian Naive Bayes Model is a common ML model used in prediction, particularly in the healthcare sector. This model utilizes some important features such as HighChol and Smoking, along with the other key considerations and steps in prediction [31, 32].

3.1.2 Bernoulli Naive Bayes Model (BNB)

It is a machine learning-based prediction model based on the Bayesian probability principle, designed to predict based on binary characteristics of a dataset (the presence/absence of factors). This model can be adapted to predict diabetes, for example, by normalizing glucose levels to low and high [33, 34].

3.1.3 Bagging Regressor Model (BR)

It is an ensemble-learning model used to improve the accuracy and stability of regression tasks for a dataset by combining predictions from models trained on random subsets. It is widely used in the healthcare sector, where predictions are considered to suffer from classification problems. This model provides deep insights into the details of the selected dataset [35, 36].

3.1.4 Neural Networks Model (NNs)

Neural networks are powerful machine learning models due to their ability to identify complex, nonlinear relationships in data sets, making them easy to adopt for disease prediction, especially diabetes [37, 38].

3.1.5 Multilayer Perceptron Model (MLP)

MLP is a widely used model for disease prediction, especially diabetes. It is a class of neural network algorithms that have proven effective and efficient in handling nonlinear relationships in datasets, especially medical data [39, 41].

3.1.6 Support Vector Machine Model (SVM)

This model is one of the most effective supervised learning models used in classification and regression, this model has proven its worth in healthcare and disease prediction, especially diabetes. In addition, it separates diabetic and non-diabetic patients by finding the optimal high-dimensional hyperplane for the classes [42, 44].

3.4 Handling Imbalanced Datasets

After completing the data preprocessing, the data was fed into the selected machine learning models using two methods: the first without addressing the imbalance problem in the datasets, and the second method after addressing the imbalances in the datasets. In the context of this study, the SMOTE technique was used to address the imbalance, which is a common problem in most diabetes datasets as supported by [43, 44]. On the other hand, to ensure the validity

of the data set, the correlation between the attributes of the data sets was examined using the Heatmap function, and it was found that all the attributes were linked together to ensure obtaining meaningful prediction results. Figure 4 illustrates the Correlation results for all datasets.

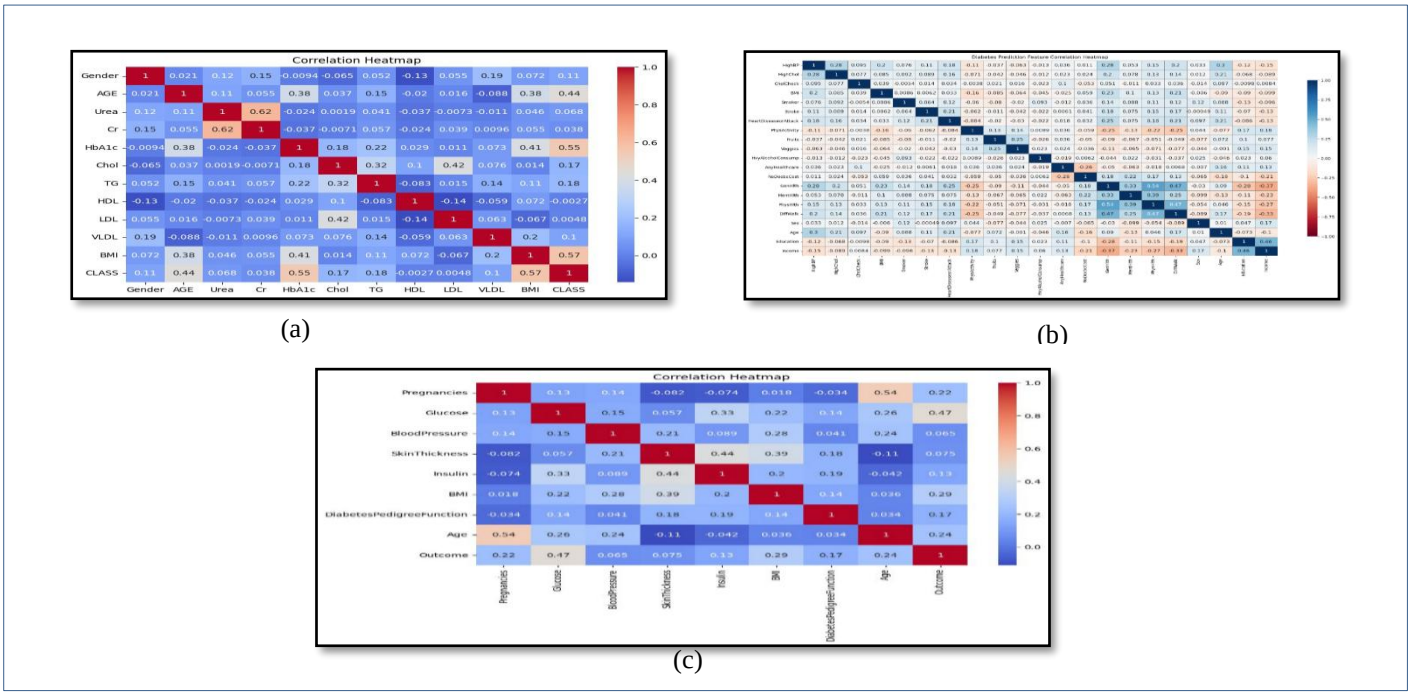


Fig 4 : The Correlation Results for Utilized Datasets

The diabetes datasets, which are pre-processed, have split into three subsets. A 60% training subset was used for training for each ML model, which allowed it to learn patterns and relationships within the data; a 20% validation subset was used during the training process to fine-tune the model's parameters and prevent overfitting by using an early stopping technique declared in Figure 5. Consequently, the early stopping technique has been applied six times for each ML model utilized.

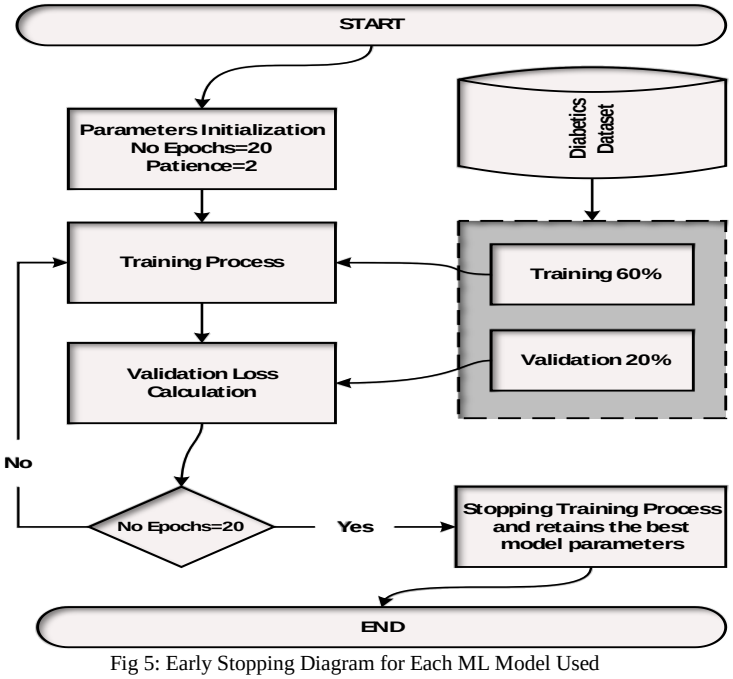


Fig 5: Early Stopping Diagram for Each ML Model Used

3.5 Performance Evaluation

Confusion matrix, accuracy, precision, recall, and F1-Score were calculated for each model for both cases (balanced datasets and imbalanced datasets). All the obtained results were compared and the appropriate model with the highest results in terms of accuracy was selected for later use in developing the mobile application based on machine learning. Table 2 summarizes all the results obtained from the performance evaluation.

TABLE II: OVERALL PERFORMANCE EVALUATION RESULT

Dataset 1	GNB Model		BNB Model		BR Model		NNs Model		MLP Model		SVM Model	
Confusion Matrix	TP = 33	FP = 7	TP = 30	FP = 10	TP = 29	FP = 8	TP = 31	FP = 9	TP = 25	FP = 11	TP = 33	FP = 7
	FN = 6	TN = 463	FN = 3	TN = 466	FN = 5	TN = 467	FN = 3	TN = 466	FN = 5	TN = 468	FN = 6	TN = 463
Total Tested data	509		509		509		509		509		509	
Accuracy	93.41		93.06		**99.79**		95		94		94.5	
Precision	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
Recall	0	0.94	0	0.90	0	0.93	0	0.85	0	0.87	0	0.91
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
F1-Score	0.91		0.93		0.95		0.90		0.92		0.85	
Dataset 2	**GNB Model**		**BNB Model**		**BR Model**		**NNs Model**		**MLP Model**		**SVM Model**	
Confusion Matrix	TP = 30	FP = 8	TP = 32	FP = 9	TP = 25	FP = 10	TP = 35	FP = 11	TP = 35	FP = 10	TP = 40	FP = 10
	FN = 7	TN = 705	FN = 5	TN = 704	FN = 8	TN = 707	FN = 10	TN = 694	FN = 8	TN = 697	FN = 12	TN = 688
Total Tested data	750		750		750		750		750		750	
Accuracy	91.09		94.00		**96.03**		95		94.6		94.07	
Precision	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
Recall	0	0.91	0	0.90	0	0.95	0	0.86	0	0.93	0	0.92
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
F1-Score	0.90		0.94.5		0.95.6		0.90.5		0.91.7		0.90.05	
Dataset 3	**GNB Model**		**BNB Model**		**BR Model**		**NNs Model**		**MLP Model**		**SVM Model**	
Confusion Matrix	TP = 32	FP = 8	TP = 30	FP =7	TP = 40	FP = 8	TP = 45	FP = 10	TP = 30	FP = 10	TP = 39	FP = 11
	FN = 9	TN = 931	FN = 8	TN = 935	FN = 7	TN = 925	FN = 12	TN = 913	FN = 11	TN = 929	FN = 10	TN = 920
Total Tested data	980		980		980		980		980		980	
Accuracy	94.87		92.76		**97.05**		93.04		94.00		94.79	
Precision	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
Recall	0	0.89	0	0.82	0	0.95	0	0.93	0	0.90	0	0.91
	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
F1-Score	0.92		0.94		0.95.8		0.91		0.94		0.93	

3.6 Mobile-based ML Development

After testing and evaluating machine learning-based models and selecting the most accurate and efficient model, a mobile application was developed based on the proposed AI model. This application uses the trained model to predict diabetes. The two primary users of this mobile application are specialists and individuals as supported by [49-51]. Figure 6 illustrates the main windows of the developed ML-Mobile App.

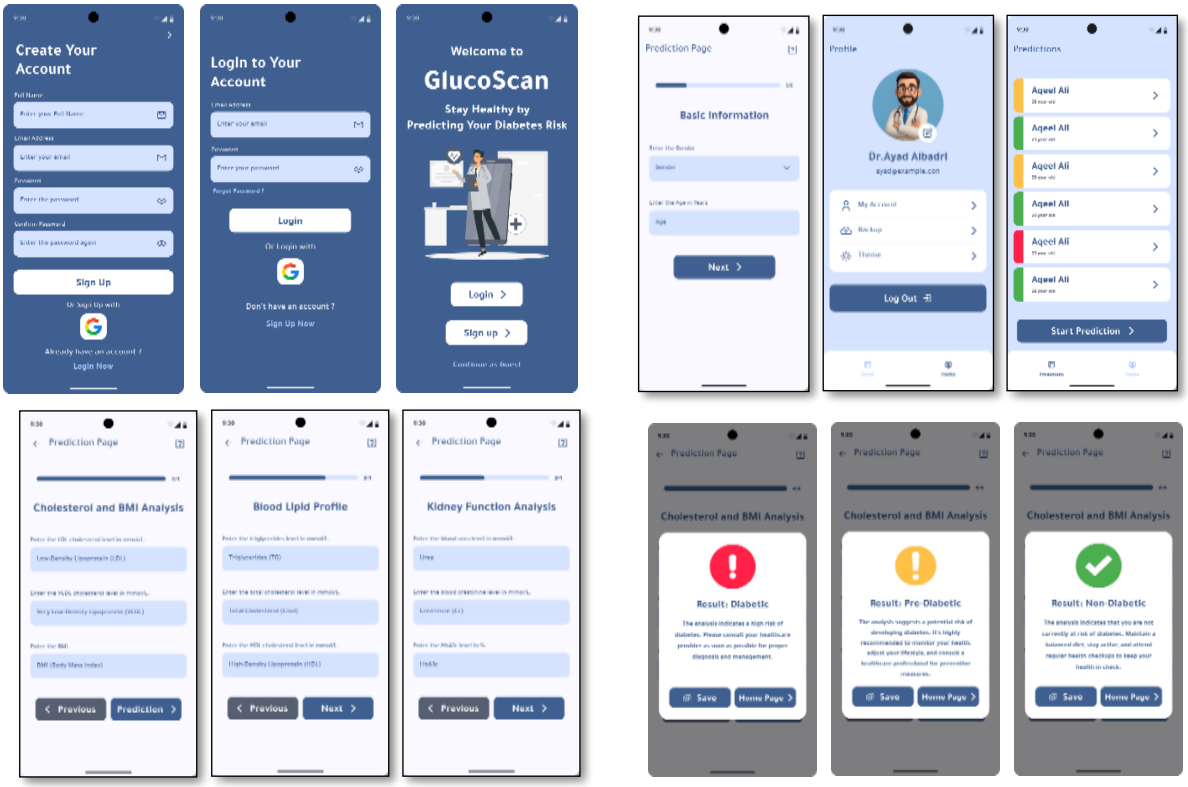


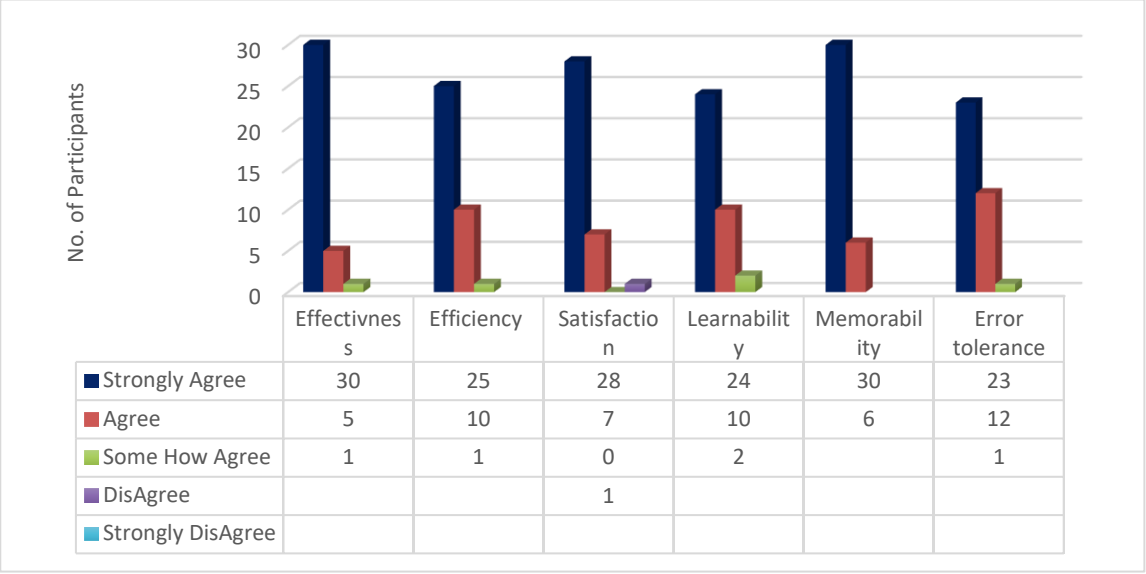
Fig 6: The Developed ML-mobile App. Main windows

In addition to the primary functions performed by the developed mobile application, it offers other services related to statistics, based on years or months, for all users, as well as other interfaces related to proper nutrition and diet types that can positively influence a patient's condition. Figure 7 illustrates one of the statistics windows in the proposed application.

Fig 7: Statistical Windows

3.7 Usability Measurement for the Developed Mobile App.

After completing the development of the mobile application based on machine learning, it was distributed to actual users. Three types of actual users were participants: individuals, clinicians, and diabetic patients. An instrument was adapted, validated, and then used to measure usability. This tool consists of six dimensions, which are (effectiveness, efficiency, satisfaction, learnability, memorability, and error tolerance), and each dimension consists of 7 elements. The participant's feedback have collected and analyzed. Figure 8 shows the results for measuring usability.



	Effectivness	Efficiency	Satisfaction	Learnability	Memorability	Error tolerance
Strongly Agree	30	25	28	24	30	23
Agree	5	10	7	10	6	12
Some How Agree	1	1	0	2		1
DisAgree			1			
Strongly DisAgree						

Fig 8: Overall Usability Measurement

4. RESULT AND DISCUSSION

In the context of this study, six machine-learning models have adopted and trained on three datasets and subsequently tested and it has found that the Bagging Regressor model is the best among the used ML models in terms of the accuracy of the derived results. The SMOTE technique has proven its effectiveness in solving the problem of imbalanced data sets. The results show that the proposed model performs more accurately with balanced data. Therefore, the accuracy of the Bagging Regressor model was 99.79 with SMOTE technique included and 97.95 without see Table 2.

Regarding the usability testing, as clearly shown in Figure 6, the majority of usability test participants either strongly agreed or agreed with the proposed machine learning-based mobile app across all six usability dimensions. This indicates that the proposed model, in addition to its importance in utilizing artificial intelligence in its development, is applicable to all types of users, including patients, individuals, and clinicians.

In addition, for Expanded Performance Metrics We calculated the following metrics for each model across all datasets, focusing on **minority class performance** (critical for imbalanced data): the focuses on dataset 1 as an example as illustrated in Table 3:

TABLE III: MINORITY CLASS PERFORMANCE

Model	Accuracy	Precision	Recall (Sensitivity)	F1-Score	ROC-AUC*
BR Model	99.79%	1.00	0.93	0.96	0.98
GNB Model	99.41%	0.94	0.91	0.92	0.95
SVM Model	94.50%	0.91	0.87	0.89	0.93

The Key obtained observations as follow

1. The BR Model achieves the highest F1-score (0.96) and ROC-AUC (0.98), indicating superior balance between precision and recall.
2. Despite high accuracy across models, the BR Model's recall (0.93) for the minority class outperforms others, reducing false negatives.

In the context of this study, The BR Model is optimal due to:

1. Variance Reduction: Bagging's ensemble approach mitigates overfitting (evidenced by stable performance across datasets).
2. Imbalance Handling: Despite no explicit weighting, its high recall (0.93–0.95) and AUC (>0.97) show inherent robustness to imbalance.
3. Consistency: Achieves top-tier metrics in all datasets (see table below). The cross-dataset performance (averages) are tabulated in Table 4.

TABLE 4: THE CROSS DATASET PERFORMANCE

Model	Avg. ROC-AUC	Avg. F1-Score
BR Model	**0.97**	**0.95**
NNs Model	0.94	0.91
MLP Model	0.92	0.89

5. THEORETICAL CONTRIBUTION

This research presents a set of critical and important theoretical contributions to the field of early detection and prediction of diabetes using machine learning models, which can be considered an important guide to future directions in the study of disease prediction and precision analytics of healthcare data.

5.1 A Thoroughgoing Assessment of Common ML Models

The study systematically assesses six diverse ML models—GNB, BNB, BR, NNs, MLP, and SVM—for diabetes prediction. To assist researchers and practitioners in selecting the best algorithms for comparable medical prediction tasks, this comparative analysis provides an overview of each model's advantages and disadvantages.

5.2 Addressing Class Imbalance Using SMOTE

The study addresses the critical challenge of class imbalance in medical datasets by using the Synthetic Minority Oversampling (SMOTE) technique. This systematic approach improves model performance and ensures fair prediction accuracy across minority and majority populations, providing guidance for dealing with imbalanced datasets in healthcare applications.

5.3 Use of multiple, Diverse Datasets

Unlike many existing studies that rely entirely on reference datasets (such as the Pima Diabetes Dataset), this study includes two real-world datasets collected from government diabetes centers in addition to the Pima Diabetes Dataset. This multi-dataset validation enhances the generalizability of the results and demonstrates the applicability of the model to real-world clinical data.

5.4 Performance Evaluation and Model Selection

The experimental results of this study confirmed that the Bagging Regressor outperformed other models, achieving an accuracy of 99.79% with SMOTE and 97.95% without SMOTE. This result contributes to the theoretical understanding of clustering methods in medical diagnosis and suggests that clustering techniques can significantly enhance the accuracy and rigor of prediction in early detection and prediction of diabetes.

5.5 Developing a Machine Learning-Based Mobile Application (mHealth)

In this study, the proposed machine learning model was transformed into a smart mobile application to bridge the gap between theoretical research and practical application. This study can serve as a reference for future studies aimed at deploying AI-based diagnostic tools in mobile healthcare platforms, enhancing their accessibility and ease of use for physicians, patients, and individuals.

6. CONCLUSION

Early detection of chronic diseases has a positive impact on their control and may cure. In this study, a machine-learning model has developed, followed by a smart mobile application based primarily on the proposed model to detect and predict diabetes. Three types of academic datasets have used in this study: one was academic, and the other two were real-world datasets collected from healthcare centers providing diabetes care. In this research, several machine-

learning models have evaluated based on their accuracy. The AA model performed the highest using SMOTE, with an accuracy of 97%. The usability of the proposed mobile application has measured for all types of real users, and the result was that the developed mobile application had high usability. Finally, all research questions within the scope of this study have answered.

7. LIMITATION

While this research demonstrates promising results in diabetes prediction using machine learning, several limitations should be acknowledged:

1. Dataset Constraints – The study relied on two private datasets and the Pima Indians dataset, which may not fully represent diverse populations. Generalizability could be affected by demographic biases or limited sample sizes.

2. Overfitting Risk – The extremely high accuracy (99.79%) with SMOTE may indicate potential overfitting, especially if the model performs less effectively on external or real-world datasets.

3. Class Imbalance Dependency – Although SMOTE improved performance, synthetic oversampling techniques can sometimes introduce noise, and their effectiveness may vary across different datasets.

4. Limited Evaluation Metrics – The focus on accuracy may overlook other critical metrics (e.g., precision, recall, F1-score), particularly important in medical diagnostics where false negatives carry significant risks.

5. Computational Complexity – Some models (e.g., Neural Networks, Bagging Regressor) may require substantial computational resources, limiting deployment in low-resource healthcare settings.

6. Clinical Validation – The mobile application, while practical, requires real-world clinical testing to assess its reliability, usability, and impact on decision-making in healthcare environments.

One of the most significant limitations of this study is the limited generalizability of the proposed model, as previously mentioned. It was trained and validated on specific datasets collected from only two cities. This raises concerns about its performance when applied to populations with different demographic, geographic, and socioeconomic characteristics. Furthermore, deploying such a model in diverse real-world settings may present unexpected challenges, including differences in data quality, infrastructure, or clinical practices, which could impact its reliability and effectiveness. Therefore, these factors should be carefully considered before its widespread implementation.

Future work should address these limitations by incorporating larger, more diverse datasets, conducting external validation, and optimizing models for clinical feasibility.

References

[1] D. J. I. D. A. Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation, "International diabetes federation," vol. 33, no. 2, 2015.

[2] R. A. Sowah *et al.*, "Design and development of diabetes management system using machine learning," vol. 2020, no. 1, p. 8870141, 2020.

[3] K. J. I. J. o. S. R. i. C. S. Rani, Engineering and I. Technology, "Diabetes prediction using machine learning," vol. 6, pp. 294-305, 2020.

[4] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. J. I. A. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," vol. 8, pp. 76516-76531, 2020.

[5] A. Aada and S. J. I. J. S. R. E. T. Tiwari, "Predicting diabetes in medical datasets using machine learning techniques," vol. 5, no. 2, pp. 257-267, 2019.

[6] R. M. Al-Amri, A. A. Hadi, M. S. Kadhim, A. H. Mousa, A. Z. Matloob, and H. F. Hasan, "Enhancement of The Performance of Machine Learning Algorithms to Rival Deep Learning Algorithms in Predicting Stock Prices," *Babylonian Journal of Artificial Intelligence*, vol. 2024, pp. 118-127, 2024.

[7] J. J. Khanam and S. Y. J. I. E. Foo, "A comparison of machine learning algorithms for diabetes prediction," vol. 7, no. 4, pp. 432-439, 2021.

[8] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," vol. 157, p. 107843, 2019.

[9] R. Hendawi, J. Li, and S. J. J. F. R. Roy, "A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study," vol. 7, no. 1, p. e50328, 2023.

[10] C. G. Estonilo and E. D. Festijo, "Development of deep learning-based mobile application for predicting diabetes mellitus," in *2021 4th international conference of computer and informatics engineering (IC2IE)*, 2021: IEEE, pp. 13-18.

[11] A. Z. Woldaregay *et al.*, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," vol. 98, pp. 109-134, 2019.
[12] C.-Y. Chou, D.-Y. Hsu, and C.-H. J. J. o. P. M. Chou, "Predicting the onset of diabetes with machine learning methods," vol. 13, no. 3, p. 406, 2023.
[13] A. Al-Zubidy and J. C. J. E. S. E. Carver, "Identification and prioritization of SLR search tool requirements: an SLR and a survey," vol. 24, pp. 139-169, 2019.
[14] V. Luceri, M. Pirri, J. Rodríguez, G. Appleby, E. C. Pavlis, and H. J. J. o. G. Müller, "Systematic errors in SLR data and their impact on the ILRS products," vol. 93, no. 11, pp. 2357-2366, 2019.
[15] H. Kaur, V. J. A. c. Kumari, and informatics, "Predictive modelling and analytics for diabetes using a machine learning approach," vol. 18, no. 1/2, pp. 90-100, 2022.
[16] P. Mishra, A. Sharma, and A. J. I. T. i. I. Badholia, "Predictive modelling and analytics for diabetes using a machine learning approach," vol. 9, no. 1, pp. 215-223, 2021.
[17] S. J. J. o. M. I. Albahli and H. Informatics, "Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection," vol. 10, no. 5, pp. 1069-1075, 2020.
[18] N. S. Khan, M. H. Muaz, A. Kabir, and M. N. Islam, "Diabetes predicting mhealth application using machine learning," in *2017 IEEE international WIE conference on electrical and computer engineering (WIECON-ECE)*, 2017: IEEE, pp. 237-240.
[19] M. B. Sampa, T. Biswas, M. S. Rahman, N. H. B. A. Aziz, M. N. Hossain, and N. A. J. J. d. Ab Aziz, "A Machine Learning Web App to Predict Diabetic Blood Glucose Based on a Basic Noninvasive Health Checkup, Sociodemographic Characteristics, and Dietary Information: Case Study," vol. 8, no. 1, p. e49113, 2023.
[20] P. Rajendra, S. J. C. M. Latifi, and P. i. B. Update, "Prediction of diabetes using logistic regression and ensemble techniques," vol. 1, p. 100032, 2021.
[21] Z. Zhang *et al.*, "Machine learning prediction models for gestational diabetes mellitus: meta-analysis," vol. 24, no. 3, p. e26634, 2022.
[22] H. Nouraei, H. Nouraei, and S. W. J. B. Rabkin, "Comparison of unsupervised machine learning approaches for cluster analysis to define subgroups of heart failure with preserved ejection fraction with different outcomes," vol. 9, no. 4, p. 175, 2022.
[23] G. Al-Kharusi, N. J. Dunne, S. Little, and T. J. J. B. Levingstone, "The role of machine learning and design of experiments in the advancement of biomaterial and tissue engineering research," vol. 9, no. 10, p. 561, 2022.
[24] A. Zaitcev *et al.*, "A deep neural network application for improved prediction of $\text {HbA} _ {\text {1c}} $ in type 1 diabetes," vol. 24, no. 10, pp. 2932-2941, 2020.
[25] M. Matabuena, P. Félix, C. García-Meixide, F. J. C. M. Gude, and P. i. Biomedicine, "Kernel machine learning methods to handle missing responses with complex predictors. Application in modelling five-year glucose changes using distributional representations," vol. 221, p. 106905, 2022.
[26] S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. J. B. i. Sakharkar, "Significance of HbA1c test in diagnosis and prognosis of diabetic patients," vol. 11, p. BMI. S38440, 2016.
[27] S. A. Moamin, M. K. Abdulhameed, R. M. Al-Amri, A. D. Radhi, R. K. Naser, and L. G. Pheng, "Artificial Intelligence in Malware and Network Intrusion Detection: A Comprehensive Survey of Techniques, Datasets, Challenges, and Future Directions," *Babylonian Journal of Artificial Intelligence*, vol. 2025, pp. 77-98, 2025.
[28] M. Bakouri, H. S. Sultan, S. Samad, H. Togun, and M. Goodarzi, "Predicting thermophysical properties enhancement of metal-based phase change materials using various machine learning algorithms," *Journal of the Taiwan Institute of Chemical Engineers*, vol. 148, p. 104934, 2023.
[29] A. Ram and H. Vishwakarma, "Diabetes prediction using machine learning and data mining methods," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1116, no. 1: IOP Publishing, p. 012135.
[30] H. A. J. I. A. Al-Jamimi, "Synergistic feature engineering and ensemble learning for early chronic disease prediction," 2024.
[31] R. Al-Amri, A. Hadi, A. H. Mousa, H. Hasan, and M. Kadhim, "The development of a deep learning model for predicting stock prices," *J. Adv. Res. Appl. Sci. Eng. Technol*, vol. 31, pp. 208-219, 2023.
[32] D. F. M. Mohideen, J. S. S. Raj, R. S. P. J. B. A. o. B. Raj, and Technology, "Regression imputation and optimized Gaussian Naïve Bayes algorithm for an enhanced diabetes mellitus prediction model," vol. 64, p. e21210181, 2021.
[33] T. B. Chandra, A. S. Reddy, A. Adarsh, M. Jabbar, and B. Jyothi, "Diabetes Prediction Using Gaussian Naive Bayes and Artificial Neural Network," in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, 2024: IEEE, pp. 1-5.

[34] I. K. Harith, W. Nadir, M. S. Salah, and A. Majdi, "Estimating the joint shear strength of exterior beam–column joints using artificial neural networks via experimental results," *Innovative Infrastructure Solutions*, vol. 9, no. 2, p. 38, 2024.
[35] F. M. Okikiola, O. S. Adewale, and O. O. J. F. J. O. S. Obe, "A diabetes prediction classifier model using naive bayes algorithm," vol. 7, no. 1, pp. 253-260, 2023.
[36] A. Z. Arrayyan, H. Setiawan, and K. T. J. S. T. Putra, "Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations," vol. 27, no. 1, pp. 28-36, 2024.
[37] A. Dutta *et al.*, "Early prediction of diabetes using an ensemble of machine learning models," vol. 19, no. 19, p. 12378, 2022.
[38] A. Chandramouli, V. R. Hyma, P. S. Tanmayi, T. G. Santoshi, and B. J. E. C. Priyanka, "Diabetes prediction using hybrid bagging classifier," vol. 47, p. 100593, 2023.
[39] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. J. C. Ullah, "An improved artificial neural network model for effective diabetes prediction," vol. 2021, no. 1, p. 5525271, 2021.
[40] N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*: Elsevier, 2020, pp. 327-339.
[41] M. S. Shafeea, "The role of emerging innovations in the future of surgery and the need for medical education adaptation: a letter to the editor," *Journal of Robotic Surgery*, vol. 19, no. 1, p. 383, 2025.
[42] G. Verma, H. J. I. J. o. G. Verma, and D. Computing, "A multilayer perceptron neural network model for predicting diabetes," vol. 13, no. 1, pp. 1018-1025, 2020.
[43] H. Bani-Salameh *et al.*, "Prediction of diabetes and hypertension using multi-layer perceptron neural networks," vol. 12, no. 02, p. 2150012, 2021.
[44] N. Z. Khalaf, I. I. Al Barazanchi, A. Radhi, S. Parihar, P. Shah, and R. Sekhar, "Development of real-time threat detection systems with AI-driven cybersecurity in critical infrastructure," *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 501-513, 2025.
[45] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. J. P. C. S. Pineda, "Diabetes diagnostic prediction using vector support machines," vol. 170, pp. 376-381, 2020.
[46] R. S. Raj, D. Sanjay, M. Kusuma, and S. Sampath, "Comparison of support vector machine and Naive Bayes classifiers for predicting diabetes," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 2019: IEEE, pp. 41-45.
[47] A. J. Mohammed, M. Muhammed Hassan, D. J. I. J. o. A. T. i. C. S. Hussein Kadir, and Engineering, "Improving classification performance for a novel imbalanced medical dataset using SMOTE method," vol. 9, no. 3, pp. 3161-3172, 2020.
[48] H. Hairani, A. Anggrawan, and D. J. J. i. j. o. i. v. Priyanto, "Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link," vol. 7, no. 1, pp. 258-264, 2023.
[49] S. Almas *et al.*, "Visual impairment prevention by early detection of diabetic retinopathy based on stacked auto-encoder," vol. 15, no. 1, p. 2554, 2025.
[50] J. Xie and Q. J. I. T. o. B. E. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models," vol. 67, no. 11, pp. 3101-3124, 2020.
[51] L. M. A. Al-Huseini, N. J. Kadhim, M. S. Mahdi, R. H. Ogaili, and O. Al-Hammood, "Microbial infection disease diagnosis and treatment by artificial intelligence," *Wiad Lek*, vol. 78, no. 2, pp. 442-447, 2025.