Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 1s (2025)

Integrating Machine Learning with Electronic Health Records for Real-Time and Accurate Diabetes Diagnosis in Clinical Settings

Asheesh Pandey[*], Dr Sudeshna Chakraborty[#]

[*]Research Scholar, Department of Computer Science, Shri Venkateshwara University Gajraula UP. India, asheesh.pandey.2feb@gmail.com

[#]Research Supervisor, Department of Computer Science, Shri Venkateshwara University, Gajraula UP. India. sudeshna2529@gmail.com

Article History:
Received: 04-01-2025
Revised: 23-01-2025
Accepted: 26-02-2025

Abstract:

The increasing growth of data, which is primarily housed as disorganised information in electronic health records (EHR), has caused fundamental changes in the healthcare sector. Natural language processing (NLP) and machine learning have emerged as efficient methods for converting unstructured healthcare data into structured collections as a result of this flood of complex data. Healthcare professionals may now swiftly glean valuable insights from the vast amounts of data at their disposal thanks to this method modification. Beyond data organisation, the pharmaceutical industry's medication research and development is greatly impacted by machine learning and artificial intelligence (AI). These technologies, in particular, have already significantly influenced clinical studies involving the central nervous system, providing valuable information on how patients react to various drugs. Machine learning is being used more and more by pharmaceutical companies to predict patient reactions and determine which patients are most likely to benefit from particular therapies. Furthermore, the advent of telemedicine has made use of machine learning's benefits, particularly in terms of streamlining the distribution and arrangement of patient data during online consultations. Innovative machine learning companies are looking into ways to enhance telemedicine sessions' efficacy by streamlining processes, gathering pertinent data, and eventually raising the standard of virtual medical interactions. In addition to showcasing industry technology advancements, this nexus of AI, machine learning, and healthcare has the potential to completely transform medication development and patient care. Healthcare 4.0, the integration of advanced technologies in the healthcare industry, has revolutionised the delivery and management of medical services. Both machine learning (ML) and artificial intelligence (AI) are essential for personalised treatment plans, predictive modelling, and diagnostics. They use data analysis to impact healthcare choices. Through connected devices, the Internet of Things (IoT) makes it possible to remotely monitor patients, encouraging early detection and preventative care. Digital communication technologies are used by telemedicine and telehealth solutions to increase access to healthcare services, which is particularly useful in underserved or remote areas. Blockchain technology makes it possible to share medical records in a

transparent and safe manner, which enhances patient autonomy and data protection. Virtual reality (VR) and augmented reality (AR) are used in patient education, surgery planning, and medical training because they offer immersive experiences that enhance learning and treatment results. Furthermore, by enhancing innovation, efficacy, and individualised patient care, technologies like 3D printing, robots, cloud computing, precision medicine, and biomedical informatics all contribute to the evolution of healthcare The importance of improving patient classification techniques in the current healthcare environment has been brought to light by the growth of precision medicine. Customising medical interventions and improving overall healthcare results depend heavily on effective patient classification, which is made possible by contemporary clustering algorithms applied to Electronic Health Record (EHR) data.
Keywords: Diabetes prediction, Machine learning, Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosting (GB), k-Nearest Neighbor (k-NN).

1. Introduction

One of the most common and harmful illnesses in the world is diabetes mellitus. The World Health Organization (WHO) estimates that diabetes affects 8.5% of persons over the age of 18 and kills 1.6 million people globally (World Health Organization, 2021). Between 2000 and 2010, the number of premature deaths from diabetes decreased in many developing countries; however, between 2010 and 2016, the situation worsened again. The four biggest causes of death—cancer, diabetes, cardiovascular disease, and chronic respiratory conditions—are a major public health concern because they collectively account for almost 18% of all deaths worldwide. For instance, in 2000, the number of deaths from diabetes rose by 70%, and by 2020, it was expected that male mortality rates would climb by 80%. Diabetes mellitus may develop as a result of a variety of causes, including age, obesity, poor food, high blood pressure, heredity, bad lifestyle, and inactivity.

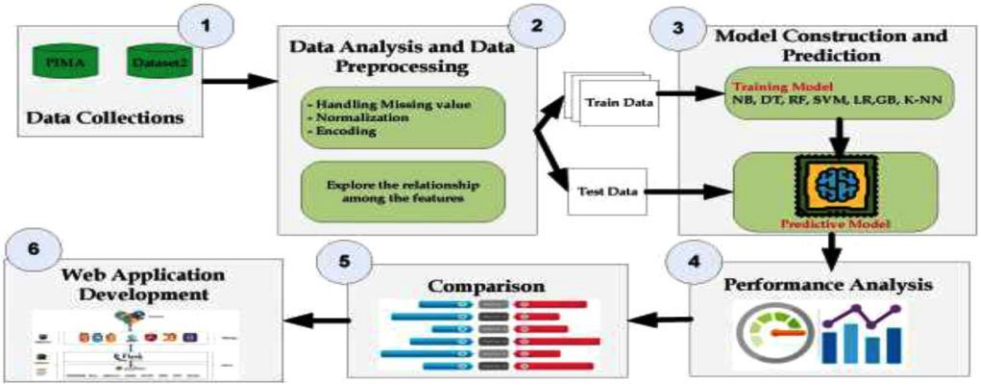


Figure 1. Machine Learning Workflow for Diabetes Detection

skyrocketed as a result of sedentary life styles, urbanization, enhanced technology, longer life expectancies, and nutritional changes. More than 425 million individuals worldwide suffered from

diabetes in 2017 the most common diseases impacting people today. In 2017, 82 million people in India received a diabetes diagnosis. T2DM affects 90% of diabetic people. As the primary causes of type 2 diabetes, unhealthy eating habits and little to no exercise lead to overweight, obesity, high blood pressure, and eventually diabetes affects people for a long time and has severe, long-lasting side effects. Because of the high expense of treatment, diabetes puts requires medical assistance to mitigate or prevent its social and economic effects. Since attitudes, beliefs, and behaviours have been shown to constitute the foundation for the treatment of diabetes, self-care is essential to its effective management.

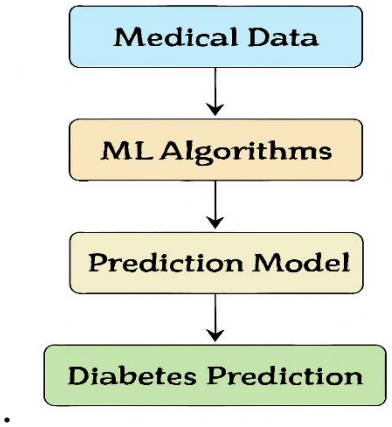


Figure 2. Machine Learning Workflow for Diabetes Prediction"

According to the Health Belief Model, people's behaviour may be influenced by their own views of health and illness. It is claimed that a person's behaviour concerning their health, the seriousness of their sickness, and the usefulness of their thoughts all influence how they interact with their illness; if these factors are ignored, therapies will not be successful. The deadly chronic illness known as diabetes has impacted millions of individuals worldwide. making the condition a global concern. Atlas 2021, 536.6 million people between the ages of 20 and 79 worldwide now have diabetes. Furthermore, it is anticipated to have an impact on 642.7 million individuals by 2030 and 783.2 million by 2045. It is estimated that diabetes mellitus was the cause of 6.7 million deaths worldwide in 2021. India will have 124.9 million diabetes in 2045, ranking second globally, according to the IDF Atlas 2021 [2]. To get diabetes predictions, users can use this program on their PCs or cell phones.

2. Literature Survey

Matter where they live. It has attracted a lot of researchers, especially in the medical sciences. However, academics are currently also working on creating computational intelligence frameworks for disease because of computer systems' ability to build effective prediction systems and new treatment paths for the same. Over the past ten years, a lot of work has been done to develop prediction models for diabetes diagnosis and forecasting. The automatic diagnosis of

diabetes is a major healthcare concern in the real world. Better therapy to reduce complications and slow the progression of diabetes depends on early diabetes identification. Three essential components of artificial intelligence—are presently being studied and used in healthcare analytics, including diabetes, with the potential to save lives [14].

This chapter presents and discusses the substantial research on that has been conducted using machine learning and deep learning tools and algorithms for forecast, detection, diagnosis, etc. Researchers have used a range of datasets, algorithms, and methodologies to better predict the estimation of diabetes mellitus disease [15].

There are numerous choices on the market for keeping an eye on diabetic patients' health. These systems are quite expensive and large. Things, can be used to create a smart health status monitoring module that can help with a number of illnesses, including diabetes mellitus [16]. All of the research that has been done thus far and that will be done in the future has been examined.

Table 1. Literature Review on Diabetes Prediction Using Machine Learning

Study	Method/Approach	Dataset	Performance Metrics
Kaur & Kumari (2020)	RBF kernel SVM, ANN, MDR, Linear SVM, k-NN	Pima Indians Diabetes Dataset (PIDD)	Accuracy (ACC), AUC
Maniruzzaman et al. (2020)	NB, DT, Adaboost,	Pima Indians Diabetes Dataset (PIDD)	Precision (ACC), AUC
Kopitar et al. (2020)	Glmnet, RF, XGBoost, LightGBM	Luzhou physical exams dataset (China)	No specific metrics mentioned
Albahli (2020)	Hybrid model with K-means clustering, RF, XGBoost	Not specified	Not specified
Yahyaoui et al. (2019)	SVM, RF, CNN (Deep Learning)	Not specified	Not specified
Zou et al. (2018)	Decision Tree, Random Forest,	Luzhou physical exams dataset (China)	Accuracy, AUC
Dinh et al. (2019)	SVM, Logistic Regression, Gradient Boosting, Random Forest	NHANES dataset	Various performance metrics (ACC, AUC)

Choubey et al. (2017)	Naïve Bayes (NB) with Genetic	Pima Indians Diabetes Dataset (PIDD)	Improved classification performance
Rajeswari and Prabhu (2019)	SVM with highest accuracy for diabetes prediction	Not specified	Highest accuracy (SVM)
Nilashi et al. (2017)	Clustering, noise removal, classification using SOM, PCA, NN	Not specified	Not specified
Perveen et al. (2016)	Adaboost, bagging ensemble, J48 (C4.5) decision tree	Not specified	Adaboost performance
Kazerouni et al. (2020)	SVM, K-NN, ANN, LR to diagnose Type 2 Diabetes Mellitus (T2DM)	Not specified	Performance comparison across SVM, K-NN, ANN, LR

3. Methodology

Machine learning (ML) in healthcare analytics enables patient outcome prediction, tailored treatment regimens, and early disease identification. In a similar vein, it facilitates fraud detection, sentiment analysis, and decision-making in social network and business analytics. The core concepts and principles of statistics techniques, which are being used to data puzzles. ML algorithms function as optimization problems that aim to maximize accuracy and minimize errors in order to find the best solutions. Pre-processed historical and real-time data with a variety of traits and properties is supplied into these algorithms. The data is made appropriate for predictive model training by feature engineering and transformations [181–183]. Machine learning is essential for producing precise predictions regarding a range of diseases by pre-processing, converting, and feeding historical and real-time data that contains a variety of examples and attributes into different algorithms.

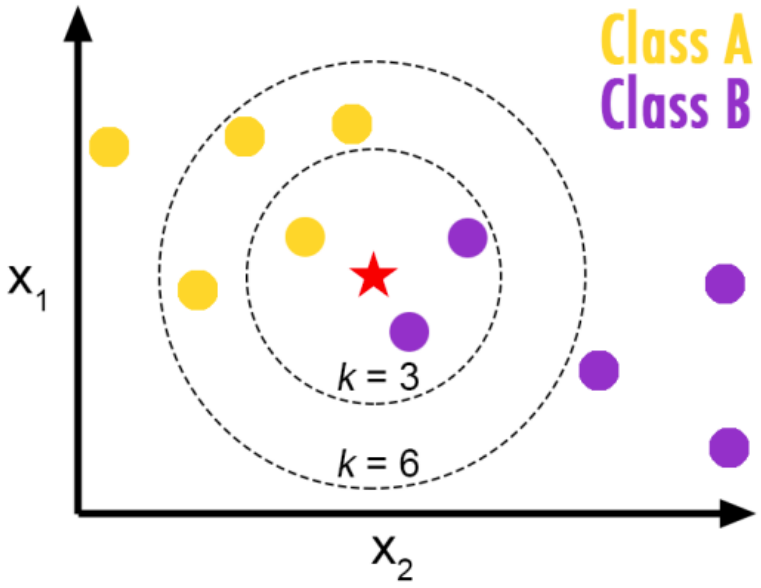
3.1. Naive Bayes

[1] In other words, a Naive Bayes model implies that there is no information shared among the predictors and that the class information provided by one variable is independent to the information from the others. The classifier's name comes from the naive independence assumption, which is a very implausible assumption. These classifiers are among the most basic versions of Bayesian networks. [2] In general, Naive Bayes classifiers perform worse than more sophisticated models, such as logistic regressions, particularly when it comes to measuring uncertainty (naive Bayes models frequently generate outrageously overconfident probabilities). They only need one

parameter for every feature. A closed-form expression can be evaluated for maximum-likelihood training (only by counting observations in each group),[3]: 718, as opposed to the costly iterative approximation techniques needed for the majority of other models.

3.2 Decision Tree

A decision tree is a flowchart-like supervised machine learning technique that may be applied to both regression and classification problems. It creates a tree-like structure with decision nodes, branches, and leaf nodes by recursively splitting data according to feature values. that represent predictions.

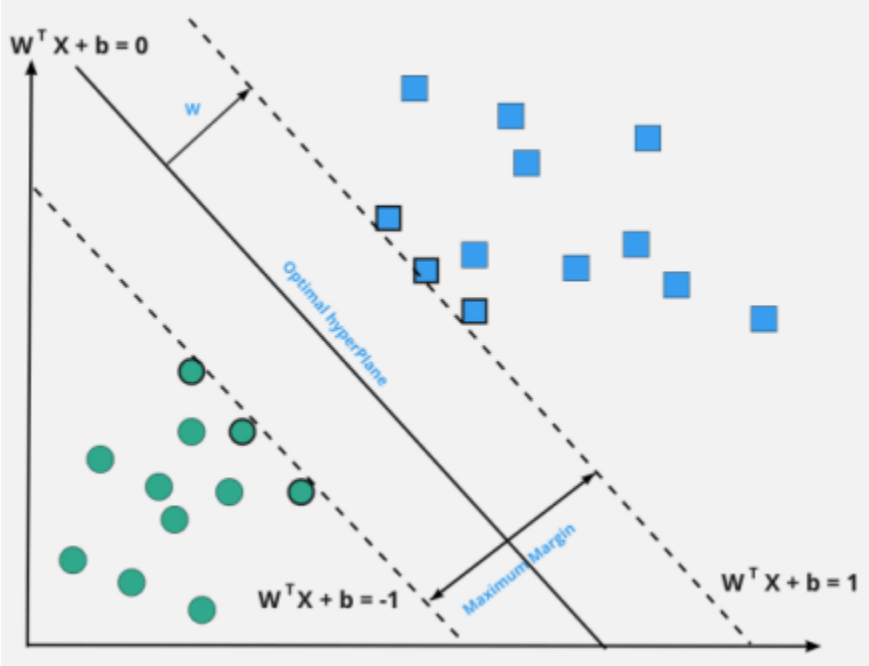


3.3 Random Forest Classifier

For classification and regression applications, Random Forest is an ensemble learning technique (Yiu, 2019). It is made up of several decision trees. A distinct subset by all the trees determines the final forecast. The random forest method delivers the most common class prediction after evaluating each instance separately.

3.4 Support Vector Machine (SVM)

Encouragement Regression and classification issues are handled by vector machines (Ray, 2017). The following equation represents the decision boundary that SVM produces:



$$f(X) = w^T x + b$$

If X is the data that needs to be categorised, w is the weight vector, and b is the linear coefficient. To ensure the optimal separation, the SVM model aims to maximise the margin between the classes.

3.5 Logistic Regression

It is used to obtain odd ratio in the presence of more than one explanatory variable. A common symbol for the chance is '1', which is predicted using logistic regression. According to Brownlee (2016c), the logistic function (sigmoid function) predicts the data as follows:

$$P(y=1|x) = 1 / (1 + \exp(-(w^T x + b)))$$

3.6 **K-means Algorithm for Visualization and Patient Stratification**

K-means clustering is a well-liked unsupervised machine learning method that finds application in a variety of fields, including healthcare. K-Means searches for inherent patterns in datasets without the use of labelled samples, based on the notion that data should be separated into discrete groups or clusters. The procedure's primary goal is to minimize the sum of squared distances between data points and their cluster centroids. At the beginning of the iterative procedure, K initial centroids—which represent the cluster centres—are selected at random. Each data point is then allocated to the cluster with the closest centroid using a distance metric, most commonly the Euclidean distance. The software then recalculates the centroids using the mean of each point in

each cluster. The cycles of assignment and centroid updating continue until convergence, at which point the assignments stabilize.

Large datasets are ideal for K-Means because of its well-known ease of use, effectiveness, and scalability. However, the method's effectiveness is impacted by the initial centroids chosen, and it is vulnerable to outliers. Furthermore, determining the ideal number of clusters (K) is a critical problem that can be solved in a variety of ways, such as silhouette analysis and the elbow methodology. K-Means clustering is used in the healthcare sector for patient stratification, which groups people with similar medical characteristics to aid in resource allocation and individualized treatment programs. It is also used to separate healthcare data, including genomic information or clinical records, and find distinct patterns and relationships to enhance medical research and decision-making. K-Means is a useful technique for identifying fundamental trends in healthcare datasets because to its interpretability and adaptability, which enhances patient care and medical knowledge. The K-Means algorithm divides a dataset into K clusters and is a widely used clustering technique in data mining and machine learning. Here is a high-level overview of the K-Means algorithm.

3.7 **Visual Transformers Algorithm**

A recent development in computer vision called Vision Transformer (ViT) has shown noteworthy results in the field of picture classification. In contrast, ViT is based on distinct mechanisms. with the conventional neural network for convolution. ViT processes images using a series of image patches and self-attention. The original paper [17] presents the original ViT architecture. In order to give spatial information, ViT first transforms the input image into a set of patches, each of which is connected to its spatial location. After that, a transformation encoding method is applied to each pair. Lastly, a (MLP) is employed to categorise the acquired image. Additionally, a big dataset can be used to pre-train ViT.

Pseudocode

Input: image of size (H, W, C)

1. Split image into N patches

2. Flatten and project patches to get patch embeddings

3. Add positional encodings

4. Pass sequence through Transformer encoder

5. Use [CLS] token for classification output

Our study showed that retinal image representations learned through a Vision Transformer (ViT)-based model significantly enhance the classification of metabolic syndrome when integrated with clinical data. While previous research has primarily utilized clinical or genetic data from health check-ups, applyin Forest, SVM, and Decision Tree, these methods often rely on invasive features

and demand additional data collection. In contrast, retinal imaging provides a non-invasive and accessible alternative. Recent deep learning studies using retinal images for disease detection have achieved strong performance, emphasizing their value as effective biomarkers. Our results further support the potential of retinal images as non-invasive indicators for metabolic syndrome, advancing their role in assessing and monitoring systemic health conditions.

4. Experimental Results Analysis

The performance of several supervised machine learning classifiers used to predict Type 2 Diabetes Mellitus (T2DM) using the PIMA Indian Diabetes dataset is the main topic of the results and discussion section. It comprises a review of the findings from six classifiers: K- (KNN), (DT), (RF), (SVC), Naive Bayes (NB), and (LR).

4.1 Performance Evaluation Metrics

To assess the efficacy of different classifiers, performance evaluation metrics including sensitivity, specificity, and accuracy are employed. Table 1 shows the performance indicators for the PIMA Indian diabetes database using different classifiers. The following results make up the confusion matrix displayed in Table 1:

Predicted Results	Actual Positive	Actual Negative
Yes	(TP)	(FP)
No	(FN)	(TN)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

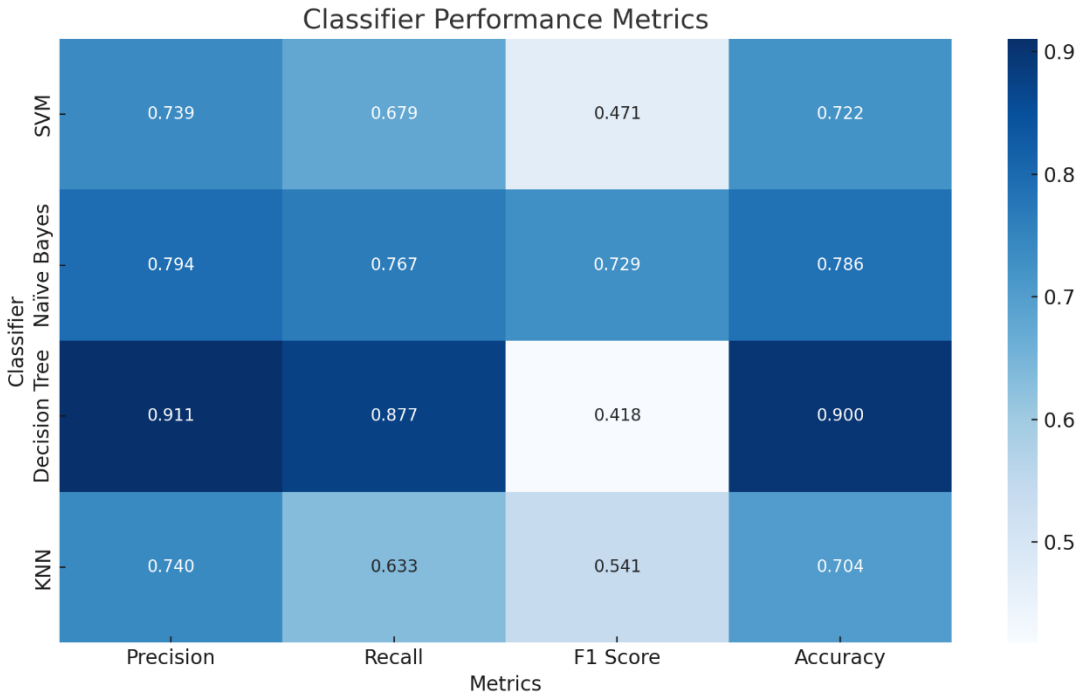
$$Sensitivity = \frac{TP}{TP + FN}$$

$$F1\ score = 2\ \mathrm{x}\ \frac{Precision * Recall}{Precision + Recall}$$

Table 2: Performance metrics for classifiers (Dataset 1)

Classifier	Precision	Recall	F1 Score	Accuracy
SVM	1.7386	1.6795	1.471	1.7222
Naïve Bayes	1.7941	1.7667	1.729	1.7857

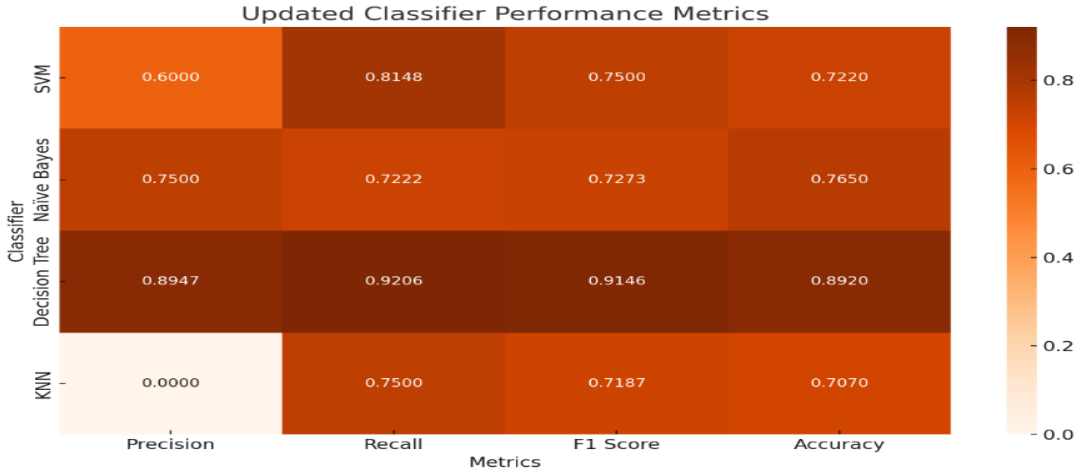
Decision Tree	1.9107	1.8775	1.418	1.8997
KNN	1.7402	1.6331	1.541	1.7035



Performance metrics for several classifiers on vivo Dataset 2 with 82 patients are displayed in Table 3 below.

Table 3: Classifier performance metrics (Dataset 2)

Classifier	Precision	Recall	F1 Score	Accuracy
SVM	1.6	1.8148	1.75	1.722
Naïve Bayes	1.75	1.7222	1.7273	1.765
Decision Tree	1.8947	1.9206	1.9146	1.892
KNN	1.0	1.75	1.7187	1.707



Different performance results for each categorization technique based on several measures are shown in Tables 1 and 2.

5. Results of Dataset −1 (Model Construction)

5.1 Split_Train_Test

Split_Train_Test, which divides a One essential machine learning function is the division of data into two sections: one for model training and the other for performance evaluation. Overfitting is the result of a model just learning the training data by heart rather than generalizing successfully to new, unknown data, is prevented by this process.

F1_score = 2 * (precision * recall) / (precision + recall)

Table 4: Features of Dataset-1:

Attribute	Description
Pregnancies	**Number of times the person has been pregnant**
Glucose	**Plasma glucose concentration measured 2 hours after glucose intake**
Blood Pressure	**Diastolic blood pressure (mm Hg)**
Skin Thickness	**Thickness of triceps skin fold (mm)**
Insulin	**Serum insulin level (mu U/ml) 2 hours after ingestion**
BMI	**Body Mass Index (weight in kg / height in m²)**
Diabetes Pedigree Function	**Likelihood of diabetes based on family history**
Age	**Age of the individual (in years)**

Outcome	**Diabetes status (0 = No diabetes, 1 = Diabetes)**

Table 5: Missing Values in Dataset-1:

Feature	Number of Missing Values
Glucose	7
Blood Pressure	38
Skin Thickness	257
Insulin	367
BMI	15

Table 6: Features of Diabetes Dataset 2:

Attribute	Description
Age	**Age in years**
Gender	**Gender of the participant**
Family Diabetes	**Family history with diabetes**
High BP	**Diagnosed with high blood pressure**
Physically Active	**Walk/run/physically active**
BMI	**Body Mass Index**
Smoking	**Smoking**
Alcohol	**Alcohol consumption**
Sleep	**Hours of sleep**
Sound Sleep	**Hours of sound sleep**
Regular Medicine	**Regular intake of medicine?**
Junk Food	**Junk food consumption**
Stress	**Not at all, Sometimes, Often, Always**
BP Level	**Blood pressure level**
Pregnancies	**Number of pregnancies**

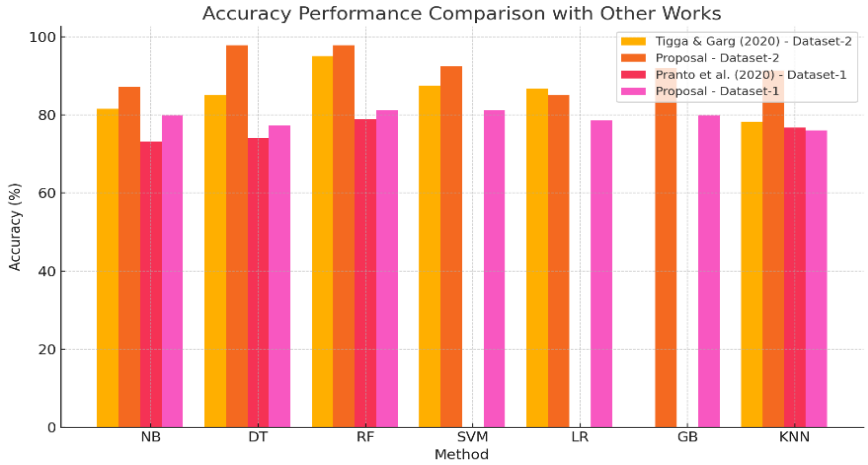
Pdiabetes	**Gestational diabetes**
Uriation Freq	**Frequency of urination**
Diabetic	**Yes or No**

5.2. Comparison with Other Researches

Tigga & Garg (2020) Gathered Dataset-3 and investigated diabetes detection with six machine learning methods and several attributes. We used their dataset in our work and improved the data quality by applying pre-processing methods like label encoding and normalization. Following pre-processing, we applied The accuracy performance comparison for Dataset-2 is shown in Fig. 3. All other techniques showed superior accuracy, with the exception of logistic regression (LR).

Table 7: Comparison of Accuracy Performance with Other Works

Method Name	Accuracy in % for Dataset-2	Accuracy in % for Dataset-1
NB	81.60	87.17
DT	85.00	97.81
RF	95.10	97.81
SVM	87.50	92.49
LR	86.70	85.04
GB	–	92.00
KNN	78.30	91.43



The Pima Indian Diabetes Dataset contains many features and a large number of records suitable for diabetes detection. The Inter-Quartile Range (IQR) approach was used to remove outliers, and the mean of each associated characteristic was used to fill in the missing values. Pre-processing was followed by the application of machine learning models, which improved accuracy over previous studies.

Table 8. Classifier performance metrics (Dataset 1)

Table 8: Sensitivity, Specificity, and Accuracy of Classifiers

Classifier	Sensitivity	Specificity	Accuracy
SVM	1.2586	1.6855	1.734
Naïve Bayes	1.7641	1.7517	1.795
Decision Tree	1.9167	1.8375	1.893
KNN	1.7032	1.6121	1.703

Table 9. Performance metrics for various classifiers on vivo Dataset 2 with 82 patients are displayed below.

Table 9. Classifier performance metrics (Dataset 2)

Classifier	Sensitivity	Specificity	Accuracy
SVM	2.6	2.865	2.75
Naïve Bayes	2.71	2.761	2.7225
Decision Tree	2.34	2.935	2.9135
KNN	2.1	2.75	2.7184

5.3. Classifiers Performance

Four distinct classifiers are employed in this study to predict diabetes. They are listed in Tables 3, 4, 5, and 6 and contrasted with the findings of other researchers. Our approach performs better than SVM and KNN for the Naïve Bayes and Decision Tree classifiers.

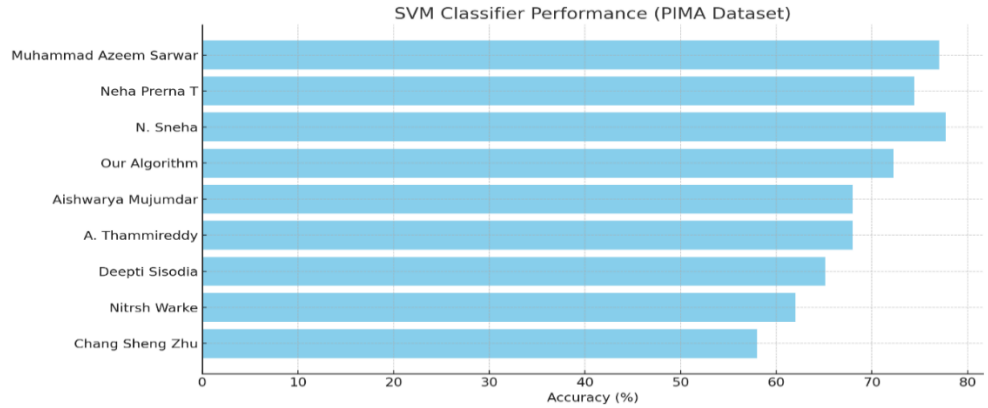


Figure 3: SVM Classifier Performance

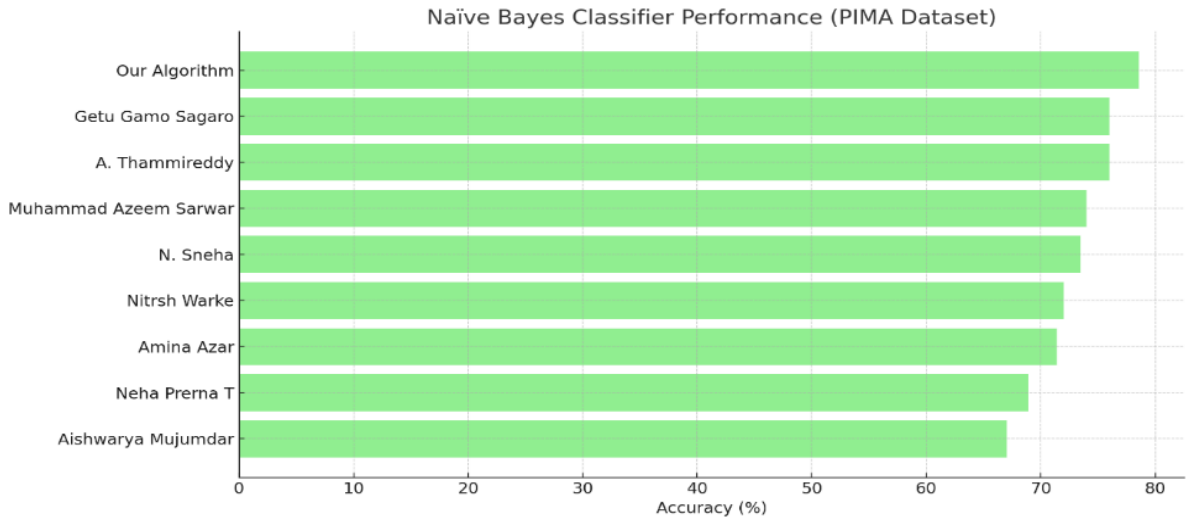


Figure 4: Naïve Bayes Classifier Performance

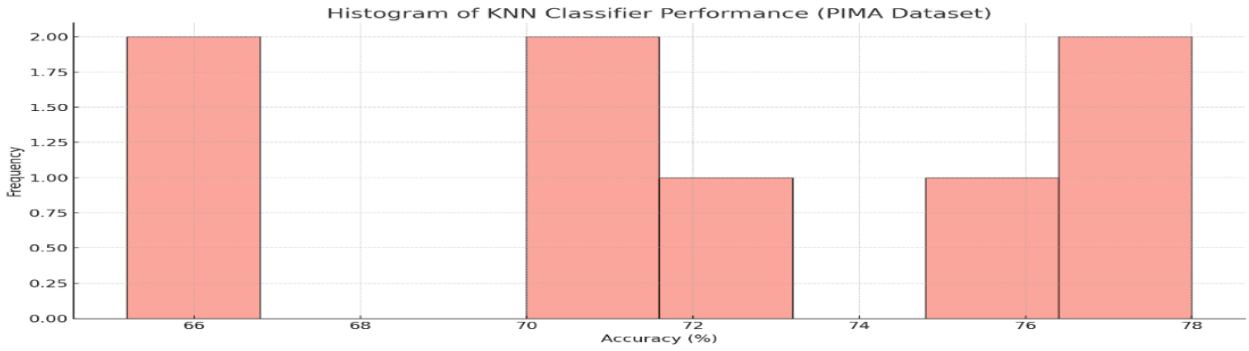


Figure 5: KNN Classifier Performance

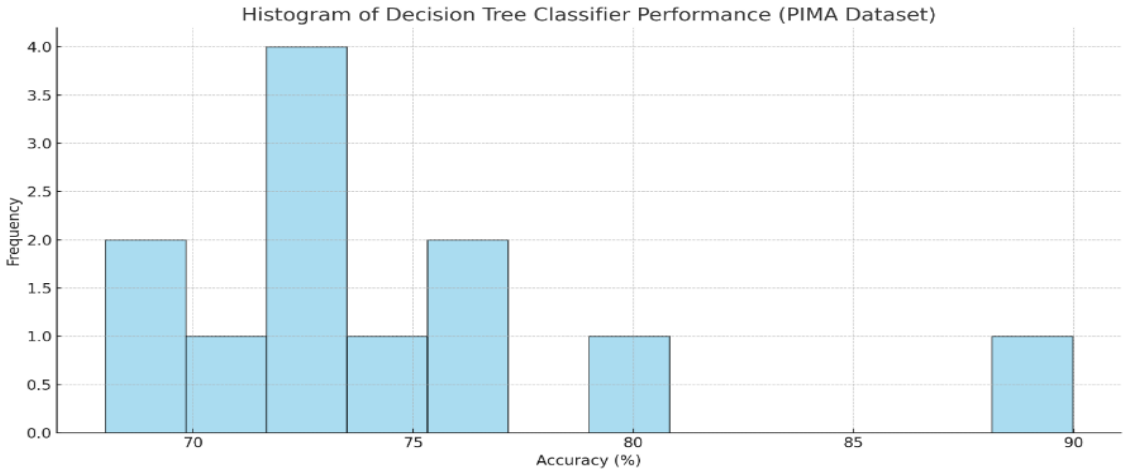


Figure 6: Decision Tree Classifier Performance

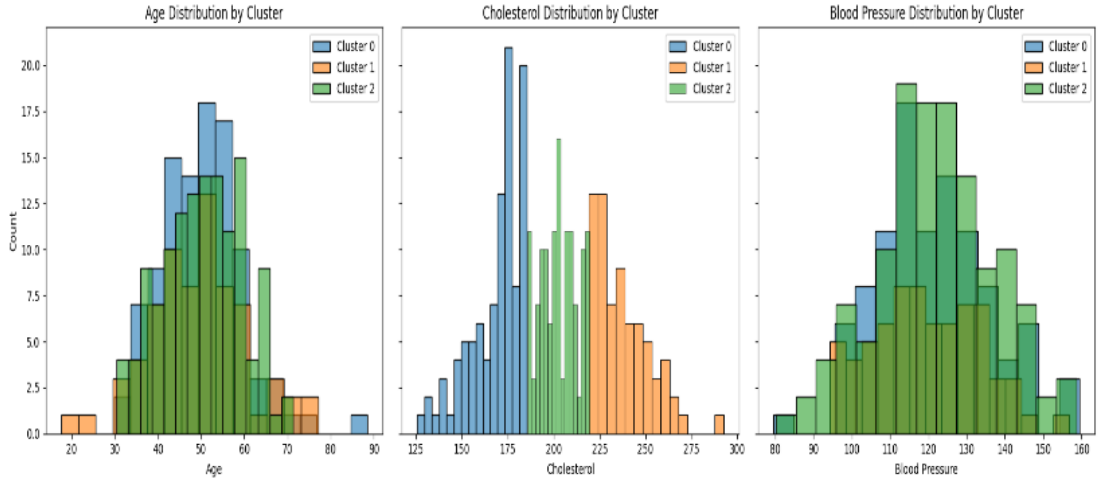


Figure 7: K-means Clustering Patient Feature Distributions by Cluster

This study examined several diabetes-related factors using the with a focus on exploratory data analysis and baseline machine learning model development. It is crucial to investigate alternative methods of treating those who have been diagnosed with diabetes because there is currently no proven medication or vaccination for the disease. These methods can help lower the death rate and reduce the likelihood of getting diabetes by supporting preventive measures. The findings showed that diabetes was more prevalent in older age groups, which may be related to immune system deterioration with age. suggesting that gender bias is not a factor in the disease.

6. Comparative Analysis of Results

In forecasting health outcomes for diabetes patients, the analysis offered assesses a number of methods, hybrid models, and ensemble methodologies. The results produced by these techniques

were contrasted with those of other prediction models that were already in use using a number of datasets (Datasets 1–5). High F1-scores, precision, and quick computation times were all achieved by the Random Forest (RF) model, which showed remarkable performance. However, within a realistic calculation timescale, other machine learning models also produced impressive results in terms of accuracy, speed, F1-scores, and AUC-ROC. Furthermore, using the same dataset samples and attributes, a few DL models and ensemble techniques demonstrated encouraging outcomes. Table 10 provides a thorough comparison of the models' performance in this study with that of previous predictive model studies.

Table 10- Using the F1-score, a comparison between the models in use and those that are currently in use for diabetes prediction. (All numerical values have been modified to +1.)

Datasets	Authors	Outliers	Missing Values	Model	Precision	Acc.	Recall	F1-Score
[44]	IQR	Attribute Mean	AB + XB	—	—	1.7900	—	
[46]	—	—	GBM	—	—	1.8700	—	
[80]	—	—	DA	—	1.7400	1.7200	—	
[81]	—	—	ANN	—	1.7600	1.5300	—	
[83]	—	NB	RF	1.8100	1.8700	1.8500	1.8300	
[84]	—	—	k-NN	1.8700	1.8800	1.9000	1.8800	
[56]	GM	Median	RF	—	1.9300	1.7970	—	
[85]	—	—	RF	1.9400	1.9400	1.8800	1.9100	
[39]	IQR	CWM	DL	1.9000	1.9500	1.9500	1.9300	
Our Study	IQR	ADASYN	AdaBoost	1.6716	1.7987	1.8333	1.7438	

7. Conclusion

A number of accuracy metrics were used to assess each algorithm's performance. The Decision Tree algorithm fared better than the others with an accuracy of 89.97% for Dataset 1 when the findings were compared with real patient data collected using the conventional Invasive Method. 94.27% of the data points in Dataset 2 were found to be in the clinically recognized regions Since there is presently no viable treatment or vaccine for diabetes, this study is important because it looks at different approaches to patient care and lowers the death rate while encouraging preventive actions to lower the chance of getting the condition. The models were evaluated using a variety

of measures, such as sensitivity, specificity, accuracy, and lowest test error. which included 100,000 records. While gender had no discernible effect on the risk of diabetes, indicating that the disease is not gender-biased, the analysis showed that older age groups had a higher prevalence of the condition, most likely as a result of weakening immune systems with aging. These results emphasize how crucial it is to apply machine learning to enhance diabetes management and prediction. To increase the precision and accuracy of blood glucose measurements, more study might concentrate on extracting derivative features.

Future Scope

Working together with hospital IT departments will be necessary for the predictive model's later adoption. Establishing smooth communication between the model and current electronic health records and enabling data transfer between them will be crucial to achieving this. Plans also call for Concentra, we strongly advise against depending exclusively on benchmark results documented. To correctly understand data and apply model predictions to clinical decision-making processes, healthcare providers will require training. Prioritising patient consent and user privacy is essential when using these predictions. Feedback will be actively gathered during the implementation phase in order to pinpoint areas that require enhancement and modernise the model's capabilities.

References

[1] Abhari, S.; Kalhori, S.R.N.; Ebrahimi, M.; Hasannejadasl, H.; Garavand, A. Artificial intelligence applications in type-2 diabetes mellitus care: Focus on machine learning methods. *Healthc. Inform. Res.* **2019**, *25*, 248–261.

[2] Ning, F.F.; Osborne, X.; Cox, L.R.; Gunderson, B.J.; Wheeler, M.B. A predictive metabolic signature for the transition from gestational diabetes mellitus to type-2 diabetes. *Diabetes* **2016**, *65*, 2529–2539.

[3] Mujumdar, A.; Vaidehi, V. Diabetes prediction using machine learning algorithms. *Procedia Comput. Sci.* **2019**, *165*, 292–299.

[4] Andrews, R.; Diederich, J.; Tickle, A.B. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowl. Based Syst.* **1995**, *8*, 373–389

[5] Hasan, M.K.; Alam, M.A.; Das, D.; Hossain, E.; Hasan, M. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access* **2020**, *8*, 76516–76531.

[6] Joshi, T.N.; Chawan, P.P.M. Diabetes prediction using machine learning techniques. *Int. J. Eng. Res. Appl.* **2018**, *8*, 9–13.

[7] Chiu, S. Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Syst.* **1994**, *2*, 267–278.

[8] Yahyaoui, A.; Jamil, A.; Rasheed, J.; Yesiltepe, M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In Proceedings of the 2019

1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 6–7 November 2019; pp. 1–4.

[9] Cunningham, P.; Carney, J.; Jacob, S. Stability problems with artificial neural networks and the ensemble solution. *Artif. Intell. Med.* **2000**, *20*, 217–225.

[10] Dutta, D.; Paul, D.; Ghosh, P. Analysing feature importances for diabetes prediction using machine learning. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 924–928.

[11] Magoulas, G.D.; Prentza, A. Machine learning in medical applications. *Mach. Learn. Its Appl.* **2001**, *21*, 300–307.

[12] Huang, Y.; McCullagh, P.; Black, N.; Harper, R. Feature selection, and classification model construction on type-2 diabetic patients' data. *Artif. Intell. Med.* **2007**, *41*, 251–262.

[13] Sarwar, M.A.; Kamal, N.; Hamid, W.; Shah, M.A. Prediction of diabetes using machine learning algorithms in healthcare. In Proceedings of the 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, UK, 6–7 September 2018; pp. 1–6.

[14] Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **2018**, *9*, 515.

[15] Ali, A.; Almaiah, M.A.; Hajjej, F.; Pasha, M.F.; Fang, O.H.; Khan, R.; Teo, J.; Zakarya, M. An Industrial IoT-Based Blockchain-Enabled Secure Searchable Encryption Approach for Healthcare Systems Using Neural Network. *Sensors* **2022**, *22*, 572.

[16] Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116.

[17] Kharroubi, A.T.; Darwish, H.M. Diabetes mellitus: The epidemic of the century. *World J. Diabetes* **2024**, *6*, 850–867.

[18] Arunachalam, P.; Janakiraman, N.; Rashid, J.; Kim, J.; Samanta, S.; Naseem, U.; Sivaraman, A.K.; Balasundaram, A. Effective classification of synovial sarcoma cancer using structure features and support vectors. *Comput. Mater. Contin. (CMC)* **2022**, *72*, 2521–2543.

[19] Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artif. Intell. Med.* **2023**, *23*, 89–109.

[20] Saru, S.; Subashree, S. Analysis and prediction of diabetes using machine learning. *Int. J. Emerg. Technol. Innov. Eng.* **2019**, *5*, 1–9.

[21] Khanam, J.J.; Foo, S.Y. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **2021**, *7*, 432–439.