# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
1	Diabetes Prediction Using Machine Learning Techniques	Tejas N. Joshi, Prof. Pramila M. Chawan	Journal	India	2018	One of its strengths is its comparative approach, evaluating multiple machine learning models—Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Networks (ANN)—to achieve higher accuracy in diabetes prediction. The study highlights the potential of machine learning in medical diagnostics, particularly in early disease detection, which can improve patient outcomes and reduce healthcare costs. Additionally, the paper emphasizes the importance of data preprocessing and feature selection to enhance model performance.	The study does not provide detailed insights into the dataset size, distribution, or potential biases, which are crucial factors in determining the generalizability of the results. Furthermore, while the paper discusses different machine learning techniques, it lacks an in-depth performance analysis, such as precision, recall, and F1-score comparisons. The absence of a discussion on real-world deployment challenges, such as interpretability and clinical validation, also limits its practical applicability. Lastly, the study does not explore ensemble learning or hybrid approaches that could further improve predictive accuracy.	The methodology of this paper revolves around the application of supervised machine learning techniques for diabetes prediction. It begins with data collection and preprocessing, where relevant medical parameters such as glucose level, blood pressure, insulin levels, BMI, and age are used as features. Preprocessing ensures that missing values are handled and data is standardized for optimal model performance. The study then employs three supervised learning algorithms—Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Network (ANN)—to classify individuals as diabetic or non-diabetic. The dataset is split into training and testing subsets, with an emphasis on how increasing the training data can improve classification accuracy. Model performance is evaluated primarily based on classification accuracy, though the paper does not provide details on other essential metrics like precision, recall, or F1-score. Finally, the study conducts a comparative analysis to determine which model achieves the highest accuracy, concluding that SVM, Logistic Regression, and ANN are well-suited for diabetes prediction. This structured methodology underscores the potential of machine learning in medical diagnostics, particularly in enabling early detection of diabetes through predictive modeling.	—
2	Deep Belief Neural Network Model for Prediction of Diabetes Mellitus	P. Prabhu; S. Selvabharathi	Conference	India	2019	The paper demonstrates several notable advantages and merits in its approach to diabetes prediction. At its core, the proposed Deep Belief Network (DBN) model shows superior performance compared to traditional classifiers like Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, and SVM. Its comprehensive three-phase methodology, incorporating pre-processing, pre-training DBN, and fine-tuning, provides a robust framework for prediction. The implementation of PCA for feature selection enhances efficiency by reducing dimensionality. Notably, the model exhibits scalability, allowing its adaptation for other medical prediction tasks. Furthermore, the DBNN's automated learning capabilities enable it to uncover hidden patterns within complex medical datasets.	However, the study presents several limitations and challenges that warrant consideration. The reliance on a single, gender-specific dataset (Pima Indians, females only) significantly limits the model's generalizability. The relatively small sample size of 768 instances may impact the model's robustness. The paper notably lacks comparison with other deep learning architectures and doesn't address computational complexity or processing time. Additionally, the absence of cross-validation results, hyperparameter optimization discussion, and error analysis leaves gaps in understanding the model's full capabilities. The study also doesn't address the potential impact of data imbalance, with 500 negative versus 268 positive samples.	The main components of the diabetes prediction model integrate various methods across three key phases. The pre-processing phase uses min-max normalization and Principal Component Analysis (PCA) for feature selection, with correlation matrix calculations and data standardization. The dataset organization includes training, validation, and test sets preparation from the Pima Indians Diabetes Dataset with 768 samples. The pre-training phase builds the Deep Belief Network using Restricted Boltzmann Machines (RBMs) stacking. The architecture includes an input layer, three hidden layers [500 500 1000], and an output layer. It implements ReLU and sigmoid activation functions, with RBM training running for 10 epochs and Gaussian Distribution weight initialization. The model processes 8 input attributes through these layers to produce the final classification output. The fine-tuning phase implements neural network classification with a [500 500 1000 2] topology. Key methods include Stochastic Gradient Descent (SGD), ReLU and softmax activations, back-propagation, and specific parameters like 0.01 learning rate and 0.5 momentum value. The model compares its performance against Naïve Bayes, RBF-NN (Radial Basis Feed Forward Neural Network), Decision Tree, Logistic Regression, Random Forest, and Support Vector Machine (SVM).	Deep Belief Networks (DBN): Recall: 1.0 Precision: 0.6791 F1 Measure: 0.808 Naïve Bayes: Recall: 0.759 Precision: 0.763 F1 Measure: 0.760 RBF-NN: Recall: 0.761 Precision: 0.756 F1 Measure: 0.757 Decision Tree: Recall: 0.738 Precision: 0.735 F1 Measure: 0.736 Logistic Regression: Recall: 0.73 Precision: 0.73 F1 Measure: 0.73 Random Forest: Recall: 0.71 Precision: 0.72 F1 Measure: 0.72 SVM: Recall: 0.424 Precision: 0.651 F1 Measure: 0.513

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
3	Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus	Md. Faisal Faruque; Asaduzzaman; Iqbal H. Sarker	Conference	Bangladesh	2019	This paper provides an overall analysis of the various techniques that machine learning techniques use for the prediction of diabetes mellitus. Among these algorithms, there are SVM, NB, KNN, and C4.5 DT algorithms that are mentioned with respect to their effectiveness. Early detection of diabetes is underlined in this study as essential in managing the disease and avoiding complications. The research also explores the possibility of integrating machine learning methods with real-time data collection for healthcare applications to improve the timeliness and accuracy of forecasts.	The study faces challenges related to the interpretability of complex models like SVM and decision trees. These models can be difficult for healthcare professionals to understand and interpret. Additionally, the study is based on a specific dataset of 200 patients from a medical center in Chittagong, Bangladesh, which may not fully represent diverse populations with varying demographics, lifestyle factors, and healthcare access. The reliance on a single dataset may limit the generalizability of the findings to other populations.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of machine learning algorithms. The dataset used comprises 200 records with 16 attributes, including age, diet, hypertension, vision problems, and genetic factors. Data preprocessing involved converting numeric attribute values into nominal categories. The dataset was split into training and testing sets. The study implemented SVM, NB, KNN, and C4.5 decision tree algorithms. Performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score were used to evaluate the models.	Support Vector Machine (SVM) Accuracy: 70% Precision: 72% Recall (Sensitivity): 68% Specificity: 74% F1 Score: 70% Naive Bayes (NB) Accuracy: 65% Precision: 67% Recall (Sensitivity): 63% Specificity: 69% F1 Score: 65% K-Nearest Neighbor (KNN) Accuracy: 68% Precision: 70% Recall (Sensitivity): 66% Specificity: 72% F1 Score: 68% C4.5 Decision Tree (DT) Accuracy: 73.5% Precision: 72% Recall (Sensitivity): 74% Specificity: 72% F1 Score: 72%
4	Classification of Diabetes Patients Using Kernel-Based Support Vector Machines	G. A. Pethunachiyar	Conference	India	2020	This paper showcases the effectiveness of using Kernel-based Support Vector Machines (SVM) for classifying diabetes patients. The study demonstrates that SVM with different kernel functions can achieve high accuracy in predicting diabetes.The use of linear, polynomial, and radial kernels allows for flexibility in handling various data distributions. The paper provides a thorough analysis of the performance of these kernel functions, making it a valuable resource for researchers and healthcare professionals. Additionally, the study emphasizes the importance of early detection and diagnosis of diabetes, which can significantly improve patient outcomes and reduce healthcare costs.	The study faces challenges related to the interpretability of complex models like SVM with polynomial and radial kernels. These models can be difficult for healthcare professionals to understand and interpret. Additionally, the study is based on a specific dataset, which may not fully represent diverse populations with varying demographics, lifestyle factors, and healthcare access. The reliance on a single dataset may limit the generalizability of the findings to other populations.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of kernel-based SVM algorithms. The dataset used comprises 332 records with seven input variables and one output variable. Data preprocessing involved cleaning, handling missing values, and normalizing features. The dataset was split into training and testing sets, with 70% of the data used for training and 30% for testing. The study implemented SVM with linear, polynomial, and radial kernels. Performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score were used to evaluate the models.	Linear Kernel SVM Accuracy: 100% Recall (Sensitivity): 1.0 Specificity: 1.0 Polynomial Kernel SVM Accuracy: 90% Recall (Sensitivity): 1.0 Specificity: 0.87 Radial Kernel SVM Accuracy: 99% Recall (Sensitivity): 0.98 Specificity: 1.0
5	Diabetes Disease Prediction Using Machine Learning Algorithms	Arwatki Chen Lyngdoh; Nurul Amin Choudhury; Soumen Moulik	Conference	Malaysia	2020	This paper dives into the world of machine learning to predict diabetes. I t showcases the effectiveness of different algorithms,including K-Nearest Neighbor (KNN),Support Vector Machine (SVM), and Gradient Boosting.The study emphasizes the importance of early prediction of diabetes, which can significantly help in managing the disease and preventing complications. The authors also discuss the potential of integrating machine learning techniques with real-time data collection for healthcare applications,enhancing the accuracyand timeliness of predictions.The detailed comparison of the performance of these algorithms makes it a valuable resource for researchers and healthcare professionals.	The study does face some challenges, particularly with the interpretability of complex models like Gradient Boosting and SVM. These models can be a bit of a black box, making it tough for healthcare professionals to understand and interpret the results. Additionally, the study relies on a specific dataset, which may not fully capture the diversity of populations with different demographics, lifestyle factors, and healthcare access. This reliance on a single dataset might limit the generalizability of the findings to other populations. The authors also acknowledge the need for further research to validate the results on larger and more diverse datasets.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of machine learning algorithms. The dataset used comprises records with various attributes, such as glucose level, blood pressure, BMI, age, and insulin levels. Data preprocessing involved cleaning, handling missing values, and normalizing features. The dataset was split into training and testing sets. The study implemented several ML algorithms, including KNN, SVM, and Gradient Boosting. Performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score were used to evaluate the models.	K-Nearest Neighbor (KNN) with K=10 Accuracy (10-fold): 76% Precision: 0.76 Sensitivity: 0.73 F1 Score: 0.75 Naive Bayes (NB) Accuracy (10-fold): 74% Precision: 0.74 Sensitivity: 0.74 F1 Score: 0.74 Decision Tree (DT) Accuracy: 71% Precision: 0.72 Sensitivity: 0.71 F1 Score: 0.71 Random Forest (RF) Accuracy: 71% Precision: 0.70 Sensitivity: 0.71 F1 Score: 0.71 Support Vector Machine (SVM) Accuracy: 75% Precision: 0.73 Sensitivity: 0.74 F1 Score: 0.73

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
6	Deep Learning Based System Design for Diabetes Prediction	R. Bhargava and J. Dinesh	Conference	India	2021	The paper successfully applies deep learning model such as neural networks which showed high accuracy in predicting diabetes by capturing the complex relationships between various health indicators. This analysis can help in creating personalized treatment plans for patients in order to improve the results. Use of deep learning models is quite scalable and cost effective. Deep learning models can continuously learn and improve from new data, enhancing their predictive capabilitites overtime.	The study is based on the PIMA dataset which may not fully represent the diverse population. While the proposed deep learning model achieves high accuracy (98.07%), it may be overly complex, leading to potential overfitting. The use of dropout regularization helps mitigate this, but the risk remains, particularly with smaller datasets. The paper emphasizes accuracy as a primary metric for model performance. However, it does not sufficiently address other critical metrics such as precision, recall, and F1 score, which are essential for understanding the model's effectiveness in real-world scenarios where false positives and negatives can have significant consequences.	The paper presents a comprehensive approach to diabetes prediction using deep learning techniques, particularly focusing on a Convolutional Neural Network (CNN). They utilize the Pima Indians Diabetes dataset, which is widely recognized for its relevance in diabetes research. The methodology involves designing a deep neural network with multiple hidden layers, incorporating dropout regularization to mitigate overfitting. The model employs a binary cross-entropy loss function to optimize performance, achieving an impressive training accuracy of 98.07%. The CNN architecture consists of convolutional layers that extract features from the input data, followed by pooling layers that reduce dimensionality while retaining essential information. The activation function used is the Rectified Linear Unit (ReLU), which introduces non-linearity into the model. The final predictions are made through fully connected layers and a Softmax function that classifies patients as diabetic or non-diabetic based on the learned features. The results indicate that the proposed deep learning model outperforms traditional machine learning methods, showcasing its potential for effective diabetes prediction and classification. This research highlights the importance of leveraging advanced computational techniques to enhance healthcare outcomes, particularly in chronic disease management.	Deep Learning (DL): Accuracy: 98.07% Precision: 95.22% Recall: 98.46% F-Measure: 96.81% Specificity: 99.29% Sensitivity: 95.52% Decision Tree (DT): Accuracy: 96.62% Precision: 94.02% Recall: 95.45% F-Measure: 94.72% Specificity: 97.86% Sensitivity: 94.03% Artificial Neural Network (ANN): Accuracy: 90.34% Precision: 88.05% Recall: 83.09% F-Measure: 85.98% Specificity: 91.43% Sensitivity: 88.06% Naive Bayes (NB): Accuracy: 76.33% Precision: 59.07% Recall: 64.51% F-Measure: 61.67% Specificity: 84.29% Sensitivity: 59.70%
7	Detecting Diabetes in Human Body using Different Machine Learning Techniques	Subrata Kumer Paul, Md. Ekramul Hamid, Rakhi Rani Paul	Conference	India	2021	The paper demonstrates diabetes detection through the application of various machine learning algorithms such as Decision Trees, K-Nearest Neighbors, Naive Bayes, and Random Forest. The authors effectively demonstrate the efficacy of these models, with Random Forest achieving the highest accuracy of 86%, thereby providing a robust solution for early diabetes detection. The methodology outlined is detailed, including data preprocessing steps and performance metrics like sensitivity, precision, and F1-score, which are essential for evaluating classifier performance.	One significant concern is the reliance on a single dataset (Pima Indians Diabetes Dataset), which may limit the generalizability of the results. The dataset comprises only female patients and may not adequately represent the broader population, potentially skewing the applicability of the findings. Furthermore, while the study compares several algorithms, it does not delve deeply into hyperparameter tuning or model optimization techniques that could further enhance performance. The lack of discussion regarding potential biases in data collection or model training also presents a limitation, as these factors can significantly impact the reliability of machine learning outcomes in clinical settings.	The core of the methodology focuses on training four different machine learning algorithms: Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest. Each model is trained using the training dataset, and 10-fold cross-validation is employed to validate the overall accuracy of these models. The performance of each classifier is subsequently assessed using metrics such as sensitivity, precision, F1-score, and overall accuracy. The results indicate that the Random Forest model outperforms others with an accuracy of 86%, showcasing its effectiveness in diabetes classification. This comprehensive methodology not only highlights the importance of data preparation but also emphasizes the comparative analysis of multiple machine learning techniques in addressing a significant health concern.	Random Forest (RF) Accuracy: 86% F1 Score: 0.91 Sensitivity: 0.90 Precision: 0.92 Specificity: 0.64 K-Nearest Neighbors (KNN) Accuracy: 76% F1 Score: 0.83 Sensitivity: 0.79 Precision: 0.88 Specificity: 0.66 Decision Tree (DT) Accuracy: 73% F1 Score: 0.81 Sensitivity: 0.75 Precision: 0.89 Specificity: 0.63 Naive Bayes (NB) Accuracy: 72% F1 Score: 0.80 Sensitivity: 0.77 Precision: 0.84 Specificity: 0.58

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
8	Analysis of Machine Learning Algorithms in Diabetes Mellitus Prediction.	Roshi Saxena, Sanjay Sharma, Manali Gupta	Journal	India	2021	The study evaluates multiple machine learning algorithms, including logistic regression, Naive Bayes, support vector machines, random forest, and neural networks, providing a broad perspective on their effectiveness in predicting diabetes. The authors use various performance metrics such as precision, recall, accuracy, F-measure, and receiver operating characteristics to evaluate the models, offering a detailed analysis of each algorithm's performance. The use of the Weka tool for implementing the algorithms makes the study practical and accessible for other researchers and practitioners.	The study relies solely on the PIMA Indians diabetes dataset, which may not be representative of the broader population. The study is conducted in a controlled environment using a specific dataset. Real-world validation with diverse datasets and clinical settings is necessary to confirm the findings. While the study evaluates several algorithms, it does not explore newer or more advanced machine learning techniques that could potentially offer better performance	The methodology used in the paper "Analysis of Machine Learning Algorithms in Diabetes Mellitus Prediction" involves several key steps. The authors utilized the PIMA Indians diabetes dataset, which includes data from 768 female Pima Indians aged 21 and older. The dataset comprises eight variables: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. The paper demonstartes five machine learning algorithms: logistic regression (LR), Naive Bayes (NB), support vector machines (SVM), random forest (RF), and neural networks (NN) using the Weka tool. They employed 10-fold cross-validation to evaluate the performance of these models. The performance metrics used for evaluation included precision, recall, accuracy, F-measure, and receiver operating characteristics (ROC). The results indicated that logistic regression outperformed the other algorithms in terms of accuracy, making it the most effective model for predicting diabetes in this study	Logistic Regression Accuracy: 77.2% Precision: 0.767 Recall: 0.772 F1 Measure: 0.765 ROC: 0.832 Support Vector Machine Accuracy: 77.08% Precision: 0.766 Recall: 0.771 F1 Measure: 0.761 ROC: 0.717 Naïve Bayes Accuracy: 76.30% Precision: 0.759 Recall: 0.763 F1 Measure: 0.760 ROC: 0.819 Random Forest Accuracy: 75.5% Precision: 0.751 Recall: 0.755 F1 Measure: 0.752 ROC: 0.820 Neural Networks Accuracy: 75.1% Precision: 0.748 Recall: 0.751 F1 Measure: 0.749 ROC: 0.791
9	Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach	Madhumita Pal; Smita Parija; Ganapati Panda	Conference	India	2021	This paper presents a comprehensive comparison of three machine learning algorithms (K-NN, Linear SVM, and Random Forest) for diabetes prediction, achieving a high accuracy of 78.57% and AUC of 95.08% with Random Forest, surpassing several previous studies using the PIMA Indian Diabetes dataset. The research incorporates essential data preprocessing steps such as null value checking, data cleaning, and outlier removal. It provides clear visualizations of data correlations and glucose level relationships. The developed model has the potential to aid healthcare professionals in early diabetes diagnosis and appropriate treatment planning. 1 Furthermore, the approach exhibits generalizability and can potentially be applied to the detection of other diseases. The inclusion of both accuracy and AUC metrics ensures a more robust evaluation of the model's performance.	This research faces several limitations. The dataset size is limited (769 samples), which may hinder the generalizability of the findings. The study utilizes only 8 features for prediction, potentially overlooking other crucial indicators of diabetes. It does not address the issue of class imbalance within the dataset, which could skew model performance. Furthermore, the research lacks a discussion of feature importance or selection methods, potentially leading to suboptimal model performance. Cross-validation results are absent, limiting the robustness of the performance evaluation. The paper does not delve into hyperparameter tuning for the employed algorithms, potentially impacting their optimal performance. Additionally, interpretability aspects of the models are not discussed, hindering the understanding of their decision-making process. The research fails to address the handling of missing values in real-world scenarios, which is crucial for practical implementation. Finally, the absence of an external validation dataset prevents a thorough assessment of the model's generalizability.	This study investigates the performance of three machine learning algorithms (K-Nearest Neighbors, Linear Support Vector Machine, and Random Forest) for diabetes prediction using the PIMA Indian Diabetes dataset comprising 769 samples and 8 features. Data preprocessing steps include null value checking, data cleaning, and outlier removal. The dataset is then split into training and testing sets with an 80-20 ratio to evaluate the performance of the implemented models.	Random Forest: Accuracy: 78.57% AUC: 95.08% True Positives: 22 True Negatives: 94 False Positives: 6 False Negatives: 25 Linear SVM: Accuracy: 77.92% AUC: 73.13% True Positives: 31 True Negatives: 89 False Positives: 10 False Negatives: 24 K-NN: Accuracy: 77.27% AUC: 68.19% True Positives: 21 True Negatives: 98 False Positives: 10 False Negatives: 25
10	Prediction of Diabetes with its Symptoms Based on Machine Learning	Xingchen Xu; Xiao Huang; Jinhui Ma; Xuejianwei Luo	Conference	USA	2021	This paper employs a multi-method approach by combining literature analysis, data analysis, and machine learning, which provides robust validation through triangulation. The researchers thoroughly examined the relationships between different symptoms and diabetes, using various visualization techniques like radar maps, heat maps, and histograms to represent the correlations clearly. The paper excels in its systematic evaluation of multiple machine learning models, starting from simple approaches like SGD and KNN, then progressing to more complex models like Random Forest and Neural Networks. This comparative approach helps establish which models perform best for diabetes prediction. Another significant merit is the careful consideration of feature importance. The researchers analyzed how different symptoms correlate with diabetes and with each other, providing valuable insights for medical professionals about which symptoms might be most indicative of diabetes.	The dataset used is relatively small with only 520 samples, which might not be representative of the global diabetic population. Additionally, all data was collected from a single hospital in Bangladesh, limiting the geographical and demographic diversity of the sample. The paper also shows some gaps in addressing confounding variables. While it considers age and gender, other important factors like family history, ethnicity, and lifestyle factors aren't included in the analysis. The binary nature of most variables (yes/no responses) might oversimplify complex medical symptoms that could exist on a spectrum. A significant challenge appears in reconciling conflicting findings between medical literature and the data analysis, particularly regarding obesity's relationship with diabetes. The paper acknowledges this limitation but doesn't fully resolve the discrepancy.	This research utilizes a dataset with 16 attributes and 1 target class, employing binary encoding for attributes and normalizing age values. The methodology encompasses Exploratory Data Analysis (EDA) alongside the evaluation of both simple (SGD, KNN, Decision Tree) and complex machine learning models (Random Forest, Neural Networks, AdaBoost). Statistical analysis, including Crammer's V coefficient and heat maps, is employed to identify key features, focusing on demographic factors (age, gender), physical symptoms (Polyuria, Polydipsia, etc.), and associated conditions (Obesity, Visual Blurring, etc.).	Simple Models: SGD: 87.5% (without normalization), 85.7% (with normalization) KNN: 88.2% (without normalization), 92.3% (with normalization) Decision Tree: 96.9% (training), 97% (testing) Complex Models: Random Forest: 98.1% (training with 1.8% standard deviation), 98% (testing) Neural Networks: 96% (both training and testing) AdaBoost with Decision Tree: 97.8% (training), 98% (testing) AdaBoost with MLP: 97.3% (training)

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
11	Diabetes Mellitus Early Stage RIsk Prediction Using ML Algorithms	Sarra Samet; Mohamed Ridda Laouar; Issam Bendib	Conference	Algeria	2021	The research presents a well-structured hybrid model approach that achieves superior accuracy (90.62%) compared to individual machine learning algorithms and previous studies. This represents a significant improvement over existing methods. The authors employ comprehensive data preprocessing techniques, particularly in handling missing values and zero entries in the dataset. Rather than simply removing incomplete records, they use mean/median imputation to preserve data integrity. The study provides extensive data visualization and analysis, including correlation heatmaps, density distributions, and detailed statistical descriptions of the dataset features. This helps readers understand the underlying data characteristics and relationships between variables. The research compares multiple machine learning algorithms (NB, RF, LR, KNN, SVM, DT) systematically, demonstrating the effectiveness of each approach before developing the hybrid model. This thorough comparison helps justify their methodological choices.	The study relies solely on the Pima Indians Diabetes Database, which is limited to female patients, reducing the model's generalizability to the broader population. The relatively small dataset (768 instances) contains a high number of zero values in critical parameters like Insulin (366 records) and Skin Thickness (220 records), potentially affecting model accuracy. Additionally, the research lacks external validation on different populations or datasets, making real-world applicability uncertain. There is no discussion on overfitting mitigation or cross-validation techniques, which are crucial for model reliability. The study also does not explore feature importance or selection methods, limiting potential performance improvements and insights for medical practitioners. Furthermore, it lacks comparisons with deep learning approaches, considerations of computational complexity, and discussions on model interpretability	The study involves a detailed examination of the PIMA dataset (768 instances with 8 attributes), including handling missing values and zero entries through comprehensive data preprocessing, cleaning, and statistical analysis. Various data visualization techniques are applied to understand feature distributions. Six supervised learning algorithms—Naive Bayes, Random Forest, Logistic Regression, KNN, SVM, and Decision Tree—are implemented, along with a hybrid stacking model. The hybrid approach uses KNN, SVM, and Decision Tree in the base layer, with Logistic Regression as the meta-layer, ensuring a systematic comparison of different approaches. The methodology follows a clear step-by-step framework, covering data preprocessing, model development, evaluation, and performance assessment.	Individual Algorithm Performance: Naive Bayes: 77.27% accuracy Random Forest: 83.76% accuracy Logistic Regression: 78.57% accuracy KNN (n=5): 88.31% accuracy SVM: 87.01% accuracy Decision Tree: 85.71% accuracy Hybrid Model Performance: Accuracy: 90.62% Precision: 0.91 Recall: 0.91 F1-score: 0.90
12	On The Analysis of Some Machine Learning Algorithms for the Prediction of Diabetes	Bello A. Bodinga, Mukhtar A. Abdulsalam, Bello A. Buhari, Muzzammil Mansur	Journal	Nigeria	2022	The research effectively employs multiple machine learning algorithms, specifically Logistic Regression, Decision Tree, and Random Forest, to analyze their performance on the PIMA Indian Diabetes Dataset. The paper utilizes well-recognized performance measures such as Accuracy, F-measure, Recall, and Precision to assess the algorithms' effectiveness.	The reliance on a single dataset (PIMA Indian Diabetes Dataset) may limit the generalizability of the results. The paper acknowledges the necessity of data preprocessing to handle missing values and inconsistencies. However, it does not delve deeply into how these preprocessing steps might impact the overall model performance or how different techniques could be employed to address these issues more effectively	The methodology outlined in the paper focuses on leveraging Convolutional Neural Networks (CNNs) for effective diabetes prediction. The research utilizes the PIMA Indian Diabetes dataset, which contains various health metrics relevant to diabetes. The data undergoes preprocessing to handle missing values and normalize inputs, ensuring a clean and structured dataset for model training. At the core of the methodology is a CNN architecture comprising multiple layers, including convolutional layers, pooling layers, and fully connected layers. Each neuron processes input data through operations such as dot products, weighted sums, and activation functions, specifically ReLU, which introduces non-linearity into the model. Convolutional layers extract key features, while pooling layers reduce dimensionality by down-sampling feature maps. The model is trained using a binary cross-entropy loss function, suitable for classification tasks involving two classes—diabetic and non-diabetic. Throughout training, parameters such as weights and biases are optimized to minimize the loss function and enhance predictive accuracy. To prevent overfitting, dropout regularization is applied by randomly deactivating certain neurons during training. The performance of the CNN model is evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, providing a comprehensive assessment of its predictive effectiveness. The methodology results in an impressive accuracy of 98.07% on the PIMA dataset, showcasing the efficiency of deep learning techniques in early diabetes prediction. Overall, the approach highlights a systematic process incorporating data preprocessing, advanced neural network architecture, rigorous training, and thorough evaluation, demonstrating the potential of CNNs in improving diabetes prediction compared to traditional machine learning methods.	Logistic Regression: Accuracy: 76% Precision: 0.77 Recall: 0.55 F-measure: 0.64 Random Forest: Accuracy: 75% Precision: 0.70 Recall: 0.51 F-measure: 0.59 Decision Tree: Accuracy: 73% Precision: 0.60 Recall: 0.58 F-measure: 0.62

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
13	Performance Analysis of Machine Learning Models for Prediction of Diabetes.	Anuj Mangal, Vinod Jain	Conference	India	2022	The paper employs K-Fold validation, which enhances the reliability of the model by reducing overfitting and providing a more robust evaluation. The research demonstrates the practical application of machine learning in healthcare, specifically for early diabetes prediction, which can potentially reduce the burden on healthcare systems	The dataset used in the study is relatively small, with only 520 instances. This may limit the generalizability of the findings to larger, more diverse populations. The data is collected from a single hospital in Sylhet, Bangladesh, which may not represent the global population. The study only evaluates two machine learning models. Including more models could provide a broader perspective on the best approaches for diabetes prediction.	The methodology of the paper is centered around the application of two machine learning models: Logistic Regression (LR) and Random Forest (RF) classifiers. The authors utilized a dataset from Kaggle, which includes data on 520 individuals aged 20 to 65 years, collected from Sylhet Diabetes Hospital in Bangladesh. The dataset was divided into training and testing sets following the 80-20 rule, where 80% of the data was used for training the models and 20% for testing their prediction accuracy. The study employed K-Fold validation with K=10 to enhance the reliability of the models by reducing overfitting. The performance of the models was evaluated based on prediction accuracy, with the Random Forest model achieving an impressive 99% accuracy, significantly outperforming the Logistic Regression model. The methodology highlights the practical application of machine learning in healthcare, demonstrating the potential of these models for early diabetes prediction without the need for invasive medical tests	Model-Accuracy Logistic Regression-94.23% Random Forest- 99.03%
14	Detecting Diabetes Using Machine Learning Algorithms	Qabeela Q. Thabit; Taqwa O. Fahad; Alyaa I. Dawood	Conference	Iraq	2022	The paper provides a comprehensive evaluation of seven different machine learning algorithms (Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbor, and Naive Bayes) for diabetes prediction. This comparative approach helps establish which algorithms perform best for this specific medical diagnosis task. The paper achieves remarkably high accuracy rates, with some algorithms reaching nearly 100% accuracy, particularly the SVM and Decision Tree classifiers. This represents a significant improvement over traditional diagnostic methods. The research also employs a systematic methodology for data preprocessing, including cleaning, feature selection, scaling, and splitting - all crucial steps for ensuring reliable machine learning results. Furthermore, the paper provides extensive performance comparisons with previous research, giving readers a clear understanding of how their results stack up against existing work in the field.	The dataset used (Pima Indian Dataset) is relatively small with only 768 records and is limited to female patients above 21 years of age from a specific ethnic group. This narrow demographic focus raises questions about the model's generalizability to other populations. The paper doesn't address potential biases in the dataset or discuss how the models might perform with more diverse patient data. Another limitation is the lack of external validation using different datasets. While the internal validation metrics are impressive, the paper doesn't test the models on completely independent data from different sources or populations.	The paper's framework includes data preprocessing, which involves data cleaning, feature selection, scaling, and splitting to ensure high-quality input for model training. It implements seven different machine learning algorithms and evaluates their performance using a confusion matrix while also conducting a comparative analysis with previous research to highlight improvements. The performance metrics used for evaluation include accuracy, sensitivity, specificity, F1-score, and ROCAUC score, providing a comprehensive assessment of each model's effectiveness in diabetes prediction	Performance metrics for different classifiers: Support Vector Machine (SVM): Accuracy: 1.0000 Sensitivity: 1.0000 Specificity: 1.0000 F1-Score: 1.0000 ROCAUC: 1.0000 Decision Tree: Accuracy: 1.0000 Sensitivity: 1.0000 Specificity: 1.0000 F1-Score: 1.0000 ROCAUC: 1.0000 Gradient Boosting: Accuracy: 0.941 Sensitivity: 0.941 Specificity: 0.941 F1-Score: 0.9411 ROCAUC: 0.954 Random Forest: Accuracy: 0.9411 Sensitivity: 0.9411 Specificity: 0.9411 F1-Score: 0.9411 ROCAUC: 0.9411 Logistic Regression: Accuracy: 0.9816 Sensitivity: 0.9816 Specificity: 0.9816 F1-Score: 0.9816 ROCAUC: 0.9714 Naive Bayes: Accuracy: 0.8974 Sensitivity: 0.8974 Specificity: 0.8974 F1-Score: 0.8974 ROCAUC: 0.8845 K-Nearest Neighbor: Accuracy: 0.9 Sensitivity: 0.9 Specificity: 0.9 F1-Score: 0.9 ROCAUC: 0.9 The paper shows that SVM and DT achieved the highest accuracy (100%), though this unusually perfect performance might warrant further investigation for potential overfitting.

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
15	A Machine Learning Technique for Detection of Diabetes Mellitus	O.M. Awoniran; M.O. Oyelami; R.N. Ikono; R. F. Famutimi; T.I. Famutimi	Conference	Nigeria	2022	The research enhances diabetes prediction accuracy by implementing advanced data preprocessing techniques, including dummy categorization and PCA for dimensionality reduction, and structuring data into age groupings (under 40, 41-59, over 60) for better analysis. It employs multiple machine learning algorithms, including SVM, Random Forest Classifier (RFC), and Deep Neural Networks (DNN), providing a comprehensive evaluation. Using a dataset of 768 instances with 8 independent variables, the study achieves a high accuracy rate of 89% with DNN. The methodology is well-structured and replicable, incorporating detailed exploratory data analysis with visualizations while considering multiple diabetes types (Type 1, Type 2, and gestational) to improve prediction reliability.	The study faces several limitations, including a small dataset size (768 instances), which is relatively limited compared to modern big data standards. It does not explore advanced ML algorithms or ensemble methods, and there is no discussion on feature importance or variable significance. The evaluation is primarily focused on accuracy, lacking comprehensive validation metrics such as precision, recall, and F1-score, and no cross-validation results are presented. Additionally, the study does not address model interpretability, lacks an external validation dataset, and does not account for class imbalance (500 negative vs. 268 positive cases). Furthermore, it provides limited discussion on hyperparameter tuning, which could impact the model's overall performance and generalizability.	The study's Data Processing and Analysis involves the collection of a diabetes dataset with 8 independent variables and 1 dependent variable, the creation of age groups and glucose level categories, and the application of dummy variables for categorical encoding. Principal Component Analysis (PCA) is implemented for dimensionality reduction to enhance model efficiency. The Machine Learning Implementation evaluates three models: Support Vector Machine (SVM), Random Forest Classifier (RFC), and Deep Neural Network (DNN) to predict diabetes effectively. For Visualization and Analysis, the study employs count plots for outcome distribution, joint plots for variable relationships, and kernel density estimation plots to explore data patterns and improve interpretability. Model performance is assessed using accuracy, negative/positive predictive values, true negative/positive rates, and F1-scores, ensuring a comprehensive evaluation of predictive reliability.	Support Vector Machine (SVM): Accuracy: 76% NPV: 0.76 PPV: 0.77 TNR: 0.93 TPR: 0.46 F1-score (negative): 0.83 F1-score (positive): 0.57 Random Forest Classifier (RFC): Accuracy: 77% NPV: 0.81 PPV: 0.68 TNR: 0.84 TPR: 0.64 F1-score (negative): 0.83 F1-score (positive): 0.66 Deep Neural Network (DNN): Accuracy: 89% NPV: 0.98 PPV: 0.87 TNR: 0.92 TPR: 0.79 F1-score (negative): 0.95 F1-score (positive): 0.83
16	Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record	Sony M Kuriakose; Peeta Basa Pati; Tripty Singh	Conference	India	2022	This paper dives deep into the world of machine learning to predict diabetes using a massive dataset of 70,000 clinical records. It showcases the effectiveness of various algorithms, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). The study underscores the importance of early diabetes prediction, which can significantly help in managing the disease and preventing complications. The paper also explores the potential of integrating machine learning techniques with real-time data collection for healthcare applications, enhancing the accuracy and timeliness of predictions.	The study does face some challenges, particularly with the interpretability of complex models like Random Forest and SVM. These models can be a bit of a black box, making it tough for healthcare professionals to understand and interpret the results. Additionally, the study relies on specific datasets, including the Diabetes 130-US hospitals dataset and the Pima Indian Diabetes Dataset, which may not fully capture the diversity of populations with different demographics, lifestyle factors, and healthcare access. This reliance on specific datasets might limit the generalizability of the findings to other populations.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of machine learning algorithms. The dataset used comprises 70,000 clinical records with various attributes, such as glucose level, blood pressure, BMI, age, and insulin levels. Data preprocessing involved cleaning, handling missing values, and normalizing features. The dataset was split into training and testing sets. The study implemented several ML algorithms, including Logistic Regression, Random Forest, SVM, and KNN. Performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score were used to evaluate the models.	Accuracy result of ML methods when applied to Health Facts database with 70000 patients. Logistic Regression Accuracy: 76% Random Forest Accuracy: 79% Support Vector Machine (SVM) Accuracy: 77% K-Nearest Neighbor (KNN) Accuracy: 69% Accuracy results of ML methods when applied to Pima Data Indian Diabetes Database with 768 patients. Logistic Regression Accuracy: 73% Random Forest Accuracy: 80% Support Vector Machine (SVM) Accuracy: 77% K-Nearest Neighbor (KNN) Accuracy: 73%
17	Diabetes Prediction based on Supervised and Unsupervised Learning Techniques - A Review	M Sivaraman; J Sumitha	Conference	India	2022	The authors of this paper identify and evaluate the performance of different supervised and unsupervised learning technologies which are used for both type 1 and type 2 diabetes. They give insights about the type of learnings and have compared different ML algorithms such as Decision Trees, Support Vector Machines (SVM), Naive Bayes, and K-means clustering. Also, they clearly explain the KDD (Knowledge Discovery in Databases) process and its steps for extraction of useful knowledge from datasets.	The study relies on databases like PIMA Indian Diabetes dataset which means that the results may not generalize well to other populations. The study does not provide in depth comparison of why certain algorithms perform better than others when performed on different dataset. Also, it does not provide a detailed cost-benefit analysis of the proposed methods. Lastly, the paper highlights gap between the research findings and the real world applications.	The authors compare various Machine Learning algorithms and provide insights into their performane. Decision Trees, Support Vector Machines (SVM), Naive Bayes, and K-means clustering are some of the algorithms which have been considered in the study. They also emphasized that when a model is used in combination with an unsupervised technique like K-Mean and PCA, the accuracy and precision of the prediction increases. They also highlight that the algorithms like Decision-Tree, Logistic-Regression, SVM fall under the supervised category and algorithms like Hierarchical clusters, K-means clusters, Apriori algorithm fall under unsupervised learning.	The paper conducts a comparitive study between various methods used by researchers and academicians and tabulates those results (both for supervised and unsupervised learning methods) AdaBoost: 98.8% Random Forest: 94.10% XGBoost: 88.1% K-means: 78% Artificial Neural Networks: 75.7% SVM and K-means combined: 99.64%

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
18	Diabetes Mellitus Prediction Based on Machine Learning Techniques	Eliana S. Omoora, Hajer A. Altaweil, Kenz A. Bozed, Tarek Nagem	Conference	Tunisia	2023	This paper highlights the effectiveness of the XGBoost algorithm by predicting diabetes at an early stage, which can help in preventing severe complications associated with the disease. The study also utilizes other machine learning algorithms like Support Vector Machines, Naïve Bayes, Decision Tree and Random Forest alongwith comparing their performance, providing a thorough analysis of their effectiveness. The paper proposes a future development plan to transform the prediction system into a cognitive artificial intelligence system, which is believed to enhance its capabilities tremendeously.	Complex models like XGBoost may lead to overfitting, particularly is hyperparameter tuning is not done adequately. This results in the model performing well on the training data but might provide poor results for unseen data. Models like Random forest and XGBoost can be difficult to interpret. This lack of transparency may hinder healthcare professionals who need to understand the reasoning behind the predictions. The study is based on two datasets which may not fully represent the diverse population varying in demographics, lifestyle factors and healthcare access.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of machine learning algorithms. Two datasets were utilized. The first dataset was split into two subsets to train Model A (Diabetes Detection) and Model C (Prediabetes Prediction), while the second dataset was used for Model B (Diabetes Type Classification - Type 1 vs. Type 2). Data Preprocessing involved checking for missing values and handling them. Followed by converting categorical data like gender, polyphagia, vision blurring etc. into numerical representations, and applying feature scaling to normalize them and ensure consistency. The dataset was then split into 80:20 for training and testing purposes. Five supervised machine learning models were used whose perfomance was evaluated and compared. The performance of these models was evaluated using key metrics such as accuracy, sensitivity (recall), specificity, precision, and confusion matrices. The evaluation results demonstrated that XGBoost consistently outperformed the other algorithms across all three models, achieving the highest accuracy and classification effectiveness.	Model A - Diabetes Prediction Algorithm: XGBoost Accuracy: 92.5% Precision: 91.0% Recall: 90.0% F1 Score: 90.5% Model B - Type-1 vs Type-2 Diabetes Classification Algorithm: Random Forest Accuracy: 89.2% Precision: 88.0% Recall: 87.5% F1 Score: 87.8% Model C - Prediabetes Prediction Algorithm: Support Vector Machines (SVM) Accuracy: 85.3% Precision: 84.5% Recall: 83.0% F1 Score: 83.7%
19	Machine Learning-based Diabetes Prediction: A Cross-Country Perspective	Sadia Afrin Shampa; Md. Saiful Islam; Ayatun Nesa	Conference	Bangladesh	2023	The paper demonstrates several notable strengths in its approach and findings. First, it presents a unique cross-country perspective by analyzing diabetes data from Bangladesh, India, and Germany, providing valuable insights into how prediction models perform across different populations. The research is particularly significant for Bangladesh, where approximately 13.13 million people suffer from diabetes. The study's use of multiple machine learning algorithms, especially boosting algorithms like AdaBoost, CatBoost, Gradient Boost, and XGBoost, shows sophisticated methodology. These algorithms proved particularly effective with the Bangladesh dataset, achieving nearly perfect accuracy scores of 0.999-1.000 in both training and testing phases. Another significant merit is the paper's thorough approach to data preprocessing, including handling class imbalance through the ADASYN oversampling method.	The most significant challenge faced was data availability and quality. The researchers note that diabetes prediction becomes challenging due to the high-dimensional nature of the data and its confidential nature, limiting access to comprehensive datasets. The study also reveals a substantial class imbalance in the original dataset, with 10,978 diabetic cases compared to only 1,552 normal cases and 1,871 pre-diabetic cases in the Bangladesh dataset. While the researchers attempted to address this through oversampling, this artificial balancing might not perfectly represent real-world conditions. The paper lacks detailed discussion of feature selection results and the relative importance of different features in prediction. Additionally, while the study includes data from three countries, the datasets vary significantly in size (14,401 from Bangladesh, 768 from India, and 2,000 from Germany), making direct comparisons potentially problematic.	The paper consists of several key components, beginning with Data Collection and Preprocessing, which involves 15,000 patient records from Bangladesh, additional data from the PIMA Indian dataset (768 patients), and the German dataset from Frankfurt hospital (2,000 patients). To enhance data quality, the study applies ADASYN oversampling and utilizes feature selection techniques such as PCA, ICA, and correlation-based methods. In the Machine Learning Implementation, nine different algorithms are evaluated, including Decision Trees, Naïve Bayes, SVM, Random Forest, ANN, CatBoost, AdaBoost, Gradient Boosting, and XGBoost. The models undergo cross-validation and performance evaluation, with results compared across different datasets to assess their effectiveness in diabetes prediction.	For Bangladesh Dataset: Boosting Algorithms (AdaBoost, CatBoost, XGBoost, Gradient Boost): Accuracy: 99.9-100% Precision: 99.7-100% Recall: 100% F1 Score: 99.8-100% For PIMA Indian Dataset: Best performing model (CatBoost): Accuracy: 83.1% Precision: 81% Recall: 81.6% F1 Score: 81.3% For German Dataset: Best performing models (AdaBoost and CatBoost): Accuracy: 99% Precision: 98.4-99.3% Recall: 98.4-99.3% F1 Score: 98.8%
20	Machine Learning based Early Predication and Detection of Diabetes Mellitus	Prosanjeet Sarkar Santosh Pawar	Conference	India	2023	This research paper implements multiple machine learning algorithms (8 different classifiers) for comprehensive comparison. It uses the established Pima Indians Diabetes Dataset (PIDD) which allows for benchmarking. The paper includes detailed data preprocessing steps including handling missing values and data imbalance. It achieves high accuracy with XGBoost (89.07%) which outperforms traditional methods. It provides a complete pipeline from data preprocessing to model evaluation. It addresses class imbalance using oversampling techniques. It implements normalization using z-score technique for better model performance. It includes visual representations of data distribution and feature analysis. It has potential for practical application for early diabetes detection.	This research paper is limited to one dataset (PIDD) which may affect generalizability. It uses a gender-specific dataset (only female patients). No cross-validation results are presented. It lacks a detailed discussion of hyperparameter optimization. It does not compare with deep learning methods. The dataset size is relatively small (768 samples). There is no discussion of computational complexity or processing time. The exploration of feature importance is limited. It lacks external validation on different populations. There is no discussion of model interpretability.	This research paper investigates the performance of various machine learning algorithms for diabetes prediction using the Pima Indians Diabetes Dataset (PIDD). The data preprocessing phase includes handling missing values through mean imputation and addressing class imbalance through oversampling techniques. Data normalization is performed using the z-score technique. The study evaluates eight different classification algorithms: Naïve Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB), and LightGBM (LGBM). The performance of these models is assessed using a comprehensive set of metrics, including Accuracy, Precision, Recall, Sensitivity, Specificity, F1-Score, and Error Rate.	Accuracy Comparison (Table 3 of the paper): XGBoost: 89.07% LightGBM: 88.28% Random Forest: 88.15% SVM: 85.39% Logistic Regression: 84.86% KNN: 84.07% Naïve Bayes: 82.13% Decision Tree: 80.12% Precision, Recall, and F1-Score: These metrics are not explicitly broken down per model in the paper, but the confusion matrix suggests they align closely with the accuracy for each model. Observations: XGBoost outperforms due to its ability to handle imbalanced data and overfitting through regularization.

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
21	Diabetes Prediction in Teenagers using Machine Learning Algorithms	Jayavrinda Vrindavanam; Raye Haarika; Sindhu MG; Kilari Sumanth Kumar	Conference	India	2023	: This paper explores the use of various machine learning algorithms to predict diabetes in teenagers. It highlights the effectiveness of different algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost. The study emphasizes the importance of early prediction of diabetes in teenagers, which can help in managing the disease and preventing complications. The paper also discusses the potential of integrating machine learning techniques with real-time data collection for healthcare applications, enhancing the accuracy and timeliness of predictions.	The study faces challenges related to the interpretability of complex models like Random Forest and XGBoost. These models can be difficult for healthcare professionals to understand and interpret. Additionally, the study is based on a specific dataset collected from 150 students at Dayananda Sagar University, which may not fully represent diverse populations with varying demographics, lifestyle factors, and healthcare access. The reliance on this dataset might limit the generalizability of the findings to other populations.	The paper follows a structured workflow that includes data collection, preprocessing, model selection, training, evaluation, and comparison of machine learning algorithms. The dataset used comprises records from 150 students with various attributes, such as age, gender, BMI, diet type, blood pressure, exercise routine, parental history, and smoking or drinking habits. Data preprocessing involved cleaning, handling missing values, and normalizing features. The dataset was split into training and testing sets. The study implemented several ML algorithms, including Logistic Regression, KNN, SVM, Random Forest, and XGBoost. Performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score were used to evaluate the models.	Logistic Regression Accuracy: 79% Sensitivity: 0.84 Specificity: 0.75 F1 Score: 0.78 AUC: 0.86 K-Nearest Neighbors (KNN) Accuracy: 58% Sensitivity: 0.69 Specificity: 0.50 F1 Score: 0.60 AUC: 0.59 Support Vector Machine (SVM) Accuracy: 82% Sensitivity: 0.84 Specificity: 0.81 F1 Score: 0.81 AUC: 0.82 Random Forest
22	Predictions of Diabetic Mellitus using ML Techniques: A Systematic Overview	T Krishna Manaswini; Padmalaya Nayak; Vanam Sri Harshitha; Shreya Barlapudi	Conference	India	2023	In order to anticipate diabetes mellitus, this paper delves into the realm of machine learning. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting are among the methods whose efficacy is demonstrated. The study highlights the value of early diabetes detection, which can greatly aid in the management of the condition and the avoidance of complications. In order to improve prediction accuracy and timeliness, the authors also go over the possibilities of combining machine learning methods with real-time data collection for healthcare applications. It is a useful tool for researchers and medical experts because it compares the performance of these algorithms in detail.	There are certain difficulties with the study, especially when it comes to the interpretability of intricate models like SVM and Gradient Boosting. It might be challenging for medical practitioners to comprehend and interpret the outcomes of these models because they can be somewhat of a mystery. The study also depends on a particular dataset, which might not adequately represent the variety of communities with varying demographics, lifestyles, and access to healthcare. This dependence on a single dataset may restrict how broadly the results may be applied to other populations. In order to confirm the findings on bigger and more varied datasets, the authors nevertheless admit that more investigation is required.	Data collection, preparation, model selection, training, assessment, and comparison of machine learning methods are all part of the paper's systematic approach. Age, blood pressure, BMI, insulin levels, glucose levels, and other variables are all included in the dataset. Cleaning, addressing missing values, and normalizing characteristics were all part of the preprocessing of the data. The dataset was divided into two parts: testing and training. A number of machine learning techniques were used in the study, including Gradient Boosting, SVM, KNN, Random Forest, Decision Trees, and Logistic Regression. The models' performance was assessed using metrics like F1-score, sensitivity, specificity, accuracy, and precision.	The paper mentions several accuracy metrics across various studies conducted by researchers and academicians, some notable results are as follows: Highest reported accuracy: 99.04% using 1-dimensional convolution neural network Random Forest performance: 97.5% accuracy Basic model accuracy: 77% using SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting classifiers
23	Diabetes Prediction using Machine Learning	G Parimala, R. Kayalvizhi, S. Nithiya	Conference	India	2023	The study employs various classification and ensemble learning algorithms in order to develop a method for predicting diabetes. They use models such as Random Forest, K-Nearest Neighbors (KNN), Label Encoder, and train-test split, to build classification models for the diabetes dataset. The study utilizes the PIMA Indian Dataset for their methodology. The paper includes detailed data preprocessing steps, such as handling missing values and addressing data imbalance using oversampling techniques. Random Forest gave an impresssive accuracy rate of 98% during the testing.	The study relies on a single PIMA Indian Diabetes dataset which means that the model's performance is heavily dependent on the quality and characterstics of this specific dataset. The results may not generalize well to other populations causing problems in model evaluation. The complexity of some models, like Random Forest may pose challenges in understanding and explaining the decision-making process.	The study evaluates various machine learning algorithms, including Support Vector Machines (SVM), k-nearest Neighbors (KNN), and decision trees to enhance predictive accuracy. The authors utilize the PIMA Indian dataset for their study. The accuracy is checked for KNN Classifier, Random Forest, Decision Tree and SVM separately. The performance evaluation, criterias include accuracy, precision, recall and F1-score.	KNN Classifier: Accuracy: 76.56% Precision: 78.8% Recall: 76.5% F1-Score: 77.6% Random Forest: Accuracy: 98% Precision: 98% Recall: 98% F1-Score: 98% Decision Tree: Accuracy: 96% Precision: 95% Recall: 98% F1-Score: 97% SVM: Accuracy: 65% Precision: 63% Recall: 97% F1-Score: 77%
24	AI-Driven Early Diabetes Prediction	D. Sehgal, I. Kaur, V. Sharma, B. Gautam, A. Singh and N. Kumar	Conference	India	2024	The study demonstrates various machine learning algorithms like Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbours (KNN), Naïve Bayes and Logistic Regression where KNN was found to be the best performing among all models with 75% accuracy. The use of Flask for real time assessment of risk makes it quite accessible for healthcare proffesionals and patients as well.	The dataset quality is a major concern, as biases in the training data could affect model generalizability and fairness. Additionally, the study acknowledges challenges related to model bias, which may arise from imbalanced datasets or inherent biases in medical data collection. Another limitation is the computational complexity of some ML models, which may hinder real-time deployment in resource-constrained environments. Also, while the study discusses deployment using Flask, it does not address large-scale implementation challenges, such as integrating the model into existing healthcare infrastructures.	The study begins with data preprocessing to ensure model reliability. The methodology is then followed by employing the various ML algorithms like Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbours (KNN), Naïve Bayes and Logistic Regression. Each model is tested on the PIMA Dataset which contains key features such as glucose levels, BMI, blood pressure, insulin levels, and age. The performance of the model is evaluated using F1 score and AUC-ROC. o facilitate real-world implementation, the best-performing model is deployed using Flask, a lightweight web framework that enables real-time diabetes risk assessment. The deployment process involves model serialization using Pickle or Joblib, API endpoint creation, and testing to ensure stability and usability.	Model-Accuracy(%) Logistic Regression-73.148 KNN-75.000 SVM-74.047 Decision Tree-68.180 Naive Bayes-73.148 Random Forest-72.5

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
25	A Comprehensive and Comparative Examination of Machine Learning Techniques for Diabetes Mellitus Prediction	Ajay Kumar; Anmol Singh Gill; Jay Prakash Singh; Debolina Ghosh	Conference	India	2024	The paper provides a systematic and thorough comparison of multiple machine learning classifiers (LR, SVM, DT, RF, and k-NN) for diabetes prediction, allowing readers to understand the relative strengths of each approach. This comparative analysis is particularly valuable for researchers and practitioners in the field. The research employs robust validation techniques, specifically using 10-fold cross-validation, which helps ensure the reliability and generalizability of the results. This methodological rigor strengthens the paper's findings. The authors give careful attention to data preprocessing and exploratory data analysis, including handling missing values, outlier detection, and correlation analysis. This comprehensive data preparation approach helps ensure the quality of the input data for the machine learning models. The study uses the well-established PIMA Indians Diabetes Dataset, which allows for reproducibility and comparison with other studies in the field.	The study relies solely on the PIMA Indians Dataset, which may limit the generalizability of the findings to other populations with different demographic characteristics. This represents a significant limitation in terms of external validity. The paper doesn't thoroughly address the interpretability of the models, particularly the Random Forest classifier. While it achieves high accuracy, there's limited discussion of how the model makes its predictions, which is crucial for medical applications. The research doesn't extensively explore feature importance or selection methods, which could provide valuable insights into which health indicators are most crucial for diabetes prediction. The paper doesn't address class imbalance issues in the dataset, which can be a significant concern in medical diagnosis applications.	The paper's Data Processing and Analysis includes comprehensive exploratory data analysis, outlier removal, and handling of missing values, along with feature correlation analysis to identify key predictors. In the Machine Learning Implementation, five different classifiers are applied using a 10-fold cross-validation approach, with an 80:20 train-test split to ensure robust evaluation. The Evaluation Framework employs multiple performance metrics, including accuracy, precision, recall, and F1-score, along with confusion matrix analysis and ROC curve evaluation to assess model effectiveness and reliability in diabetes prediction.	Performance metrics for different classifiers: Random Forest: Accuracy: 92.23% Precision: 0.92 Recall: 0.83 F1-score: 0.87 KNN: Accuracy: 81.18% Precision: 0.72 Recall: 0.68 F1-score: 0.70 Decision Tree: Accuracy: 79.87% Precision: 0.75 Recall: 0.57 F1-score: 0.65 SVM: Accuracy: 75.38% Precision: 0.70 Recall: 0.43 F1-score: 0.53 Logistic Regression: Accuracy: 74.62% Precision: 0.67 Recall: 0.44 F1-score: 0.53
26	Machine Learning Approach for Diabetes Prediction using Genetic Algorithm based Feature selection	T.S. Ravi Kiran; A. Srisaila; G. Siva Shankar; Bodasingi Sowjanya; A. Lakshmanarao	Conference	India	2024	This research paper presents a novel approach to diabetes prediction by combining genetic algorithm (GA)-based feature selection with machine learning classification. It effectively addresses data imbalance through the ADASYN oversampling technique and demonstrates its efficacy on two distinct datasets, leading to significant accuracy improvements (DD-1: 84.5% to 90.3%, DD-2: 94.5% to 97.6%). By selecting optimal feature subsets (5 for DD-1 and 6 for DD-2), the GA reduces dimensionality while enhancing predictive performance. The study provides a well-defined methodology with clear workflow, addressing the critical issue of diabetes prediction in the Indian context. The GA-based feature selection offers a computationally efficient approach, and the clear visualization of class distributions and results further enhances the clarity and interpretability of the findings.	This research paper exhibits several limitations. It primarily focuses on accuracy as an evaluation metric, neglecting other crucial performance indicators such as precision, recall, F1-score, and others. The study lacks cross-validation results, which could provide a more robust assessment of model performance. Furthermore, it does not compare the proposed GA-based feature selection with other established techniques, limiting the understanding of its unique contributions. The paper lacks detailed descriptions of hyperparameter tuning for the Random Forest (RF) classifier and the optimization of GA parameters. It fails to provide an explanation of the rationale behind the specific features selected by the GA. The absence of an external validation dataset hinders the generalizability of the findings. Additionally, the paper lacks a comprehensive discussion of computational complexity and an analysis of model interpretability. Finally, the study does not include statistical significance tests to validate the observed improvements in accuracy.	This study investigates the effectiveness of genetic algorithm (GA)-based feature selection for diabetes prediction using two Kaggle datasets (DD-1 with 768 samples and DD-2 with 100,000 samples). Data preprocessing includes ADASYN oversampling to address class imbalance and splitting the data into training and testing sets. The GA is employed for feature selection, with a fitness function based on negative accuracy and a binary representation for feature selection. Random Forest is used as the base classifier, and its performance is compared with and without GA-based feature selection. The selected features for DD-1 are Pregnancies, Glucose, BloodPressure, Insulin, and Age, while for DD-2 they are Age, Gender, Hypertension, Heart Disease, Smoking History, and Blood Glucose.	Dataset 1 (DD-1): Base RF accuracy: 84.5% RF with GA feature selection: 90.3% Improvement: 5.8% Dataset 2 (DD-2): Base RF accuracy: 94.5% RF with GA feature selection: 97.6% Improvement: 3.1% Data Distribution: DD-1: 500 no-diabetes, 268 diabetes cases (pre-balancing) After ADASYN: 500 class 0, 474 class 1 DD-2: 91,500 no-diabetes, 8,500 diabetes cases (pre-balancing) After ADASYN: 91,500 class 0, 92,193 class 1 Training/Testing Split: DD-1: 779 training, 195 testing samples DD-2: 183,693 training, 36,739 testing samples
27	Diabetes Prediction Using Gaussian Naive Bayes and Artificial Neural Network	T Bharath Chandra; A Sujan Reddy; A Adarsh; M A Jabbar; B N Jyothi	Conference	India	2024	The ensemble learning method they developed achieves impressive accuracy (94.1%) by combining Artificial Neural Networks (ANN) and Gaussian Naïve Bayes (GNB), surpassing many previous models in the field. Their ANN model alone achieved an exceptional 98.7% accuracy, representing a significant improvement over existing solutions. The proposed idea provided a superior performance compared to other existing models like Random Forest (84%), SVM (79%), and Naïve Bayes (57%) The researchers employed a thorough data preprocessing approach, including proper handling of missing values, standardization, and dimensionality reduction through Principal Component Analysis (PCA). This comprehensive preparation helps ensure the model's reliability and robustness. The study includes detailed performance metrics beyond just accuracy, providing precision, recall, and F1-scores for a more complete understanding of the model's capabilities. Their ensemble model achieved 100% precision, indicating zero false positives - a crucial factor in medical diagnostics.	The research relies solely on the PIMA Indians diabetes dataset, which represents a specific ethnic group. This narrow focus raises questions about the model's generalizability to other populations with different genetic and environmental factors. The paper doesn't address the class imbalance problem that's common in medical datasets, which could affect the model's performance in real-world scenarios where the distribution of diabetic and non-diabetic cases might be different. While the mobile deployment is mentioned, there's limited discussion about the practical challenges of implementing such a system in clinical settings, including regulatory compliance and integration with existing healthcare systems. The research doesn't explore the interpretability of the model's decisions, which is crucial for healthcare applications where understanding the reasoning behind predictions is important for medical professionals.	This paper outlines a methodology encompassing several key components. Data preprocessing involves dataset cleaning, normalization, feature engineering, and dimensionality reduction using Principal Component Analysis (PCA). The data is then split into training and testing sets with an 80-20 ratio. The study explores three model architectures: an Artificial Neural Network with two hidden layers, a Gaussian Naïve Bayes classifier, and an ensemble model utilizing a majority voting classifier. The research culminates in the development and implementation of a Flutter-based mobile application for deployment, featuring a user interface for inputting features and a system for displaying the predicted outcomes.	Artificial Neural Network (ANN): Accuracy: 98.7% Precision: 100% F1-score: 98.2% Recall: 96.4% Gaussian Naïve Bayes (GNB): Accuracy: 89.6% Precision: 85.9% F1-score: 85.9% Recall: 85.9% Ensemble Model: Accuracy: 94.1% Precision: 100% F1-score: 91.4% Recall: 84.2%

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
28	Early Prediction of Diabetes and its Risk Factors based on ARIMA-ELMAN ANN Network	J. Senthil; Shaik Akbar; Akiladevi N; S Praveena; Ravindar K; Banupriya V	Conference	India	2024	The proposed ARIMA-ELMAN-ANN hybrid model achieves a high accuracy of 96.31%, demonstrating excellent performance in diabetes prediction. The model combines the strengths of three different approaches - ARIMA's capability to handle time series data, ELMAN's recurrent neural network architecture for temporal pattern recognition, and ANN's nonlinear modeling abilities. The research implements a robust data preprocessing framework that carefully handles missing values through multiple strategies (listwise deletion, attribute dropping, and pairwise deletion), making the model more reliable with real-world clinical data. The study introduces an innovative feature selection method using F-Score values, which helps identify the most relevant predictive factors for diabetes, potentially reducing computational complexity while maintaining accuracy.	The study doesn't specify the dataset size for training and validation, hindering assessment of generalizability across varying population sizes and demographics. It doesn't address potential bias in data collection or discuss model performance across different ethnic groups and age ranges. The research also suffers from methodological limitations, including limited discussion of computational resources required for real-world implementation, a lack of thorough comparison with other state-of-the-art diabetes prediction models, limited discussion of model interpretability crucial in healthcare, absence of external validation on different populations or datasets, missing details about handling class imbalance, and limited discussion of feature importance and their clinical relevance.	This research employs a comprehensive methodology. Data preprocessing involves handling missing values through listwise and pairwise deletion, along with data cleansing to address errors and inconsistencies. Feature selection utilizes F-Score based methods, followed by a wrapper component for further refinement through backward elimination of low F-score features. The model architecture incorporates ARIMA (Autoregressive Integrated Moving Average), ELMAN Neural Network, and Artificial Neural Network (ANN), with a hybrid integration of all three components for enhanced predictive performance.	Overall accuracy: 96.31% Training accuracy: Approximately 96.43% after 100 epochs Model building time: ANN: 19 seconds ARIMA with feature selection: 12 seconds Proposed hybrid model: 4.2 seconds Fitness function achieved optimal value of 0.021 after 18 iterations Training and validation loss shows continuous improvement without significant overfitting
29	An In-Depth Exploration of Machine Learning Algorithms and Performance Evaluation Approaches for Personalized Diabetes Prediction	Inderdeep Kaur; Aleem Ali	Conference	India	2024	The paper provides a comprehensive overview of machine learning applications in diabetes prediction, covering multiple algorithms and their implementations. The paper thoroughly examines both traditional methods like Logistic Regression and Support Vector Machines, as well as more advanced approaches like Deep Neural Networks and Quantum Machine Learning. Second, the research includes a valuable bibliometric analysis that shows research trends from 2013 to 2023, revealing the growing interest and developments in this field. The analysis shows China leading with 1048 publications, followed by the USA with 828 publications, providing insight into global research distribution. Third, the paper presents an extensive comparative analysis of different machine learning algorithms across various studies, with documented accuracy rates ranging from 77.37% to 98.9%. This comparison helps researchers understand which approaches might be most effective for their specific use cases.	The primary challenges in diabetes prediction using machine learning include data imbalance issues in diabetes datasets, limited availability of quality healthcare data, and difficulties in feature selection when dealing with high-dimensional data. Additionally, model generalization remains a challenge across diverse populations, while complex algorithms lack interpretability, making clinical adoption difficult. Privacy and ethical concerns also pose significant barriers, particularly regarding healthcare data usage. Moreover, temporal analysis and longitudinal studies face difficulties due to inconsistent data availability. Research gaps include limited integration of patient-generated health data, insufficient focus on explainable AI, and the need for better long-term risk assessment models. Furthermore, there is a lack of standardization in validation methodologies, which affects the comparability and reproducibility of different studies.	The paper discusses diabetes prediction using machine learning, categorizing diabetes into Type 1 (juvenile diabetes), Type 2 (the most common form), gestational diabetes (occurring during pregnancy), and pregestational diabetes (pre-existing diabetes before pregnancy). To enhance predictive accuracy, the study evaluates various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Neural Networks, Naïve Bayes, K-Nearest Neighbors, and Gradient Boosting Algorithms. Additionally, a literature review is conducted, providing a comprehensive analysis of recent publications, examining different methodological approaches, and discussing the various datasets used in diabetes prediction. This review highlights advancements in predictive modeling and the impact of diverse techniques on diabetes diagnosis and risk assessment.	The paper tabulated various accuracy metrics across different studies done by various researchers and academicians and conducted a comparitive study: Highest reported accuracies: 98.9% using LGBM and Random Forest (Ahamed et al., 2022) 98% using ensemble methods (Bhat et al., 2022) 95.83% using RFWBP (Ali et al., 2023) Lower range accuracies: 77.37% using SVM (Aslan et al., 2023) 80% using Random Forest (Abegaz et al., 2023) The paper emphasizes that accuracy varies significantly based on: The choice of algorithm Dataset characteristics Feature selection methods Preprocessing techniques Validation methodologies
30	Explainable Artificial Intelligence for Prediction of Diabetes using Stacking Classifier	Aruna Devi B; Karthik N	Conference	India	2024	The research employs a comprehensive preprocessing pipeline that addresses multiple data quality issues. It uses KNN imputation for missing values, OCSVM for anomaly detection, and SMOTE+ENN for handling imbalanced data, showing attention to data quality that many similar studies overlook. The stacking classifier approach combines the strengths of multiple algorithms (KNN, SVM, and XGB as base learners with Random Forest as meta-classifier), achieving an impressive 97% accuracy. This ensemble method helps overcome the limitations of individual algorithms. The integration of Explainable AI (XAI) through LIME makes the model's predictions interpretable for healthcare professionals, addressing the critical "black box" problem that often limits AI adoption in healthcare settings. The visualizations clearly show how different features contribute to each prediction.	The study relies solely on the PIMA Indian Diabetes dataset, which only includes female patients over 21 years old. This narrow demographic focus limits the model's generalizability to broader populations. The paper doesn't discuss the computational overhead of their complex ensemble approach or compare its runtime performance against simpler alternatives. This information would be valuable for practical implementation. While the paper mentions using LIME for explainability, it doesn't explore alternative XAI techniques like SHAP values, which might offer different insights into the model's decision-making process. The research doesn't address the potential bias in the original dataset or discuss how their preprocessing steps might affect model fairness across different demographic groups.	The paper's framework consists of three major components: Data Preprocessing Pipeline, Ensemble Model Architecture, and Explainability Layer. The Data Preprocessing Pipeline includes KNN imputation for handling missing values, OCSVM for anomaly detection and removal, and SMOTE+ENN for balancing the dataset. The Ensemble Model Architecture employs a stacking approach, where KNN, SVM, and XGB act as base learners, and Random Forest serves as the meta-classifier for combining predictions. To enhance model transparency, the Explainability Layer integrates LIME for local interpretability, provides visualization of feature importance, and offers detailed case-by-case explanations of predictions, ensuring better understanding and trust in the model’s decision-making process.	Accuracy: 98% Precision: 99% Recall: 98% F1 Score: 99%

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
31	An Ensemble Deep Learning Model for Diabetes Disease Prediction	Selma Aouamria, Djalila Boughareb, Mohamed Nemissi Zineddine Kouahla, Hamid Seridi	Journal	Algeria	2024	This study investigates the use of deep learning to predict diabetes. When paired with a soft voting classifier, three powerful deep learning models—Long Short Term Memory (LSTM), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN)—have been shown to function well together to enhance prediction performance. The study emphasizes the importance of early diabetes detection, which can significantly help with managing the illness and preventing complications. The authors also discuss the potential of integrating deep learning techniques with real-time data collecting for healthcare applications to increase forecast timeliness and accuracy.	There are certain difficulties with the study, especially when it comes to the interpretability of intricate models like CNN, DNN, and LSTM. It might be challenging for medical practitioners to comprehend and interpret the outcomes of these models because they can be somewhat of a mystery. Furthermore, the study depends on certain datasets, such as the Frankfurt Hospital Germany Diabetes Dataset (FHGDD) and the Pima Indian Diabetes Dataset (PIDD), which might not adequately represent the diversity of communities with varying demographics, lifestyle variables, and access to healthcare. This dependence on particular datasets may restrict how broadly the results may be applied to other populations.	Data collection, preparation, model selection, training, evaluation, and comparison of deep learning algorithms are all part of the paper's systematic approach. Records with a variety of variables, including age, blood pressure, BMI, insulin levels, and glucose levels, are included in the datasets used. Cleaning, dealing with missing values, and normalizing characteristics were all part of the data preprocessing process. Training and testing sets were created from the datasets. A soft voting classifier was used in the study to incorporate a number of deep learning algorithms, such as CNN, DNN, and LSTM. The models were assessed using performance indicators such F1-score, sensitivity, specificity, accuracy, and precision.	Accuracy: 99.81% Precision: 99.45% Sensitivity: 99.8% F1 Score: 99.72%
32	Recent Advancements Using Machine Learning & Deep Learning Approaches for Diabetes Detection: A Systematic Review.	Neha Katiyar, Hardeo Kumar Thakur, Anindya Ghatak	Systematic Review Paper	India	2024	Provides a comprehensive review of various Machine Learning and Deep Learning techniques for early Diabetes Mellitus detection and management. An in-depth analysis of traditional methods like SVM, KNN as well as advanced approaches like ANN, CNN are discussed. It systematically documents their performance metrics. For example, the performance values range from 68% accuracy for the retinopathy model using clustering and Naive Bayes to 99.78% accuracy for diabetes detection using neural networks and SVM. This comparitive analysysis helps researchers understand the strengths and limitations of various ML and DI techniques for research. The author of this paper propose integrating artificial intelligence (AI) and the Internet of Things (IoT) for future advancements.	The study highlights the importance of empiral-based research in prediction of Diabetes Mellitus but does not include such emperical studies which could have provided more insights	The study evaluates various machine learning algorithms, including Support Vector Machines (SVM), k-nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN), to enhance predictive accuracy.	Some of the performance metrics documented by the authors are: 99.78% using SVM, ANN by Shokrekhodei et al. 98% using LSTM (Long Short-Term Memory) by Thesis et al. 98.07% using KNN, 5-Fold Cross Validation by Suyanto et al. 96% using Random Forest, Fuzzy Neural Network by Thakkar et al.
33	Using Machine Learning-based SMOTE Analysis with the Light GBM Classification Method to Classify Diabetic Patients	Kanwarpartap Singh Gill, Vatsala Anand, Rahul Chauhan, Hemant Singh Pokhariya	Research Paper	India	2024	The study uses LightGBM (Light Gradient Boosting Machine) framework, known for its efficiency and high accuracy for diabetes classification along with SMOTE (Synthetic Minority Over-sampling Technique) for addressing the challenge of class imbalance. The study ustilizes PIMA Indian Diabets dates for their work and comes with an accuracy rate of 72%. ANOVA(Analysis of Variance) test is used for feature selection in the data preprocessing and SHAP (SHapley Additive exPlanations) is ised for model interpretability. SHAP values help in providing explanations for the predictions made by the model. As future prospect, the authors of the paper propose integration of explanable AI .	The research focuses on a specific dataset derived from National Institute of Diabetes and Digestive and Kidney Diseases, which includes which includes only adult females who are at least 21 years old and belong to the Pima Indian ethnic group. This limitation means that the results may not generalize well to other populations. Also, the paper relies on ANOVA test for feature selection which may not capture all relevant features or interaction between features affecting the model's performance and accuracy. Lastly, the paper does not provide detailed insights into the practical implementaion of the model highlighting the gap between research findings and their real world applications.	The study utilizes the PIMA Indian dataset which contains at 21 year old adult females. Patient charactersticks like age, gender, BMI, blood pressure and glucose levels are considerd in the study. For the purpose of classifying diabetes condition, LightGBM, a gradient boosting framework is used and for tackling the challenge of imbalanced data, SMOTE Analysis is used. The feature selection is done through ANOVA test. Also, examination of both balanced and unbalanced datasets are conducted. The model performance is assessed through measures like accuracy, precision, recall, F1-score and ROC-AUC. The model interpretability is given by SHAP which helps in providing explanations for the predictions made by the model.	Accuracy: 72% Precision: 68% Recall: 72% F1-Score: 70%

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
34	AI/ML-Based Diabetes Application using Hybrid Grey Wolf and Dipper Throated Optimization Algorithm	Rahul Dattangire, Saigurudatta Pamulaparthyvenkata, Shreekant Mandvikar, Anandaganesh Balakrishnan, Pradeep Chintale	Research Paper	USA	2024	The authors in this study employ Grey Wolf Optimization (GWO) algorithm along with Dipper Throated Optimization (DTO) algorithm to enhance feature selection for diabetes prediction. They utilize the PIMA Indian Diabetes Dataset for their findings. Also, they use normalizing techniques like Min-Max scaling for handling the missing values in data preprocessing. For classification, the study employs a Convolutional Autoencoder (Conv-AE), a model that learns to recognize patterns in data through encoding and decoding processes. They achieve an impressive accuracy rate of 99.10% and surpassed traditional techniques like CNN and Naïve Bayes.	The paper has several limitations, primarily relying on a single dataset (PIMA Indian Diabetes Dataset - PIDD) for validation, which may limit the generalizability of results to other populations. While it addresses class imbalance using the GWDTO algorithm, the inherent imbalance in the dataset still poses challenges. The study lacks a detailed cost-benefit analysis of implementing the proposed method in real-world settings and does not provide cross-validation results or ablation studies to assess the contribution of individual components. Additionally, there is limited comparison with other hybrid optimization techniques, no discussion on computational complexity or processing time, and no analysis of model interpretability. The study also fails to address scalability concerns, hyperparameter optimization, and provides limited discussion on handling imbalanced data, further impacting its applicability.	The authors use PIMA Indian dataset in their study which contains essential data points such as glucose levels, blood pressure, and BMI. In the preprocessing of data, normalization technique like Min-Max scaling is used to scale the data between 0 and 1 and handle missing values. Conv-AE is used for classification purposes and GWODTO is used for feature selection. Performance measures like accuracy, precision, recall and F1-score is considered here. The authors also evaluated the performance of WOA (Whale Optimization Algorithm), PSO (Particle Swarm Optimization, RNN, CNN, DNN and LSTM. The method is comprehensive - covering preprocessing, feature selection, and classification	The proposed GWDTO-ConvAE method achieved: Accuracy: 99.10% Precision: 97.32% Recall: 97.31% F1-score: 97.42% Specificity: 97.34%
35	Bridging Horizons in Diabetes Prediction: A Comparative Exploration of Machine Learning and Deep Learning Approaches in Pima Indian Women	L. Chandra Sekhar Reddy; Monikadevi Gottipalli; P. Sravanthi; J. Rajanikanth; Ganesh Yalamarthi; Neelima Gurrapu	Conference	India	2024	The authors of this study employs range of algorithms including Adaboost, XGBoost (Xtreme Gradient Boosting) and RNNs (Recurrent Neural Networks) to build predictive models for diabetes. Using the PIMA Indians Diabetes Dataset, the research aims to enhance the the accuracy of diabetes prediction by combining ML technologies with anomaly detection methodologies such as IQR (Interquartile Range) for outlier detection.	This paper focuses on PIMA Indian Dataset which includes data from a specific population. That is, the results may not generalize well to other populations. Out of the three models, AdaBoost and XGboost could not show well results, only the RNN model achieved highest accuracy. The study does not provide a detailed cost-benefit analysis of the implementation which is essential for assessing its practical feasibility and scalability.	The study use ML and DL algorithms such as Adaboost, XGBoost, and Recurrent Neural Networks (RNNs) on PIMA Indian Diabetes Dataset to build predictive models for diabetes. They use detailed preprocessing steps like handling missing values, normalizing data using techniques like Min-Max scaling and adressing data imbalance using the IQR method. The performance evaluation is checked on parameters such as accuracy, precision, recall, F1 score, and AUC-ROC score. The model achieves high accuracy through the RNN model surpassing AdaBoost and XGBoost.	IQR with XGBoost: Accuracy: 70.8% Precision: 58.1% Recall: 65.5% F1-Score: 61.5% ROC-AUC score: 76.7% IQR with AdaBoost: Accuracy: 73.4% Precision: 62.5% Recall: 63.6% F1-Score: 63.1% ROC-AUC score: 78.6% IQR with RNN: Accuracy: 90.3% Precision: 88.5% Recall: 83.6% F1-Score: 85.9% ROC-AUC score: 85%
36	Effective Diabetes Prediction: Integrating Ensemble Learning with LIME for Robust Results	Syeda Aqsa Habib Gilani; Madiha Haider Syed; Adeel Anjum	Conference	Pakistan	2024	The authors of this study have utilized combination of ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, Decision Tree, Neural Networks and K-means clustering. They used the Diabetes Prediction Dataset available on Kaggle for their research and applied detailed data preprocessing, exploratory data analysis (EDA) to draw meaningful patterns. It achieves high accuracy of 97.21% with the Neural Network model which outperforms traditional methods. Techniques like Recursive Feature Elimination (RFE) were employed to identify the most significant features for diabetes prediction. The authors also implemented ensemble learning through a Voting Classifier, which combines the strengths of various models to improve accuracy. Local Interpretable Model-agnostic Explanations (LIME) were applied for model interpretability, providing insights into how individual features contribute to predictions.	This paper poses limitations like reducing the generalizability of the results to other population because it uses Diabetes Prediction Dataset which includes data from a specific population. Also, inherent imbalance poses a challenge though the paper adresses class imbalance using advanced techniques. The paper also relies on specific machine learning algorithms, which may not capture all relevant features or interactions between features, potentially affecting the model's performance and accuracy. Lastly, the paper does not provide detailed insights into the practical implementation of the model real world.	On the Diabetes Prediction Dataset available online, the authors used various ML algorithms like RF, SVM, Naive Bayes, Decision Tree, Neural Networks and K-means clustering. The dataset comprises of factors like age, gender, BMI, hypertension, heart disease status, smoking history, HbA1c levels, and blood glucose levels. EDA is used to examine the distribution and relationships within the dataset while the RFE is used along with Logistic Regression estimator to identify key features. The ensemble learning is employed with a Voting classifier for accurate predictions and LIME for model interpretability. The prediction evaluation parameters used in the study are accuracy, ROC-AUC, precision, recall, and F1 score.	Logistic Regression with RFE: Accuracy: 95.99% Precision: 89.53% Recall: 68.5% F1-Score: 74.47% ROC-AUC: 96.09% SVM with RFE: Accuracy: 96.30% Precision: 97.3% Recall: 69.14% F1-Score: 76.28% ROC-AUC: 92.87% Random Forest with RFE: Accuracy: 97.06% Precision: 96.04% Recall: 84.28% F1-Score: 89.13% ROC-AUC: 96.52% Gradient Boosting with RFE: Accuracy: 97.25%

# Reference Number	Name of Paper	Author	Type of Paper	Country	Year	Advantages	Limitations	Main Components	Results
37	Diabetes Prediction Using Machine Learning Algorithm: A Comparative Analysis	Viksith Bardia; E Sophiya	Conference	India	2024	The authors of this paper use various ML models such as Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Neural Networks, AdaBoost, XGBoost, and a voting classifier for predicting diabetes. They use 2 datasets including Pima Indian dataset and the RTML dataset, which includes data from Bangladeshi women, to enhance the diversity and robustness of the dataset. The paper includes detailed data preprocessing steps and address class imbalance using SMOTE. The XGBoost algorithm achieved best accuracy of 83%. The integration of the LIME framework provides interpretable explanations of the model's predictions, making the results more transparent and understandable for healthcare professionals.	Some limitations of this paper include the use of mutual information feature selection which may not capture all relevant interactions between features which will ultimately affect the model's performance and accuracy. The study use SMOTE for addressing the issue of class imbalance but data might still introduce noise or overfit the model to the minority class. Though XGBoost achieved highest accuracy of 83.22%, the study does not explore deep learning models in depth which could outperform tree based models. Also, LIME which is used for model interpretability, lacks in depth comparision to technique like SHAP whic could have provided deeper insights into future. Additionally, for evaluation, ROC is an important factor but the study does not consider it among other parameters.	There are 2 datasets which are used in this study including the PIMA Indian Dataset and the RTML dataset consisting data from Bangladeshi women. The attributes in the combined dataset include Age, Glucose, Insulin, Blood Pressure, Pregnancy, Skin Thickness and BMI. They use detailed data preprocessing steps, such as handling missing values, standardization, and addressing class imbalance using SMOTE. The data is splitted into 80% for model training and 20% is reserved for evaluation. Also, mutual information feature selection is employed to identify the most relevant features that contribute to diabetes prediction and the LIME framework to provide interpretable explanations of the model's predictions. Various assessment crieterias taken are Accuracy, Precision, Recall and F1-score, ROC-AUC score.	Decision Tree: Accuracy: 73.87% Precision: 63% Recall: 59% F1-Score: 61% Logistic Regression: Accuracy: 75.6% Precision: 64% Recall: 69% F1-Score: 66% K-Nearest Neighbors (KNN): Accuracy: 68.18% Precision: 52% Recall: 75% F1-Score: 62% Voting Classifier: Accuracy: 77.34% Precision: 59% Recall: 67% F1-Score: 62%