Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00130-x

Review Article

Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights

Meenu Goyal [1], Dilbag Singh [1]

1. *Computer Science and Engineering, Chaudhary Devi Lal University, Sirsa, IND*

Received: April 28, 2026 | Review began: May 20, 2026 | Review ended: June 17, 2026 | Published: June 22, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Diabetes is a chronic metabolic disorder associated with high blood sugar that can adversely affect the eyes, heart, kidneys, and nerves of our body. Early prediction can be valuable for effective management and better treatment of the disease. Here, machine learning plays a significant role in analyzing large amounts of data and finding hidden information from it and predicting outcomes accordingly. This study reviews various classification and preprocessing techniques used for diabetes prediction in recent years. It covers machine learning methods, including decision trees, Naive Bayes, Support Vector Machine, random forests, ensemble learning methods, deep learning methods like Convolutional Neural Network and Long Short-Term Memory. It also discusses the impact of various preprocessing techniques like missing value imputation, data balancing, and feature selection on model performance. Visualization charts of related work are also presented for better understanding. Small dataset, limited features, quality of data, and lack of model interpretability are some key challenges in this area. Future research in the area should focus on improving these challenges and provide a better tool for early management and treatment of diabetes.

Categories: AI/ML-based decision support systems, Machine Learning (ML), Data Visualization

Keywords: diabetes prediction, machine learning, deep learning, ensemble learning, pre-processing techniques

Introduction And Background

Diabetes is a health issue characterized by high blood sugar (glucose) levels, with glucose originating from the carbohydrates present in our diet, acting as the primary energy source for the body. Insulin, a hormone, plays a vital role in controlling blood sugar levels within the body. The pancreas releases insulin into the bloodstream, lowering the amount of sugar. When the pancreas body does not make enough insulin or cannot use it effectively, sugar levels start increasing in your bloodstream, which is called hyperglycemia. It affects people of all ages. Diabetes is a chronic disease that can result in long-term complications, such as cardiovascular disease, nerve damage, kidney problems, vision issues, and foot-related complications [1]. In India, with a population of approx. 1.3 billion, 75 million diabetes cases are reported with a prevalence rate of 9.6%. It is also noted that 43.9 million cases are undiagnosed. The mean health expenditure on diabetes is $92 per person annually. One million deaths happen in India each year just because of diabetes. Globally, it is around 7.9 billion, and there are 537 million diabetes cases reported with a prevalence rate of 10.5% among adults. Worldwide, 232 million cases remain undiagnosed. Healthcare expenditure on diabetes varies by country. Diabetes is the cause of approx. 6.7 million deaths annually worldwide. The statistical analysis shows that non-communicable diseases like diabetes contribute significantly to the overall disease burden and mortality in both India and the world. It emphasizes the need for early detection and accurate prediction for effective management and to address the growing public health concern associated with the disease [2,3]. Diabetes is of different types - Type 1 diabetes, Type 2 diabetes, prediabetes, and gestational diabetes. Type 1 diabetes occurs mainly in infancy. In Type 1 diabetes, the body generates

no insulin or very little insulin. To control diabetes insulin injection is needed. Type 1 diabetes increases the chances of heart-related disease and stroke. The main symptoms of Type 1 diabetes include thirst, hunger, and eye-related disorders. Type 2 diabetes mainly occurs in adults, and it results when the body does not react to insulin. Type 2 diabetes also increases the chances of heart-related problems. Type 2 diabetes symptoms include high blood pressure and weight gain. Prediabetes occurs if the blood glucose level increases from the normal level, but symptoms are very low and difficult for the doctors to identify. With the help of exercise and weight loss, the risk of prediabetes can be lowered. Prediabetes increases the chances of heart-related problems and Type 2 diabetes. Gestational diabetes occurs during pregnancy in women [4]. Type 2 diabetes adds up to approximately 90% of all diabetes cases. This form of the disease is defined by persistent high blood sugar levels and the body's inability to effectively regulate glucose levels, resulting in excessive amounts of sugar in the blood stream. Type 2 diabetes is more common in older adults and is often associated with obesity or being overweight. Early detection of the disease can significantly improve the chances of successful management and recovery. Advancements in technology, including machine learning (ML) and deep learning (DL) methods have become valuable tools in the early prediction and analysis of this condition [5].

With the advancements in technology, ML plays an important role to enhance the accuracy of medical diagnoses, personalize healthcare, and find critical insights from large datasets provided by the healthcare industry. These numerous datasets are analyzed using ML algorithms to identify important patterns and forecast results. The use of ML algorithms can help healthcare providers to improve decision-making, thereby advancing the quality of healthcare. In the healthcare sector, ML can be used in many different forms [6]. ML techniques can analyze data from various sources, like electronic health records, to identify the likelihood of hospital readmission or the development of chronic conditions. It helps in the identification of patients with high risk and provides a complete plan to avoid serious issues. ML algorithms trained on medical data, such as electronic health records and medical imaging like X-rays and Computed Tomography (CT) scan, help in diagnosing the disease and deciding appropriate treatment strategies. ML techniques can be integrated with clinical decision systems that help healthcare professionals make better-informed decisions about patient care [7].

The study discusses a detailed review of numerous classification and preprocessing techniques used in diabetes prediction. Also, the impact of preprocessing techniques to improve the model performance by improving the data quality is highlighted.

The study contributes to:

• Systematic analysis of recent ML/DL-based diabetes prediction studies.

• Comparing preprocessing strategies and their impact on performance.

• Presenting visual insights to highlight research trends.

• Identifying gaps related to dataset diversity and explainability.

Methodology

The study followed a structured approach to analyze various techniques used in diabetes prediction. The study conducted a review using papers collected from Scopus (Elsevier), selected as the primary database due to its extensive multidisciplinary coverage, including indirect indexing of content from major sources such as PubMed (biomedical literature), Web of Science (multidisciplinary sciences), and IEEE Xplore (engineering and computing). Furthermore, preprocessing and classification techniques used in the paper are analyzed in detail. Finally, a comparative discussion is conducted based on the findings and key insights of the study. In the end, the work is concluded with the future scope of the study.

The study selection process is shown in Figure *1*.

Search Strategy

Primary Keywords using (OR)	**"Early diabetes prediction" OR "blood sugar prediction" "diabetes prediction"**
AND	
Secondary Keywords with (OR)	**"Machine Learning" OR "Deep learning"**

FIGURE 1: Primary and secondary keywords used in search.

Eligibility Criteria

Inclusion criteria: The inclusion criteria are as follows:

• Studies published from 2018 onwards (to capture recent advancements in ML/DL for diabetes prediction)

• Studies written in English.

• Peer-reviewed journal or conference articles (ensuring rigorous validation and reproducibility).

• Studies available in open access (for equitable full-text review and accessibility).

• Studies that focus on ML, DL, preprocessing techniques, or equivalent Artificial Intelligence (AI) methods for the prediction of diabetes mellitus (including early detection or blood sugar forecasting).

Exclusion criteria: The exclusion criteria are as follows:

• Studies published earlier than 2018.

• Studies not written in English.

• Studies other than peer-reviewed journal or conference articles (e.g., books, reviews, editorials, or preprints).

• Studies not available in open access.

• Studies that do not apply ML, DL, preprocessing techniques, or equivalent AI methods for the prediction of diabetes mellitus (e.g., non-predictive or traditional statistical analyses only).

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

Figure 2 presents the distribution of articles included in the systematic review. Out of a total of 98 references included in this study, 95 studies were considered for the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and systematic review process. The remaining three references consist of general background, including diabetes facts and Intent Detection Framework (IDF) reports, and were therefore not included in the PRISMA study count.

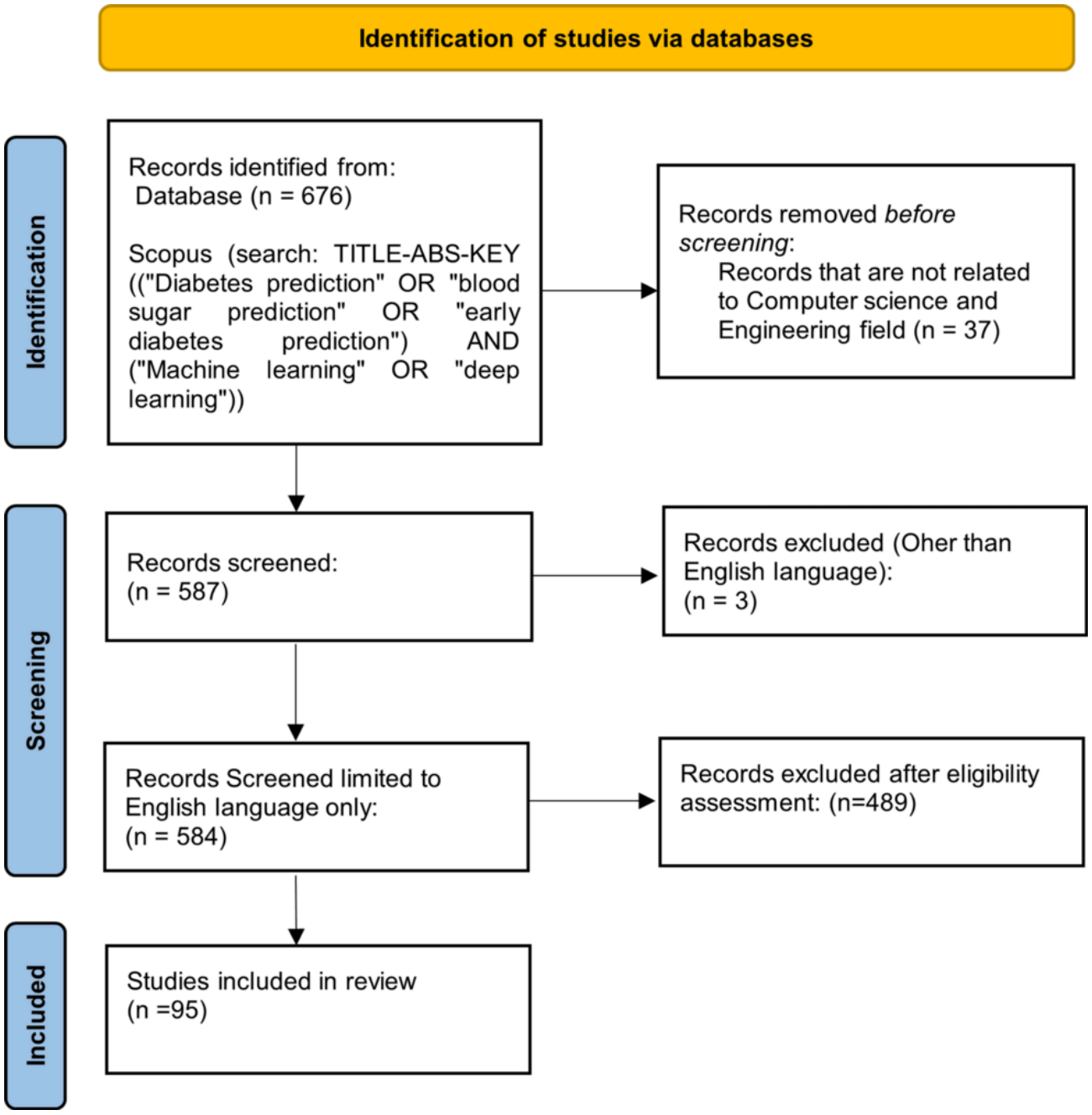


FIGURE 2: PRISMA flow diagram of study search and selection.

PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Review

Researchers developed various models using ML techniques such as logistic regression, XGBoost, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayes, and gradient boosting for diabetes prediction. Ensemble learning techniques such as bagging and boosting ensemble models are effective for the classification. DL techniques such as deep neural networks (DNN) are also used for diabetes prediction and have shown better performance. Different feature selection methods are studied to find the most significant features for diabetes prediction. The challenges related to imbalanced and missing datasets are handled using different imputation techniques and data augmentation methods, as discussed in the studies given below.

El-Sofany et al. [8] proposed a system for diabetes prediction at its early stage, using distinct ML approaches. The publicly available Pima dataset was used for the study, having 768 samples and 8 attributes. The dataset consists of missing values, and these missing values are imputed by the mean value, and it is an imbalanced dataset, having 268 positive samples and 500 negative samples. The Synthetic Minority Oversampling Technique (SMOTE) is used to handle the imbalanced dataset. The experiment was performed with and without SMOTE, and there was a large difference in the performance. With balanced data using SMOTE, it gives much better performance. Out of all the ML techniques used in the study, XG Boost performed the best. Additionally, a mobile app had been created to predict diabetes symptoms on the spot. Although the study has some limitations, such as dataset size, imbalanced classes, and selection of the best algorithm, this limitation will be addressed in future research by incorporating advanced techniques for improving accuracy and usability [8].

Tripathy et al. [9] used distinct ML and DL algorithms for diabetes prediction. The Pima dataset was used to conduct the experiment. The study then concluded that the performance of the Long Short-Term Memory (LSTM) model was better than the standard ML algorithm in the prediction of diabetes. The performance of the model can be further enhanced with correct preprocessing and increased data size.

Prabhakar et al. [10] developed a framework for diabetes prediction using a cloud-based ensemble learning technique (DPUCE). A bagging ensemble classifier with three base learners, DT, SVM, and Artificial Neural Network (ANN), and the Pima dataset were used for the study. The proposed ensemble model gives good results as compared to single classifiers. The work can be extended with multiple classes of datasets and also used image and textual datasets [10].

Alnowaiser proposed a tri-ensemble approach to diagnose the disease at its early stage that is necessary for timely treatment. In the tri-ensemble model, three different classifiers (XGB, ETC, RF) with voting mechanisms are used. For any model to achieve good performance, handling the missing values of the data is necessary. In that study, the research's main focus was on the missing values of the Pima dataset. To handle the missing value, the KNN imputation method was used. The Pima dataset with deleted missing values and with KNN-imputed missing values is trained on the proposed tri-ensemble classifier. It was found that the KNN imputation method gave better results. Furthermore, to improve the performance, explore DL techniques and use a large dataset with large dimensions [11].

El-Bashbishy & El-Bakry [12] proposed a diabetes prediction model based on DL technique. The dataset collected for the study from MUCHD Hospital consists of 548 samples and 18 attributes, and missing values of the data are replaced with the mean. Standardization and normalization methods are used for feature scaling. For feature selection, the Pearson correlation method was used and found that seven features are most significant and relevant. The Deep Neural Network (DNN)-based Multi-Layer Perceptron (MLP) method was used for prediction with batch size of 10, a 50-epoch size, and Stochastic Gradient Descent (SGD) optimizer to tune the parameters of DNN algorithm. Two activation functions, Rectified Linear Unit (ReLU) and Sigmoid, are used for 1 to 9 hidden layers. The relay activation function was used, and for the 10th hidden layer, the sigmoid function was used. The proposed model showed better performance. To achieve further improvement, the model is trained with a large dataset. The work is extended to determine the type of diabetes. It is also advised to explore the system by combining it with other medical scan images of eyes, chest, etc. [12].

Tiwaskar et al. [13] mainly focused on imputation methods for medical datasets. There are two approaches to handle the missing values in the dataset. First, eliminate the missing values, and second is to impute the missing values with some imputation techniques. The first method has some limitations, such as reduction of data size, elimination of some important row and wastage of effort, time, and cost of the data collection. The alternate approach to impute the missing value was used in that study. For imputation, three methods, KNNI, MICE, Miss Forest, were used on the Pima dataset with

complete dataset and imputed dataset. Four classifiers named SVM, RF, AdaBoost, and XGBoost and 11 evaluation metrics were used to measure the performance. It was found that the Pearson correlation coefficient (PCC) is the best criterion for finding the relation before and after imputation. The Miss Forest imputation method achieves best result with the SVM classifier.

Ma et al. [14] proposed a system to predict coronary heart disease (CHD) in diabetic patients. For that experiment, a dataset was collected from Luzhou Hospital in China. It consists of 300,089 samples with 9,851 positive samples, 290,238 negative samples, and 76 attributes. Duplicate and irrelevant records are removed from the data. Analysis of Variance (ANOVA) and IF feature selection methods are used, and the top 9 features are selected. To train the dataset, XGB, LR, and RF methods were used, and RF achieved the best performance. Findings from the study help to understand the risk factors for CHD related to diabetes. In that study, an imbalanced dataset and limited features were used; in the future, explore DL-based ensemble learning methods to overcome the problems.

Qi et al. [15] proposed an ensemble framework to predict diabetes on fusion of key features (KF Predict Framework). Pima dataset was used and the missing value of the data was imputed with the mean. For feature selection, the correlation coefficient method was used, and three features were chosen as key features and five were chosen as second features. A further multi-input-based neural network was used to fuse the key features and secondary features for classification. Finally, the soft voting ensemble classifier containing three methods (SVM, RF, and KNN) is used for prediction and achieves good performance. The study was biased towards only one gender, so use a different gender dataset and the blood glucose fluctuation problem.

Chandramouli et al. [16] proposed a hybrid bagging classifier for diabetes prediction on a Kaggle dataset. The median value is used for missing-value imputation, and optimal features are selected using the Recursive Feature Elimination (RFE) technique. The hybrid bagging classifier has the highest accuracy of 92%.

Houari et al. [17] proposed an automated diabetes prediction system with risk factor assessment using algorithms such as SVM, KNN, DT, and LR. The dataset used in the study was collected from Algeria. For data preprocessing, median values are used for missing-value imputation, and the ANOVA and RFE methods were used for optimal feature selection. It is observed that KNN method gave the best result due to its advantage of the nearest-neighbor approach.

Roobini et al. [18] proposed a diabetes prediction model and analyzed it with two datasets. Different ML classifiers, DT, NB, MLP, and RF are experimented with Pima and SYLHET datasets. It is observed that the RF classifier had the best performance.

Atif et al. [19] conducted a study to determine the best model for early diabetes prediction on the SYLHET dataset. They used different ML and ensemble learning classifiers to measure the performance. The chi-square technique was used to select the optimal features. It is observed that ensemble learning classifiers such as gradient boosting, AdaBoost, and random forests gave the best results.

Saihood & Sonuc [20] developed a diabetes prediction model based on three stages: data preprocessing, hyperparameter tuning, and ensemble learning. In data preprocessing, the mean value is used for missing-value imputation, RFE is used for optimal feature selection, SMOTE techniques are used for data balancing, and the grid search method was used for hyperparameter tuning. The experiment was performed on the Pima dataset. It is observed that the stacking ensemble method gave the best performance.

Sharma et al. [21] proposed an automated cloud-based e-health system for diabetes detection at an early stage. Extreme Learning Machine (ELM, a type of ANN) method was used for training. The Pima dataset was used for the study, and optimal features are selected by Principal Component Analysis (PCA). The experiment was conducted in a local environment as well as in a cloud-based environment. It is observed that the proposed model gave an accuracy of 90% and performed faster in a cloud-based environment.

Assegie et al. [22] developed a gestational diabetes prediction model using the KNN method. The Pima dataset was used for the study. It is observed that KNN method with grid search gave an accuracy of 82.5%.

Flores et al. [23] compared different ML approaches for early-stage diabetes prediction. Here, the classifiers that are used are RF, SVM, and NN. The SYLHET Bangladesh dataset was used for the study and relied on the method used for optimal feature selection. It is observed that RF had the best performance.

Yadav et al. [24] proposed a model for diabetes risk prediction to make clinical decision. The experiment was performed on multiple datasets using ML and ensemble learning classifiers. Optimal features are selected using Boruta approach,

SMOTE used for data balancing, and GS_GWO method was used for hyperparameter tuning. It is observed that staking classifier had the best performance.

Vijiyakumar et al. [25] proposed a model to predict diabetes using feature importance-based fusion technique. The experiment performed on a private dataset was collected from Saudi Hospital using ML and ensemble learning classifiers. Permutation, SHapley Additive exPlanation (SHAP), and Internet Gateway (IG)-based fusion method was used for features selection. It is observed that SVM and RF gave the best performance. Shimpi et al. [26] developed decision-level fusion-based hybrid classifier for diabetes prediction. Pima dataset was used for the study. Different ML classifier for training and Program-Specific Outcome (PSO) was used for hyperparameter tuning. It is observed that SVM, KNN, RF based fusion had the best performance. Li et al. [27] proposed a model to find the key influential factors of diabetes. LR model was used for the study, and the dataset was collected from China Hospital. Root-Mean-Square Error (RMSE) with LR model had the accuracy of 93%, age and blood sugar are key influencing factor. Dogru et al. [28] proposed a hybrid super ensemble model for early-stage diabetes prediction on Pima and SYLHET dataset. Chi-square approach was used for feature selection and grid search-based hyperparameter tuning was performed. It is observed that proposed Super learning model [at level 0 base learners (LR, RF, GB, DT) are used and at level 1 SVM meta learner is used] had the best performance. Zhou et al. [29] proposed BROUTE-based features selection method and ensemble learning for early diabetes prediction on Pima and SYLHET dataset. It is observed that proposed model provides best results as compared to traditional method. Zohair et al. [30] proposed diabetes prediction model in binary and multiclass classification using fusion-based approach. It is observed that stacking ensemble method had the best performance. Kakoly et al. [31] proposed diabetes prediction using ML and features selection technique on Bangladesh dataset. It is observed that LR obtained 82.2% accuracy after balancing the dataset using SMOTE and glucose is the most relevant factor. Gupta et al. [32] DL model for diabetes prediction suing Pima dataset and obtained 84.42% accuracy. Rupapara et al. [33] proposed ensemble learning model based on logistic regression and extra tree classifier for diabetes prediction on Pima dataset. Proposed model obtained 85% accuracy. AlZu'bi et al. [34] experimented with different classifier on Pima dataset and obtained highest accuracy of 82% using Adaboost technique. Priyanka et al. [35] focused on diabetes prediction using DL-based Computational Neural Network (CNN) approach with Pima dataset and it improved the performance.

Aslan & Sabanci [36] proposed DL technique for diabetes prediction. In this study, numerical Pima dataset is converted into Image dataset and CNN-SVM-based model is used for training. The proposed model obtained accuracy of 92.19%.

Alghamdi study predicts diabetes using different data mining classification techniques. Using the XGB model on the Pima dataset, it obtained 89% accuracy [37].

Healthcare Engineering used different ML classifiers on the Pima dataset. Using grid search with LR model, it obtained 86% accuracy [38].

Roobini & Lakshmi [39] used the DL-based model LSTM for diabetes prediction. Along with LSTM PCA and grid search used for feature selection and hyperparameter tuning. The proposed model reduces the execution time and achieves the best performance.

Ahamed et al. [40] proposed a diabetes prediction model with feature augmentation. The Pima dataset was used, and the random oversampling method was used and obtained 89.5% without augmentation and 92.5% with augmentation.

Naz et al. [41] proposed a hybrid model based on DT-DL and the SMOTE method for diabetes prediction on the Pima dataset. The proposed method shows better performance compared to the individual.

Alex et al. [42] focused on diabetes prediction using deep CNN. The Pima dataset was used with removed missing values and balanced using the SMOTE method. The proposed model with the Adam optimizer obtained 86.29% accuracy.

Dutta et al. [43] applied ensemble learning methods on the DDC-2011 and DDC-2017 datasets. The experiment was performed with multiple preprocessing techniques. The proposed ensemble model with grid search optimization obtained 73.5% accuracy.

Jagini et al. [44] used ANN on Pima dataset and obtained 86.2% accuracy with 20 neurons in two hidden layers. Li et al. [45] focused on the ML and neural network method for diabetes prediction using the Pima dataset. Genetic algorithm used for feature selection. The neural network with two hidden layers obtained the highest accuracy of 88.6%.

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

Model	References
LR	[27, 31,38,46,47,48,49]
DT	[50,51]
SVM	[13,25,36,52,53,54,55,56]
KNN	[17,22]
RF	[14,18,23,25,57,58,59,60,61,62,63,64,65]
Ensemble Learning (Bagging, Boosting, Stacking, XGB, GB, SGB, LGB, AdaBoost)	[8,10,11,15,16,19,20,24,26,28,29,30,34,37,40,43,45,66,33,67,68,69,70,71,72,73,74,75,76,77,78 ,79,80,81,82,83,84]
ANN	[21,44,85,86]
MLP	[12,87,88]
DNN	[32,57,89,90,91,92,93,94]
CNN	[35,36,42,95,96]
LSTM	[9,39,97]
DT-DL	[41]

TABLE 1: Different classifiers used for early diabetes prediction.

LR, Logistic Regression; DT, Decision Tree; SVM, Support Vector Machine; KNN, K-Nearest Neighbor; SGB, Stochastic Gradient Boosting; LGB, Light Gradient Boosting; ANN, Artificial Neural Network; DNN, Deep Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; DT-DL, Decision Tree-Deep Learning

Figure *3* shows the frequency of different classification techniques used in diabetes prediction studies.

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

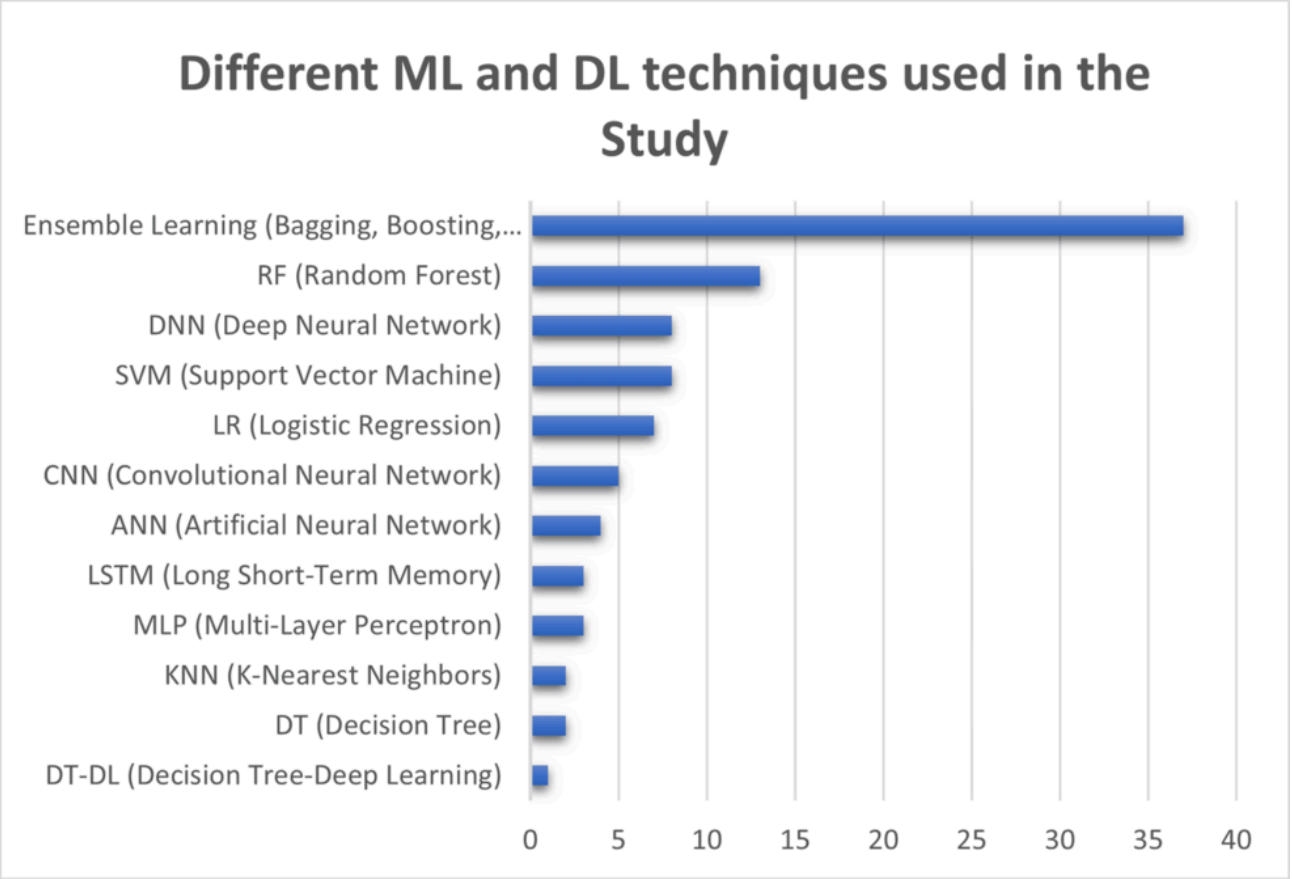


FIGURE 3: Frequency of different classification techniques used in the studies.

LR, Logistic Regression; DT, Decision Tree; SVM, Support Vector Machine; KNN, K-Nearest Neighbor; ANN, Artificial Neural Network; DNN, Deep Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; DT-DL, Decision Tree-Deep Learning

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

Ref.	Dataset	Missing value	Data balancing	Feature selection
[8]	PIDD	MEAN	SMOTE	-
[11]	PIDD	KNN IMPUTER	-	-
[12]	MUCHD	MEAN	-	PEARSON CORRELATION
[13]	PIDD	KNNI, MICR, MISS FOREST	-	-
[14]	LUZHOU HOSPITAL IN CHINA	-	-	ANOVA, IFS
[15]	KAGGLE	MEDIAN VALUE	-	RPE
[16]	ALGERIA DATASET	MEDIAN VALUE	-	ANOVA, RFE
[17]	SYLHET BANGLADESH	-	-	CHI-SQUARE
[19]	PIDD	MEAN	SMOTE	RFE
[20]	PIDD	-	-	PCA
[21]	SYLHET BANGLADESH	-	-	RELIEF
[23]	PIMA, ADRC, FHD	-	SMOTE	BROUTE
[24]	SAUDI HOSPITAL	-	-	FUSION OF PERMUTATION, SHAP AND IG METHODS
[25]	CHINA HOSPITAL	-	-	RMSE
[27]	PIDD SYLHET BANGLADESH DIABETES2	-	-	CHI-SQUARE
[28]	PIDD, FRANKFURT GERMANY	AVERAGE	-	BROUTE
[30]	BANGLADESH QUESTIONNARE	-	SMOTE	PCA
[31]	PIDD	-	-	PCA + IG
[32]	PIDD	-	-	RELIEF

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

[33]	PIDD	MEAN	SMOTE	PCA, RFE, RELIEF
[36]	PIDD	-	-	PCA
[39]	PIDD, DMS (SURVEY)	-	OVERSAMPLING	-
[40]	PIDD	-	SMOTE	-
[41]	PIDD	-	SMOTE	-
[42]	DDC-2011 DDC-2017	MEDIAN/KNN IMPUTATION	-	RF, IG, XGB, LGB
[43]	PRIVATE DATASET	-	-	CHI-SQUARE, PCA
[45]	PIDD	-	-	FTWGA-FSA
[49]	PIDD	MEDIAN	-	RF AND XG BOOST
[50]	UCI	-	-	PCA, ICA, CORRELATION BASED
[51]	KAU QUESTIONARE	MEAN	-	PCA, ICA AND CORRELATION METHOD
[53]	PIDD	-	-	CORRELATION METHOD
[57]	PIDD	MEAN	-	CORRELATION COEFFICIENT
[60]	PIDD	MEAN	-	CORRELATION BASED, PCA, IG
[64]	NHANES(US)	-	-	LR
[67]	PIDD	-	-	FTR-FOM
[69]	SYLHET BANGLADESH	MEAN/MEDIAN	RESAMPLING	-
[72]	PIDD,130-US	MEDIAN	-	GA
[73]	PIDD	MEDIAN	-	-
[75]	CPWCS	-	SMOTE	RFECV
[77]	PIDD	-	SMOTE	-
[79]	PIDD, MANDELA	MEDIAN	-	RFE
[82]	PIDD	MEDIAN	-	-

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

[83]	SYLHET BANGLADESH	KNN IMPUTER	-	-
[84]	PIDD	-	-	GWO, APGWO
[85]	PIDD	MEDIAN	-	-
[86]	PIDD	MEAN	-	PEARSON CORRELATION
[88]	FIMMG	-	-	APCA
[93]	PIDD	-	-	FEATURE IMPORTANT SCORE
[94]	NHANES	-	-	BIVARIATE ANALYSIS USING LOGISTIC REGRESSION
[95]	PIDD	MEAN	-	-
[96]	PIDD	-	-	SAE, VAE
[97]	CGMS	-	SMOTE	INFORMATION GAIN
[98]	PIDD, SYLHET BANGLADESH	-	SMOTE	T-DISTRIBUTED SNE

TABLE 2: Different preprocessing techniques used in the studies.

LR, Logistic Regression; LGB, Light Gradient Boosting; SMOTE, Synthetic Minority Oversampling Technique; ANOVA, Analysis of Variance; PCA, Principal Component Analysis; RFE, Recursive Feature Elimination; RPE, Reward Prediction Errors; SHAP, SHapley Additive exPlanations; RMSE, Root-Mean-Square Error; RF, Random Forest; ICA, Independent Component Analysis; SAE, Sparse Autoencoders; VAE, Variational Autoencoder; SNE, Stochastic Neighbor Embedding

Figure *4* shows datasets and preprocessing techniques used in the reviewed literature.

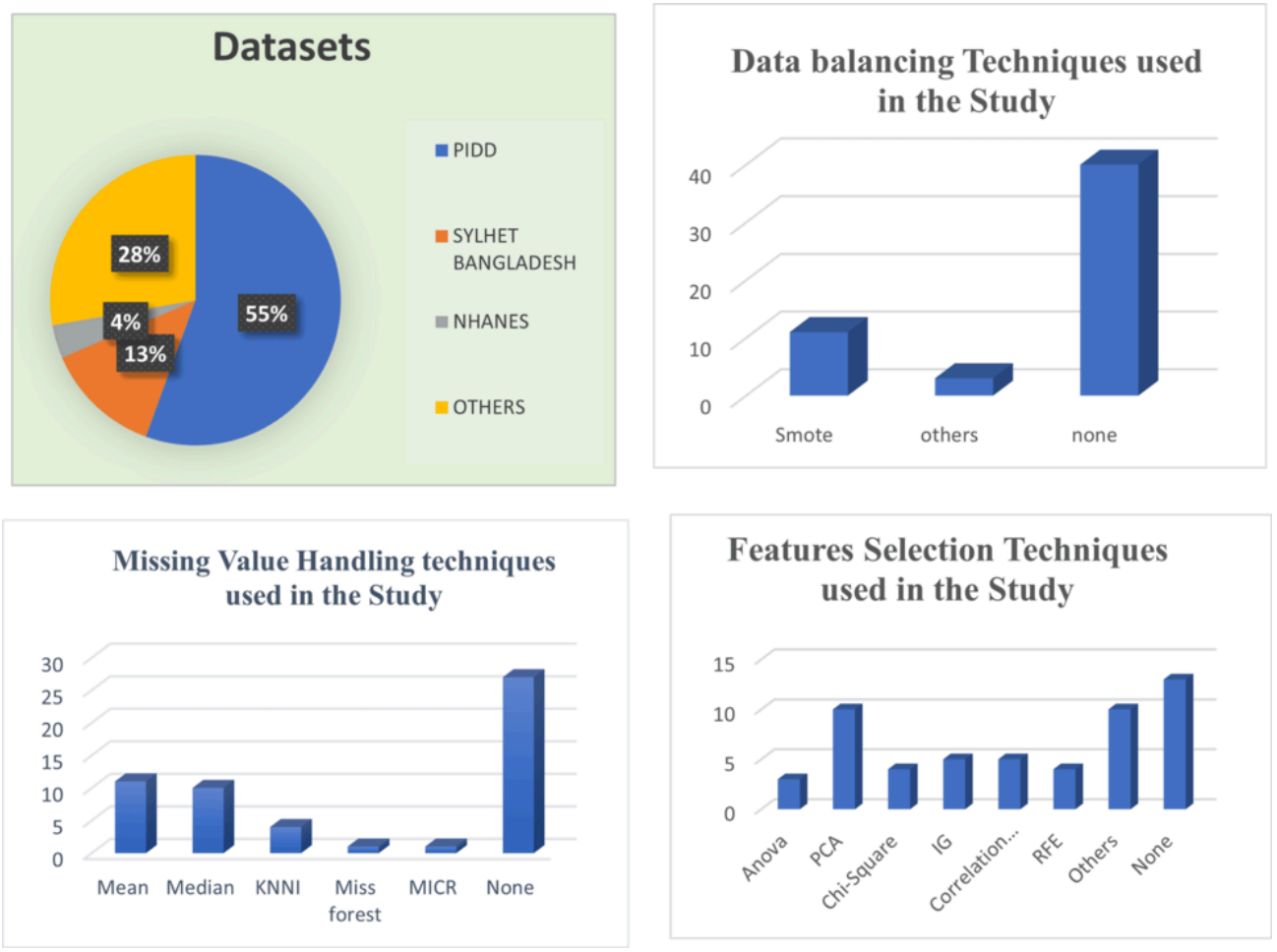


FIGURE 4: Dataset and different preprocessing techniques used in literature.

ANOVA, Analysis of Variance; SMOTE, Synthetic Minority Oversampling Technique; RFE, Recursive Feature Elimination; PCA, Principal Component Analysis

Discussion

Dataset

The majority of studies in the field of diabetes prediction have utilized the publicly available Pima Indian Diabetes Dataset. It was referred to as PIMA or PIDD. The Pima Indian repository dataset is one of the oldest datasets available in the repository. It contains 768 samples; out of it, 268 are positive samples and 500 are negative samples. It consists of eight variables that are used to predict diabetes, including the number of pregnancies (PRG), triceps skin-fold thickness (thickness), insulin level (insulin), family history (pedigree), diastolic blood pressure (BP), plasma glucose concentration after 2 hours of oral glucose tolerance test (OGTT; plasma), age, and body mass index (BMI; body), and one output variable, diabetes (yes or no). The sample size of the Pima dataset is very small. Along with it, some researchers use different private and public datasets.

Data quality issues: raw datasets face several challenges that should be addressed before applying machine learning classifiers for disease prediction. And to solve the challenges, different techniques are used in the literature as discussed further.

Missing Value

Missing values are simply the data that have no value, and they are denoted by "N/A" or null. Sometimes missing values are caused by human error during data entry, or the patient does not provide that information. Large numbers of missing values in datasets affect the performance of the model. It is important to handle the missing data effectively. One way to handle the missing values is to simply remove the missing data, but in most cases that is not suitable. There is another method, the imputation process, to impute the missing value. In literature, several imputation techniques are used to impute the missing data. In literature, different imputation techniques are used to handle missing values, like mean [8,12], median [16,17], and KNN imputation [11]. The study [11] used the KNN imputation method and improved the accuracy by 7%.

Data Balancing

When the samples in different classes are uneven, the problem will occur. In the medical dataset, mostly healthy patients are more than unhealthy patients, and it is divided into two classes: minority class and majority class. It affects the model's performance as the model performs poorly on the minority class and is biased towards the majority class. In literature, to handle imbalanced data, the SMOTE technique is used as given in El-Sofany et al., Saihood & Sonuç, and Kakoly et al. [8,20,31], and the El-Sofany et al. study [8] showed significant improvement in performance after balancing the dataset, as shown in Figure *5*.

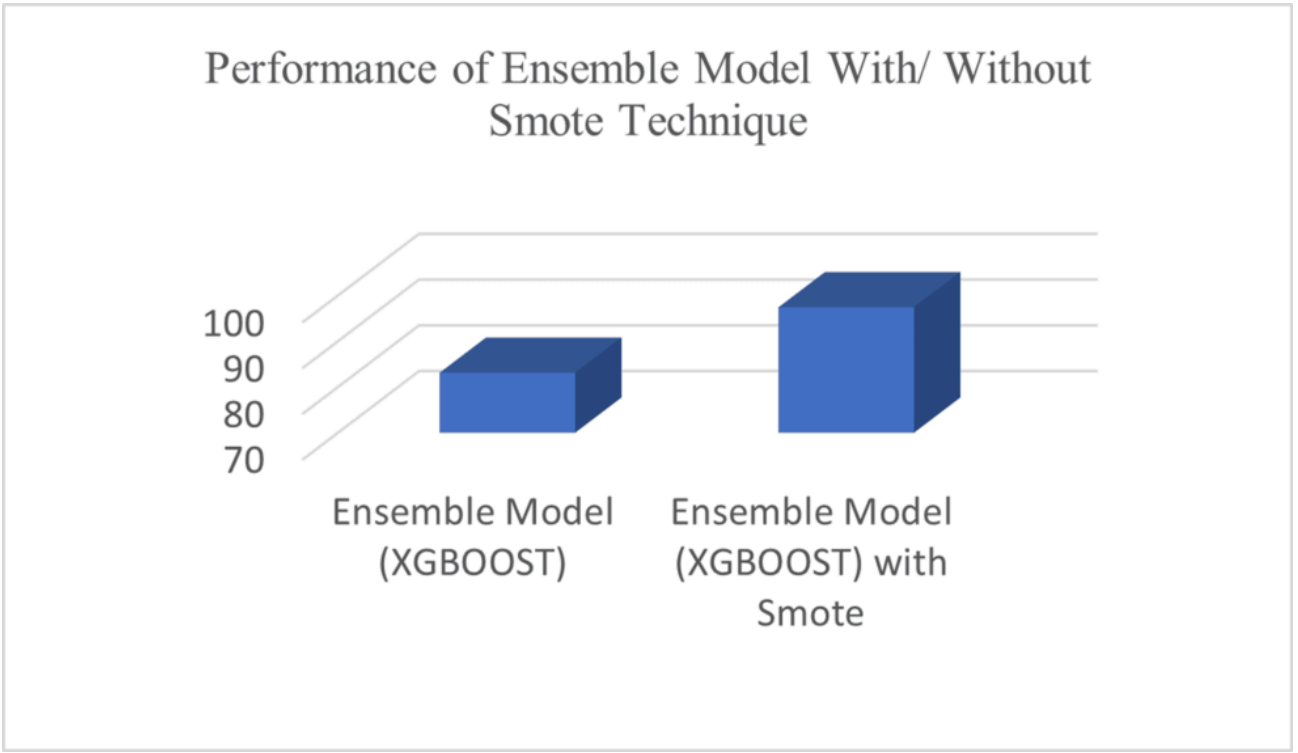


FIGURE 5: Ensemble model performance with or without missing-value imputation in literature.

Feature Selection and Extraction

Feature engineering is a process to find out the most relevant features that contribute to the prediction and remove irrelevant, redundant, or noisy features that are not contributing to the prediction process. It simplifies the model for interpretation and reduces the training time. In literature, different feature selection techniques are used, such as chi-square, ANOVA, correlation analysis, RFE, etc. Feature extraction involves converting the original set of attributes into a new set to reduce the dimensionality of the data. In literature, various techniques were used for feature selection. Chandramouli et al. [16] used the RFE method to select a maximum of nine features for training. Houari et al. [17] used

ANOVA and RFE methods to choose relevant features for the study and selected nine optimal features based on the Pima dataset. There is a significant improvement in the performance after applying feature selection. The performance of KNN was 92.80% (ACC), whereas with ANOVA it was 97.60% and with RFE it was 95.2%. Atif et al. [19] used the Chi-square feature selection method to remove two features with the least Chi-square value. Sharma et al. [21] used PCA and selected five optimal features. Assegie et al. [23] used the relief-based feature selection method, and one attribute weakness was eliminated because of unlikely symptoms of diabetes, affecting the results. Doğru et al. [28] selected five and nine features using the Chi-square value method.

Algorithm Selection

There is a need to improve the performance of an algorithm for better diabetes prediction and monitoring. Thus, optimization techniques are an essential part of the working of ML algorithms. Optimization techniques are important for tuning parameters of a model in a manner to minimize the loss function. A loss function is a way of measuring how good the model performs by determining the difference between the predicted and actual values. Hence, it is an iterative approach to adjusting the parameters of fine-tuning the model. Suitable parameter adjustments ultimately provide better cumulative performance. El-Sofany et al. [8] used the ensemble learning technique XGBoost, Alnowaiser [11] used a tri-ensemble classifier, Saihood & Sonuç [20] used a stacked model, FLores et al. [23] applied the RF method, and in Verma et al. [67], the gradient boosting approach outperforms traditional ML algorithms. In addition, DL techniques were used. Tripathy et al. and Roobini & Lakshmi [9,39] used the LSTM, El-Bashbishy & El-Bakry study [12] used the MLP method, and Priyanka et al. [35] used the CNN method and showed better performance compared to traditional machine learning techniques. For hyperparameter tuning, different optimizer techniques were used. El-Bashbishy & El-Bakrystudy and Singh et al. [12,78] used the SGD optimizer, in Kakoly et al. and Saihood & Sonuç [20,31] grid search-based methods were used, in Dutta et al. and Patro et al. [95,43] the Adam optimizer was used, and in Joseph et al. [98] the AdaGrad optimizer was used to adjust the hyperparameters effectively. The frequency of the best-performing models is shown in Figure *6*.

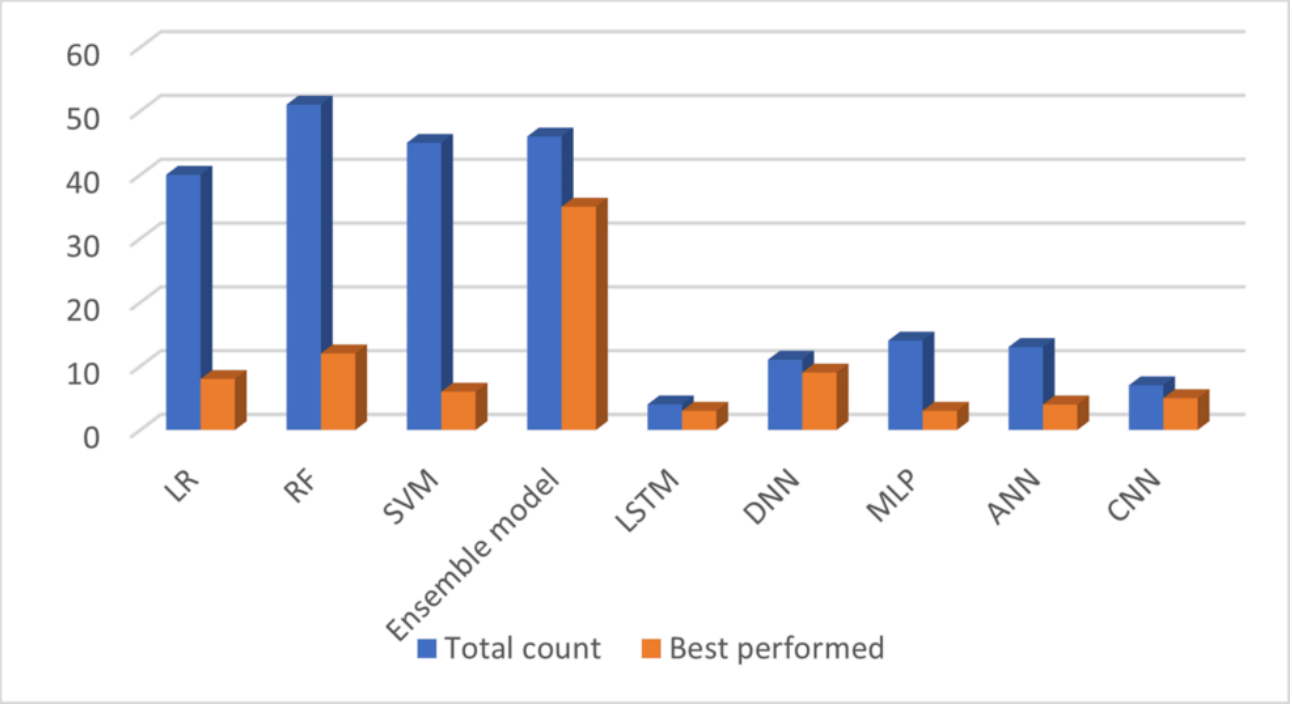


FIGURE 6: Frequency of best performed model in literature.

LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; ANN, Artificial Neural Network; DNN, Deep Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; MLP, Multi-Layer Perceptron

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

Conclusions

A detailed review of diabetes prediction using various ML methods has been analyzed in the study. Literature provides different ML, ensemble learning, and DL methods. Ensemble learning techniques have shown promising results, as they combine multiple base learners that overcome the weaknesses of single classifiers and provide good results. DL techniques also perform better. Secondly, the quality of the dataset is a major factor in achieving the best results with the minimum prediction error. It is found from the literature that by imputing missing values, accuracy increased by 7%, and with balancing the data, performance improved by 14%. Thus, data quality plays a major role in model performance. To reduce the dimensionality of the dataset, various feature selection techniques are used to select the most relevant features and remove others on the basis of the respective technique used in the study. There are several challenges that should be focused on in future studies.

Future work should prioritize developing a more generalized and robust model that can perform well on multi-center datasets. Use advanced techniques like ensemble methods with deep learning methods as a base learner and explainable AI to enhance the performance and interpretability of the model. The black-box nature of the AI system affects transparency, interpretability, and trust in the model. These models just provide the predictions but do not give the reasoning behind the prediction that creates a gap between the machine and humans. To develop a user-oriented system, integration of the model with explainable AI is needed in critical fields like healthcare, where each decision should be taken very precisely. By addressing these challenges, the diabetes prediction field can progress in creating models that not only provide accurate predictions but also provide insights that help in effective disease management and prevention and improve the overall healthcare sector.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Meenu Goyal, Dilbag Singh

Acquisition, analysis, or interpretation of data: Meenu Goyal, Dilbag Singh

Drafting of the manuscript: Meenu Goyal, Dilbag Singh

Critical review of the manuscript for important intellectual content: Meenu Goyal, Dilbag Singh

Supervision: Dilbag Singh

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

References

1. What is Diabetes | International Federation of Diabetes [Internet]. (2024). Accessed: April 28, 2026: https://idf.org/about-diabetes/what-is-diabetes/.
2. Kumar A, Gangwar R, Zargar AA, Kumar R, Sharma A: Prevalence of diabetes in India: a review of IDF Diabetes Atlas 10th Edition. Current Diabetes Reviews. 2024, 20:1-10. 10.2174/1573399819666230413094200
3. Diabetes Facts and Figures. International Diabetes Federation. (2024). Accessed: April 28, 2026: https://idf.org/about-diabetes/diabetes-facts-figures/.
4. American Diabetes Association: 2. Classification and diagnosis of diabetes: standards of medical care in Diabetes-2021. Diabetes Care. 2020, 44:S15-S33. 10.2337/dc21-s002
5. Pradeepa R, Mohan V: Epidemiology of type 2 diabetes in India. Indian Journal of Ophthalmology. 2021, 69:2932-2938. 10.4103/ijo.ijo_1627_21
6. Javaid M, Haleem A, Pratap RS, Suman R, Rab S: Significance of machine learning in healthcare: features, pillars and applications. International Journal of Intelligent Networks. 2022, 3:58-73. 10.1016/j.ijin.2022.05.002
7. Swapna G, Vinayakumar R, Soman KP: Diabetes detection using deep learning algorithms. ICT Express. 2018, 4:243-246. 10.1016/j.icte.2018.10.005
8. El-Sofany H, El-Seoud SA, Karam OH, El-Latif YMA, Taj-Eddin IATF: A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. International Journal of Intelligent Systems. 2024, 2024:1-13. 10.1155/2024/6688934
9. Tripathy N, Moharana B, Balabantaray SK, Nayak SK, Pati A, Panigrahi AA: Comparative analysis of diabetes prediction using machine learning and deep learning algorithms in healthcare. International Conference on Advancements in Smart, Secure and Intelligent Computing. 2024, 1-6.
10. Prabhakar G, Chintala VR, Reddy T, Ruchitha T: User-cloud-based ensemble framework for type-2 diabetes prediction with diet plan suggestion. e-Prime - Advances in Electrical Engineering, Electronics and Energy. 2024, 7:1-10. 10.1016/j.prime.2024.100423
11. Alnowaiser K: Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. IEEE Access. 2024, 12:16783-16793. 10.1109/access.2024.3359760
12. El-Bashbishy AES, El-Bakry HM: Pediatric diabetes prediction using deep learning. Scientific Reports. 2024, 14:1-20. 10.1038/s41598-024-51438-4
13. Tiwaskar S, Rashid M, Gokhale P: Impact of machine learning-based imputation techniques on medical datasets- a comparative analysis. Multimedia Tools and Applications. 2024, 84:5905-5925. 10.1007/s11042-024-19103-0
14. Ma CY, Luo YM, Zhang TY, et al.: Predicting coronary heart disease in Chinese diabetics using machine learning. Computers in Biology and Medicine. 2024, 169:107952. 10.1016/j.compbiomed.2024.107952
15. Qi H, Song X, Liu S, Zhang Y, Wong KKL: KFPredict: an ensemble learning prediction framework for diabetes based on fusion of key features. Computer Methods and Programs in Biomedicine. 2023, 231:107378. 10.1016/j.cmpb.2023.107378
16. Chandramouli A, Hyma VR, Tanmayi PS, Santoshi TG, Priyanka B: Diabetes prediction using hybrid bagging classifier. Entertainment Computing. 2023, 47:100593. 10.1016/j.entcom.2023.100593
17. Houari N, Reguieg H, Bachiri C, Alioua M: Automated diabetes disease prediction system based on risk factors assessment: taking charge of your health. Jordanian Journal of Computers and Information Technology. 2023, 9:377-394. 10.5455/jjcit.71-1694107771
18. Roobini MS, Lakshmi M, Rajalakshmi R, Sujihelen L, Babu K: RETRACTED ARTICLE: Type 2 diabetes mellitus classification using predictive supervised learning model. Soft Computing. 2023, N/A:1-15. 10.1007/s00500-023-08726-4
19. Atif M, Anwer F, Talib F, Alam R, Masood F: Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage. IAES International Journal of Artificial Intelligence. 2023, 12:1302-1311. 10.11591/ijai.v12.i3.pp1302-1311
20. Saihood Q, Sonuç E: A practical framework for early detection of diabetes using ensemble machine learning models. Turkish Journal of Electrical Engineering and Computer Sciences. 2023, 31:722-738. 10.55730/1300-0632.4013
21. Sharma SK, Zamani AT, Abdelsalam A, Muduli D, Alabrah AA, Parveen N, Alanazi SM: A diabetes monitoring system and health-medical service composition model in cloud environment. IEEE Access. 2023, 11:32804-32819. 10.1109/access.2023.3258549

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

22. Assegie TA, Suresh T, Purushothaman R, Ganesan S, Kumar NK: Early prediction of gestational diabetes with parameter-tuned K-nearest neighbor classifier. Journal of Robotics and Control. 2023, 4:452-457. 10.18196/jrc.v4i4.18412

23. Flores L, Hernandez RM, Macatangay LH, Garcia SMG, Melo JR: Comparative analysis in the prediction of early-stage diabetes using multiple machine learning techniques. Indonesian Journal of Electrical Engineering and Computer Science. 2023, 32:887-899.

24. Yadav P, Sharma SC, Mahadeva R, Patole SP: Exploring hyper-parameters and feature selection for predicting non-communicable chronic disease using stacking classifier. IEEE Access. 2023, 11:80030-80055. 10.1109/access.2023.3299332

25. Vijiyakumar K, Lavanya B, Nirmala I, Caroline SS: Random forest algorithm for the prediction of diabetes. IEEE International Conference on System, Computation, Automation and Networking. 2019, 1-5. 10.1109/ICSCAN.2019.8878802

26. Shimpi JK, Shanmugam P, Stonier AA: Analytical model to predict diabetic patients using an optimized hybrid classifier. Soft Computing. 2023, 28:1883-1892. 10.1007/s00500-023-09487-w

27. Li Z, Pang S, Qu H, Lian W: Logistic regression prediction models and key influencing factors analysis of diabetes based on algorithm design. Neural Computing and Applications. 2023, 35:25249-25261. 10.1007/s00521-023-08447-7

28. Doğru A, Buyrukoğlu S, Arı M: A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing. 2023, 61:785-797. 10.1007/s11517-022-02749-z

29. Zhou H, Xin Y, Li S: A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics. 2023, 24:1-34. 10.1186/s12859-023-05300-5

30. Zohair M, Chandra R, Tiwari S, Agarwal S: A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification. International Journal of Information Technology. 2023, 16:1955-1965. 10.1007/s41870-023-01463-9

31. Kakoly IJ, Hoque MR, Hasan N: Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. Sustainability. 2023, 15:1-15. 10.3390/su15064930

32. Gupta N, Kaushik B, Rahmani MKI, Anwar LS: Performance evaluation of deep dense layer neural network for diabetes prediction. Computers, Materials and Continua. 2023, 76:347-366.

33. Rupapara V, Rustam F, Ishaq A, Lee E, Ashraf I: Chi-square and PCA based feature selection for diabetes detection with ensemble classifier. Intelligent Automation and Soft Computing. 2023, 36:1931-1949.

34. AlZu'bi S, Elbes M, Mughaid A, Bdair N, Abualigah L, Forestiero A, Zitar RA: Diabetes monitoring system in smart health cities based on big data intelligence. Future Internet. 2023, 15:1-17. 10.3390/fi15020085

35. Priyanka S, Kavitha C, Kumar MP: Learning based approach for prediction of diabetes. International Conference for Innovation in Technology. 2023, 1-6. 10.1109/INOCON57975.2023.10101241

36. Aslan MF, Sabanci K: A novel proposal for deep learning-based diabetes prediction: converting clinical data to image data. Diagnostics. 2023, 13:1-19. 10.3390/diagnostics13040796

37. Alghamdi T: Prediction of diabetes complications using computational intelligence techniques. Applied Sciences. 2023, 13:1-17. 10.3390/app13053030

38. Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B: Retracted: A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering. 2023, 2023:1. 10.1155/2023/9872970

39. Roobini MS, Lakshmi M: Autonomous prediction of type 2 diabetes with high impact of glucose level. Computers & Electrical Engineering. 2022, 101:108082. 10.1016/j.compeleceng.2022.108082

40. Ahamed BS, Arya MS, Nancy AOV: Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation. Advances in Human-Computer Interaction. 2022, 2022:1-14. 10.1155/2022/9220560

41. Naz H, Nijhawan R, Ahuja NJ: DT-DL based hybrid approach for early detection of diabetes using PIMA dataset. International Conference on Reliability, Infocom Technologies and Optimization. 2022, 1-6. 10.1109/ICRITO56286.2022.9964904

42. Alex SA, Nayahi JJV, Shine H, Gopirekha V: RETRACTED ARTICLE: Deep convolutional neural network for diabetes mellitus prediction. Neural Computing and Applications. 2021, 34:1319-1327. 10.1007/s00521-021-06431-7

43. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, Meshref H: Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health. 2022, 19:1-25. 10.3390/ijerph191912378

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

44. Jagini A, Gayatri N, Nagaraju S, Rao VS, Rao KS, Sumanth B: Prediction of diabetes at early stages using ANN & outlier exposure. International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems. 2022, 1-6. 10.1109/ICSES55317.2022.9914091

45. Li J, Yuan P, Hu X, et al.: A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. Journal of Biomedical Informatics. 2021, 115:1-13. 10.1016/j.jbi.2021.103693

46. Ahamed BS, Arya MS, Sangeetha SKB, Osvin NVA: Diabetes mellitus disease prediction and type classification involving predictive modeling using machine learning techniques and classifiers. Applied Computational Intelligence and Soft Computing. 2022, 2022:1-11. 10.1155/2022/7899364

47. Mounika V, Neeli DS, Sree GS, Mourya P, Babu MA: Prediction of type-2 diabetes using machine learning algorithms. International Conference on Artificial Intelligence and Smart Systems. 2021, 127-131. 10.1109/ICAIS50930.2021.9395985

48. Joshi RD, Dhakal CK: Predicting type 2 diabetes using logistic regression and machine learning approaches. International Journal of Environmental Research and Public Health. 2021, 18:1-17. 10.3390/ijerph18147346

49. Albahli S: Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. Journal of Medical Imaging and Health Informatics. 2020, 10:1069-1075. 10.1166/jmihi.2020.3000

50. Bhargavi VS, Madheswari SP: IoT enabled healthcare system for predicting the diseases using feature optimization, decision tree, neural network, and fuzzy temporal rules. Concurrency and Computation: Practice and Experience. 2022, 34:e7327. 10.1002/cpe.7327

51. Syed AH, Khan T: Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: a retrospective cross-sectional study. IEEE Access. 2020, 8:199539-199561. 10.1109/access.2020.3035026

52. Mallika C, Selvamuthukumaran S: Hybrid online model for predicting diabetes mellitus. Intelligent Automation and Soft Computing. 2022, 31:1873-1885.

53. Ahmed N, Ahammed R, Islam MM, Uddin MA, Akhter A, Talukder MA, Paul BK: Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering. 2021, 2:229-241. 10.1016/j.ijcce.2021.12.001

54. Chatrati SP, Hossain G, Goyal A, Bhan A, Bhattacharya S, Gaurav D, Tiwari SM: Smart home health monitoring system for predicting type 2 diabetes and hypertension. Journal of King Saud University Computer and Information Sciences. 2022, 34:862-870. 10.1016/j.jksuci.2020.01.010

55. Mohan N, Jain V: Performance analysis of support vector machine in diabetes prediction. 4th International Conference on Electronics, Communication and Aerospace Technology. 2020, 1-3. 10.1109/ICECA49313.2020.9297411

56. Sneha N, Gangil T: Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data. 2019, 6:1-19. 10.1186/s40537-019-0175-6

57. Wee BF, Sivakumar S, Lim KH, Wong WK, Juwono FH: Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications. 2023, 83:24153-24185. 10.1007/s11042-023-16407-5

58. Azbeg K, Boudhane M, Ouchetto O, Andaloussi SJ: Diabetes emergency cases identification based on a statistical predictive model. Journal of Big Data. 2022, 9:1-25. 10.1186/s40537-022-00582-7

59. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH: Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district Bandipora. Computational Intelligence and Neuroscience. 2022, 2022:1-12. 10.1155/2022/2789760

60. Saxena R, Sharma SK, Gupta M, Sampada GC: A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Computational Intelligence and Neuroscience. 2022, 2022:1-11. 10.1155/2022/3820360

61. Samet S, Laouar MR, Bendib I, Eom S: Analysis and prediction of diabetes disease using machine learning methods. International Journal of Decision Support System Technology. 2022, 14:1-19. 10.4018/ijdsst.303943

62. Ferdousi R, Hossain MA, El Saddik A: Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access. 2021, 9:96823-96837. 10.1109/access.2021.3094063

63. Jackins V, Vimal S, Kaliappan M, Lee MY: AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 2020, 77:5198-5219. 10.1007/s11227-020-03481-x

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

64. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM: Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems. 2020, 8:1-14. 10.1007/s13755-019-0095-z

65. Wu CC, Yeh WC, Hsu WD, et al.: Prediction of fatty liver disease using machine learning algorithms. Computer Methods and Programs in Biomedicine. 2019, 170:23-29. 10.1016/j.cmpb.2018.12.032

66. Modak SKS, Jha VK: RETRACTED ARTICLE: Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications. 2023, 83:38523-38549. 10.1007/s11042-023-16745-4

67. Verma N, Singh S, Prasad D: Machine learning and IoT-based model for patient monitoring and early prediction of diabetes. Concurrency and Computation: Practice and Experience. 2022, 34:e7219. 10.1002/cpe.7219

68. Kumar M, Bajaj K, Sharma B, Narang S: A comparative performance assessment of optimized multilevel ensemble learning model with existing classifier models. Big Data. 2021, 10:371-387. 10.1089/big.2021.0257

69. Ahmed U, Issa GF, Khan MA, et al.: Prediction of diabetes empowered with fused machine learning. IEEE Access. 2022, 10:8529-8538. 10.1109/access.2022.3142097

70. Nanda R, Nath A, Patel S, Mohapatra E: Machine learning algorithm to evaluate risk factors of diabetic foot ulcers and its severity. Medical & Biological Engineering & Computing. 2022, 60:2349-2357. 10.1007/s11517-022-02617-w

71. Singh S, Vazirani V: Classification vs clustering: ways for diabetes detection. 7th International Conference for Convergence in Technology. 2022, 1-8. 10.1109/I2CT54291.2022.9825053

72. Abdollahi J, Nouri-Moghaddam B: Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran Journal of Computer Science. 2022, 5:205-220. 10.1007/s42044-022-00100-1

73. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J: An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors. 2022, 22:1-37. 10.3390/s22197268

74. Padhy S, Dash S, Routray S, Ahmad S, Nazeer J, Alam A: IoT-based hybrid ensemble machine learning model for efficient diabetes mellitus prediction. Computational Intelligence and Neuroscience. 2022, 2022:1-11. 10.1155/2022/2389636

75. Liu R, Zhan Y, Liu X, et al.: Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: a prospective cohort study. Journal of Healthcare Engineering. 2022, 2022:1-14. 10.1155/2022/8948082

76. Zhang J, Wang F: Prediction of gestational diabetes mellitus under cascade and ensemble learning algorithm. Computational Intelligence and Neuroscience. 2022, 2022:1-9. 10.1155/2022/3212738

77. Jaiswal S, Gupta P: Ensemble approach: XGBoost, CATBoost, and LightGBM for diabetes mellitus risk prediction. Second International Conference on Computer Science, Engineering and Applications. 2022, 1-6. 10.1109/ICCSEA54677.2022.9936130

78. Singh A, Dhillon A, Kumar N, Hossain MS, Muhammad G, Kumar M: eDiaPredict: an ensemble-based framework for diabetes prediction. ACM Transactions on Multimedia Computing, Communications, and Applications. 2021, 17:1-26. 10.1145/3415155

79. Nuankaew P, Chaising S, Temdee P: Average weighted objective distance-based method for type 2 diabetes prediction. IEEE Access. 2021, 9:137015-137028. 10.1109/access.2021.3117269

80. Das SK, Roy P, Mishra AK: Oversample-select-tune: a machine learning pipeline for improving diabetes identification. Concurrency and Computation: Practice and Experience. 2021, 34:e6741. 10.1002/cpe.6741

81. Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, Moustakas K: Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access. 2021, 9:103737-103757. 10.1109/access.2021.3098691

82. Kumari S, Kumar D, Mittal M: An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering. 2021, 2:40-46. 10.1016/j.ijcce.2021.01.001

83. Bernardini M, Morettini M, Romeo L, Frontoni E, Burattini L: Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach. Artificial Intelligence in Medicine. 2020, 105:101847. 10.1016/j.artmed.2020.101847

84. Hasan MK, Alam MA, Das D, Hossain E, Hasan M: Diabetes prediction using ensembling of different machine learning classifiers. ieee access : practical innovations, open solutions. 2020, 8:76516-76531. 10.1109/access.2020.2989857

85. Vedaraj M, Ezhumalai P: A secure IoT-cloud based healthcare system for disease classification using neural network. Computer Systems Science and Engineering. 2022, 41:95-108.

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x

86. Khanam JJ, Foo SY: A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021, 7:432-439. 10.1016/j.icte.2021.02.004

87. Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR: Machine learning based diabetes classification and prediction for healthcare applications. Journal of Healthcare Engineering. 2021, 2021:1-17. 10.1155/2021/9930985

88. Le TM, Vo TM, Pham TN, Dao SVT: A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access. 2021, 9:7869-7884. 10.1109/access.2020.3047942

89. Thasil MA, Santhoshkumar S: Deep learning based process analytics model for predicting type 2 diabetes mellitus. Computer Systems Science and Engineering. 2022, 40:191-205.

90. Yang B, Xiang L, Chen X, Jia W: An online chronic disease prediction system based on incremental deep neural network. Computers, Materials and Continua. 2021, 67:951-964.

91. Annamalai R, Nedunchelian R: Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. Computational Intelligence and Neuroscience. 2021, 2021:1-11. 10.1155/2021/5573179

92. Zhou H, Myrzashova R, Zheng R: Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking. 2020, 2020:1-13. 10.1186/s13638-020-01765-7

93. Kumar PBM, Perumal RS, Nadesh RK, Arivuselvan K: Type 2: diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering. 2020, 1:55-61. 10.1016/j.ijcce.2020.10.002

94. Ryu KS, Lee SW, Batbaatar E, Lee JW, Choi KS, Cha HS: A deep learning model for estimation of patients with undiagnosed diabetes. Applied Sciences. 2020, 10:1-12. 10.3390/app10010421

95. Patro KK, Allam JP, Sanapala U, Marpu CK, Samee NA, Alabdulhafith M, Plawiak P: An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques. BMC Bioinformatics. 2023, 24:1-21. 10.1186/s12859-023-05488-6

96. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I: Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine. 2021, 202:105968. 10.1016/j.cmpb.2021.105968

97. Aliberti A, Pupillo I, Terna S, Macii E, Di Cataldo S, Patti E, Acquaviva A: A multi-patient data-driven approach to blood glucose prediction. IEEE Access. 2019, 7:69311-69325. 10.1109/access.2019.2919184

98. Joseph LP, Joseph EA, Prasad R: Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. Computers in Biology and Medicine. 2022, 151:106178. 10.1016/j.compbiomed.2022.106178

How to cite this article:
Goyal M, Singh D (June 22, 2026) Machine Learning and Data Preprocessing Methods for Diabetes Prediction: A Systematic Review With Visual Insights. Cureus J Comput Sci 3 : es44389-026-00130-x. DOI https://doi.org/10.7759/s44389-026-00130-x