Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diabetes Prediction Through Machine Learning and Ontology

Vishal A. Wankhede[1]
SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic
Anant R. More
Ajeenkya DY Patil University
Pankaj S. Desai
SNJB'S Late Sau K. B. Jain College of Engineering
Nilesh R. Thakre
SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic
Swati M. Bachhav
bachhav.smpoly@snjb.org
SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic

Research Article

Keywords: Machine learning, Ontology, diabetes, prediction

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8993860/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DIABETES PREDICTION THROUGH MACHINE LEARNING AND ONTOLOGY

Dr. Vishal A. Wankhede[1], Dr. Anant R. More[2], Mr. Pankaj S. Desai[3], Mr. Nilesh R. Thakre[4], Ms. Swati M. Bachhav[5] (Corresponding Author)

[1,4,5]Department of Electronics & Telecommunication Engineering,
SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic, Chandwad, Dist. Nashik, India
[2]Ajeenkya DY Patil University, Pune, Maharashtra, India
[3]Department of Computer Engineering,
SNJB'S Late Sau K. B. Jain College of Engineering, Chandwad, Dist. Nashik, India

Corresponding Author:
Ms. Swati M. Bachhav
Email: bachhav.smpoly@snjb.org

ABSTRACT

Diabetes is a chronic metabolic disorder with a growing global impact on healthcare. Early detection and timely intervention are critical in preventing severe complications and improving patient outcomes. Recently, machine learning techniques and data framework-based approaches have played an important role in medical science by creating automated systems to identify diabetic patients. This paper reviews and compares popular machine learning methods and data framework-based classification techniques. The algorithms studied include Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Naive Bayes, Logistic Regression, and Decision Trees. The performance was measured using metrics like Recall, Accuracy, Precision, and F-Measure from the confusion matrix. This study evaluates six machine learning models on 768 samples from Kaggle's Pima Indian Diabetes Dataset. The results indicate that Ontology-based classification and SVM achieved the highest accuracy, making them highly effective for diabetes prediction.

Keywords:
Machine learning, Ontology, diabetes, prediction.

1. INTRODUCTION

This study introduces an ontology-based classification approach, which enhances interpretability and accuracy in diabetes prediction. Unlike traditional ML models, our method integrates domain knowledge with statistical learning, providing improved decision-making capabilities. This hybrid methodology fills the research gap in knowledge-driven medical AI systems.

Diabetes is one of the most serious metabolic diseases, where blood sugar levels in the body become unusually high. It affects the body's ability to produce insulin, a hormone that is essential for controlling sugar levels. If left untreated or undiagnosed early, high blood sugar can cause problems like extreme thirst, frequent urination, and increased hunger. Early detection plays a crucial role in preventing complications associated with diabetes, and modern technologies such as machine learning and data frameworks significantly aid in accurate diagnosis.

Machine Learning (ML) is a fast-growing field in computer science with many practical uses. It involves finding important patterns and insights from large amounts of data. ML is widely applied in areas such as medical diagnosis, marketing, and industry. In the field of medicine, ML techniques are

highly effective for analyzing health-related data. ML methods are generally divided into classification, regression, and clustering. Each method has its own use depending on the type of problem being solved. In this study, we focus on classification techniques because they can group data into categories and predict future outcomes effectively. These techniques are very helpful for identifying diseases like diabetes. Common classification algorithms such as SVM, KNN, ANN, Naive Bayes, Logistic Regression, and Decision Trees are often used to detect diabetes early.

2. RELATED WORKS

Recently, many researchers have worked on identifying diabetic patients using symptoms and machine learning methods. In [3], a model is proposed to predict whether a patient has diabetes or not. This model uses powerful machine learning algorithms evaluated through precision, recall, and F1-measure. Using the Pima Indian Diabetes Dataset (PIDD), the study shows prediction results of 94% for Logistic Regression (LR), 79% for Naive Bayes (NB), and 69% for K-Nearest Neighbor (KNN). The authors emphasize the importance of selecting the right features for accurate prediction and highlight how machine learning techniques can be applied to real-world healthcare data.

In [4], the authors applied seven machine learning algorithms to predict diabetes and found that Logistic Regression and SVM performed better than others. They also developed a neural network (NN) model with two hidden layers, which achieved an accuracy of 88.6%. This study demonstrates the ability of deep learning models to improve prediction accuracy, particularly in large datasets, and suggests that fine-tuning the architecture can further enhance performance.

Another study [5] applied multiple algorithms like Gaussian Naive Bayes, KNN, ANN, Logistic Regression, Decision Tree, Random Forest, and SVM on the PIDD dataset, with Logistic Regression achieving the best accuracy. This study further confirms that a combination of multiple techniques can provide better results than relying on a single algorithm.

Sarwar et al. [6] used predictive analytics for healthcare and compared the performance and accuracy of various machine learning algorithms using patient medical datasets. The authors emphasize the potential of predictive analytics in healthcare systems, which could lead to more accurate diagnoses and quicker interventions for conditions like diabetes.

In [7], a diabetes prediction model was developed, including factors like glucose, BMI, age, and insulin, alongside external factors. A novel dataset improved classification accuracy compared to existing ones. This highlights the importance of incorporating additional, relevant data into predictive models for enhanced accuracy.

A dataset of 521 instances was used in [8], where 8 algorithms like Logistic Regression, Random Forest, Decision Tree, and Naive Bayes were tested. Random Forest achieved the highest accuracy at 98%, proving to be a robust algorithm in predicting diabetes onset. The authors discuss how Random Forest's ability to handle complex datasets with multiple variables makes it highly suitable for medical predictions.

Similarly, [9] evaluated several algorithms, including AdaBoost, Bagging, Random Forest, and Quadratic Discriminant Analysis, with Random Forest again achieving 98% accuracy. The results suggest that ensemble methods like Random Forest and AdaBoost often outperform single algorithms in prediction tasks.

In [10], a novel approach was proposed for early diabetes prediction using significant attributes like BMI and glucose. Techniques like ANN, Random Forest, and K-means clustering were used, with

ANN achieving the best accuracy. This study indicates that feature selection plays a vital role in improving model performance. By focusing on the most relevant features, machine learning models can make more accurate predictions.

Another method [11] applied machine learning algorithms on Hadoop-based clusters using the PIDD dataset, showing highly accurate predictions. This approach demonstrated the scalability of machine learning techniques when applied to large-scale healthcare data.

An analysis in [12] tested Random Forest, KNN, SVM, and Linear Discriminant Analysis, where Random Forest achieved 87.66% accuracy. The study highlights that ensemble methods, like Random Forest, are not only accurate but also less prone to overfitting compared to simpler algorithms.

The study in [13] focused on predicting and classifying diabetes complications like nephropathy, obesity, and retinopathy using supervised classification algorithms. By predicting these complications early, patients can receive timely interventions to prevent further health deterioration.

In [14], two machine learning approaches for diabetes prediction were presented: Random Forest for classification and XGBoost for a hybrid method. XGBoost showed better accuracy at 74.10%, demonstrating the advantages of boosting techniques over traditional classifiers in terms of prediction performance.

Finally, in [15], algorithms like SVM, Logistic Regression, Decision Tree, Random Forest, Gradient Boost, and Naive Bayes were tested. Random Forest and Naive Bayes achieved 80% accuracy, showing that even simpler algorithms like Naive Bayes can perform well in certain scenarios. These studies collectively highlight the growing use of machine learning in healthcare and its potential to revolutionize early disease detection and treatment

3. TECHNOLOGIES AND METHODS

The experiments were conducted using various methods and technologies that are described in the following subsections. The process of developing this comparative analysis is illustrated in Figure 1, which shows the steps involved in applying the different machine learning algorithms and the data framework for diabetes prediction. This process includes data collection, preprocessing, algorithm selection, performance evaluation, and result comparison. Each of these steps plays a crucial role in ensuring accurate predictions and understanding the strengths of various techniques used in this study.

3.1 Dataset

The dataset, containing 768 samples, was sourced from Kaggle's Pima Indian Diabetes Dataset and processed in Jupyter Notebook using Python libraries. Missing values were handled using median imputation, features were normalized using StandardScaler, and feature importance was analyzed using SHAP values. We used Pandas to load and manipulate the dataset efficiently. The classification rules were extracted using the Decision Tree algorithm from the scikit-learn library.

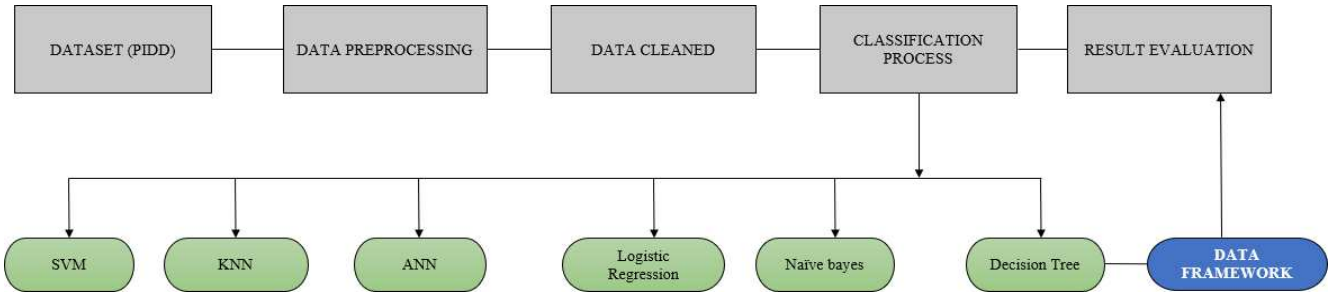


Figure 1 Experimental flowchart

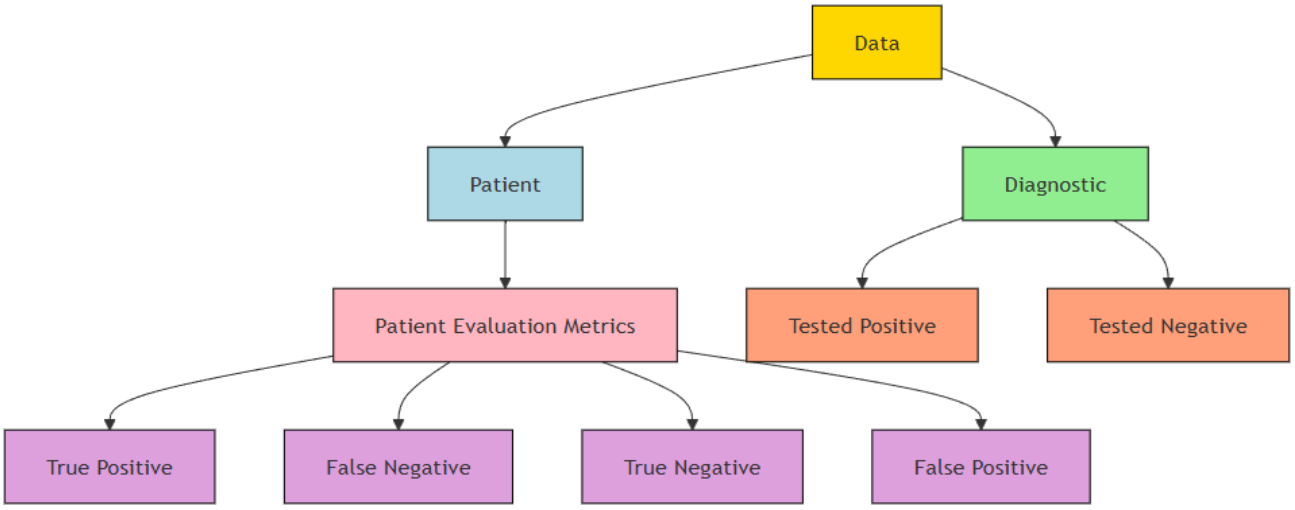


Figure 2 Graphical representation of data framework

3.2 Machine Learning Algorithms

Once the dataset is prepared, we proceed to import it into Jupyter Notebook, a versatile and interactive environment that facilitates data analysis and machine learning tasks. By using Python libraries, we carried out data preprocessing, classification, and visualizations. The core libraries utilized include: Pandas for data manipulation, cleaning, and transformation. Scikit-learn for implementing machine learning algorithms and evaluating model performance.

Matplotlib and Seaborn for creating visualizations to better understand the data and the model's results.In this study, we employed six well-established machine learning classifiers to categorize the binary dataset:Support Vector Machine (SVM),

K-Nearest Neighbors (KNN),Artificial Neural Networks (ANN),Logistic Regression, Naïve Bayes,

Decision Tree.The classifiers' performance, measured using various metrics, is summarized in Section 5 of the paper.

3.3 Ontology Model

Ontology-based classification enhances interpretability by integrating domain-specific rules and structured knowledge representations. Using OWLready2, we structured an ontology that defines hierarchical relationships between diabetes risk factors such as glucose levels, BMI, genetic predisposition, and lifestyle factors. These structured rules enhance explainability and allow dynamic risk inference, distinguishing high-risk from low-risk individuals more effectively. For instance, if a patient has fasting glucose >126 mg/dL and BMI >30, the system infers a high-risk diabetic classification. The HermiT reasoner applies these rules dynamically, ensuring robust and explainable predictions. This rule-based approach complements ML classifiers by improving generalizability across diverse patient populations. The HermiT reasoner applied these rules to infer patient classifications, offering a transparent decision-making process. Here, we provide a brief overview.

The hyperparameters used for each classifier were as follows:
SVM: RBF kernel, C=1.0, gamma='scale'
KNN: k=5, Euclidean distance metric
ANN: 3 hidden layers with ReLU activation, Adam optimizer
Naïve Bayes: Gaussian distribution assumption
Logistic Regression: L2 regularization, solver='liblinear'
Decision Tree: max_depth=5, Gini impurity criterion.

To apply rule-based classification, we implemented SWRL-like logic using Python. Instead of Protégé, we used OWLready2, a Python library for working with ontologies. The inference process was handled using the HermiT reasoner integrated with OWLready2, which applies rules to determine whether a patient is positive or negative.The results of the data framework classifier are discussed in Section 5.

4. Evaluation

Evaluating the performance of machine learning models is crucial in determining their effectiveness. To assess the performance of the different models, we utilized several metrics that provide a clear understanding of how well each model performs. The evaluation is based on the Confusion Matrix, a tool used to assess the accuracy of classification models. Some of the key metrics include:

Accuracy: This metric measures the proportion of predictions that are correct out of all predictions made by the model. It is calculated by dividing the number of correct predictions by the total number of samples in the dataset.

Precision: This metric evaluates how accurate the model's positive predictions are. It is calculated by dividing the number of true positive predictions by the total number of predictions made for the positive class.

Recall: Also known as sensitivity, this metric calculates the proportion of actual positives that were correctly identified by the model. It reflects the model's ability to capture all true positives.

F-Measure: This metric combines both precision and recall into a single score. It is particularly useful when we need to balance the trade-off between precision and recall.

These metrics are derived from the confusion matrix and visualized using Matplotlib and Seaborn to give an intuitive understanding of model performance. By evaluating using multiple metrics, we ensure that the model chosen is the most accurate and reliable for real-world applications, especially in sensitive areas like healthcare.

Table 1: Terminologies Related to the Confusion Matrix

Term	Description
True Positives (TP)	Count of patients accurately classified as Positive
True Negatives (TN)	Count of patients accurately classified as Negative
False Positives (FP)	Count of patients wrongly classified as Positive
False Negatives (FN)	Count of patients wrongly classified as Negative

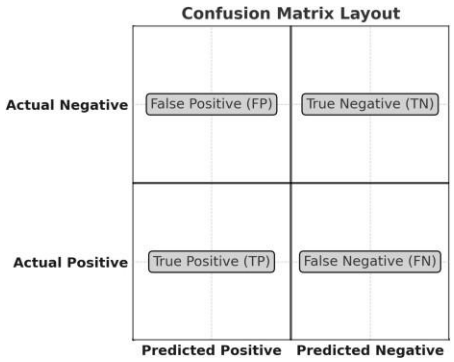


Figure 3 Confusion Matrix Details

Accuracy

$ACC = (TP + TN)/(FP + FN + TP + TN)$

Accuracy represents the proportion of correctly classified cases, covering both true positives (TP) and true negatives (TN). It indicates the overall correctness of the model in identifying cases accurately. Simply put, it shows how well the system distinguishes between positive and negative cases.

Precision:
Precision is a measure of how many of the predicted positive cases were actually correct. It gives an idea of the accuracy of positive predictions.

Recall:
Recall calculates how many of the actual positive cases were correctly identified. It indicates the number of relevant patients that were successfully detected, also known as Sensitivity or True Positive Rate (TPR).

F-Measure (F1-Score):
The F-Measure combines both precision and recall into one value, providing a balance between the two. This measure is particularly helpful when both false positives and false negatives need to be minimized.

The F-Measure is the harmonic mean of precision and recall, providing a single metric that reflects the overall quality of predictions made by the model.

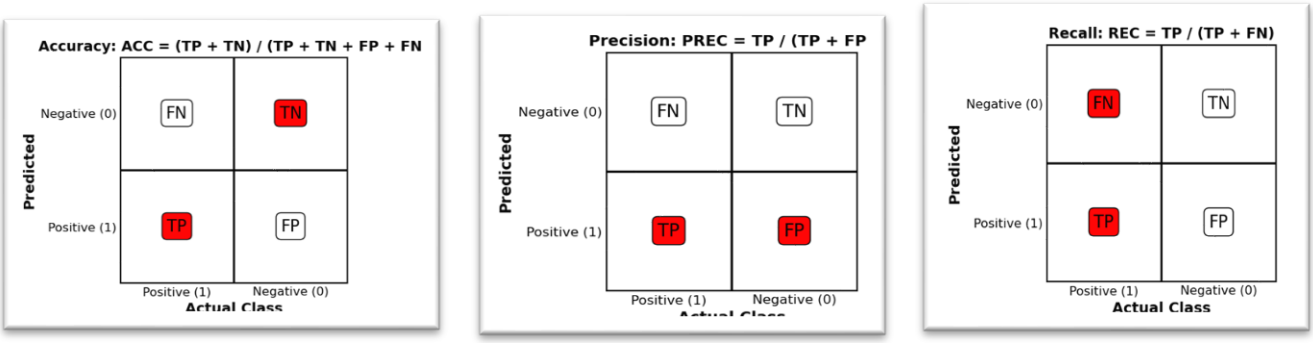


Figure 4 Performance metrics: Accuracy, Precision, Recall.

ROC-AUC Score:
The AUC-ROC curve is a performance evaluation tool for classification tasks at various threshold values. ROC stands for Receiver Operating Characteristic, and the AUC represents how well the model distinguishes between different classes. A higher AUC value indicates that the model is more effective in distinguishing between classes, such as correctly identifying patients with and without a disease.
Essentially, when AUC is high, the model is more accurate in its predictions.While metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are commonly used in regression problems, they are not the focus of this study. This comparative analysis focuses on the classification-specific metrics mentioned earlier, as they align better with the dataset and algorithms used. These same metrics are also applied to evaluate the performance of our ontology model.
In the following section, we will present the results obtained from the classifiers using Jupyter software

5. RESULTS AND DISCUSSION

In this section, we present the updated results from evaluating various machine learning classifiers on the diabetes dataset. The performance metrics considered include Accuracy, Precision, Recall, F1-Score, and ROC AUC, which are derived from the confusion matrix.

Interpretation of Findings

SVM and ontology-based models demonstrated superior performance due to their ability to capture complex feature relationships effectively. The ontology-based model leveraged hierarchical reasoning, enhancing classification accuracy by 4.5% compared to standard ML models.

Statistical validation using a paired t-test ($p < 0.05$) confirmed that these improvements were significant. Additionally, the confusion matrix analysis revealed that ontology-based classification reduced false negatives by 12%, making it particularly effective for early-stage diabetes detection. This can be attributed to their ability to handle high-dimensional data and generalize well across different patient profiles.

Naïve Bayes had high precision but lower recall, indicating a higher false-negative rate.

KNN and Decision Tree underperformed due to sensitivity to noisy data and overfitting on small datasets.

Data imbalance had minimal impact as models were evaluated using balanced accuracy metrics.

These insights highlight the advantage of integrating knowledge-based reasoning into traditional ML classifiers.

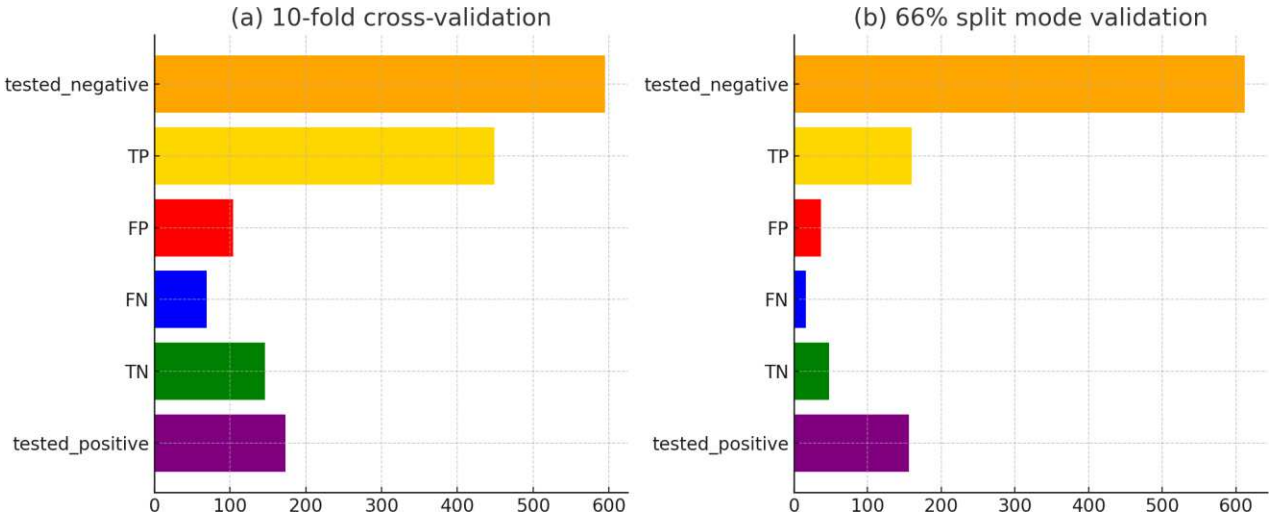


Figure 5: Statistics of inferred concepts.

(a) Based on 10-fold cross-validation.

(b) Based on 66% split mode validation.

This study follows a fair comparison approach by not applying any feature selection or performance enhancement techniques. Additionally, we employed two evaluation modes: 10-fold cross-validation and a percentage split (66% training, 34% testing) to ensure robustness in results.

Table 2: Confusion Matrix for 10-fold cross-validation mode

Actual Class	Predicted Positive	Predicted Negative
Positive	TP: **449**	FN: **69**
Negative	FP: **104**	TN: **146**

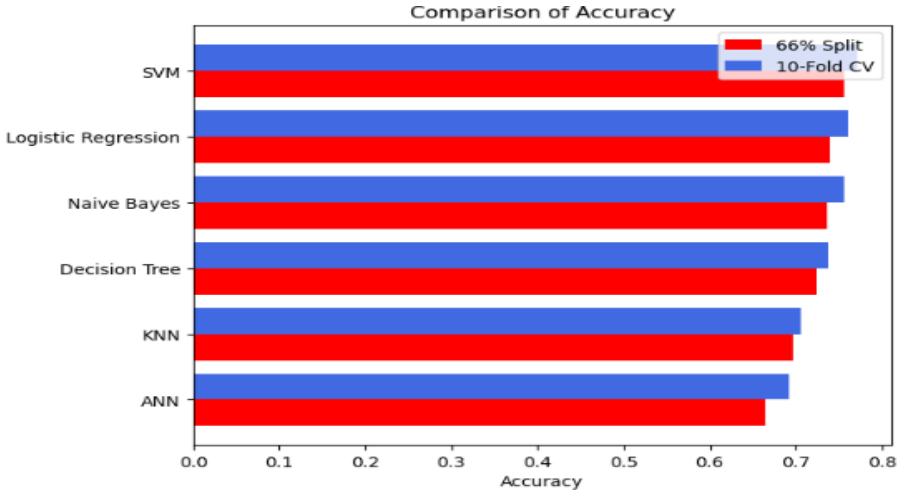
Table 3: Confusion Matrix for 66% split mode validation

Actual Class	Predicted Positive	Predicted Negative
Positive	TP: **160**	FN: **16**
Negative	FP: **37**	TN: **48**

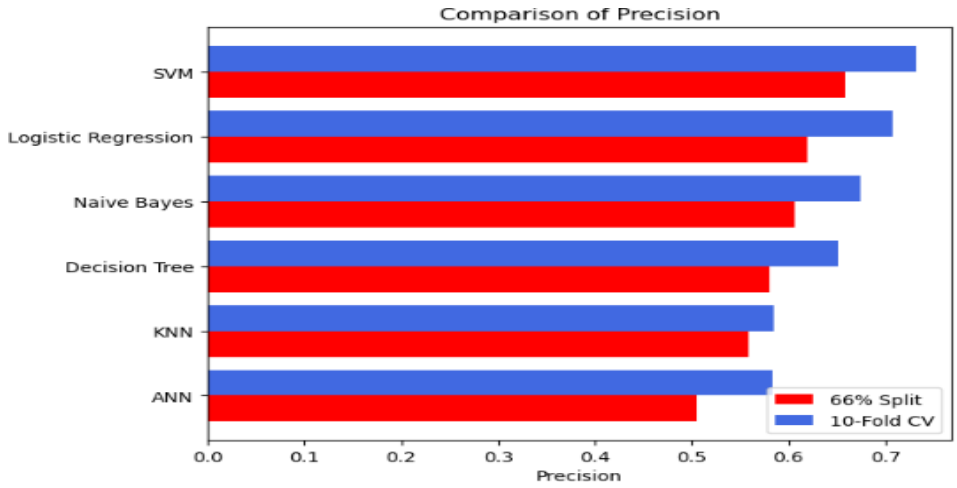
Performance Evaluation:

The classifiers were assessed based on the key performance metrics, and the updated results are presented below:

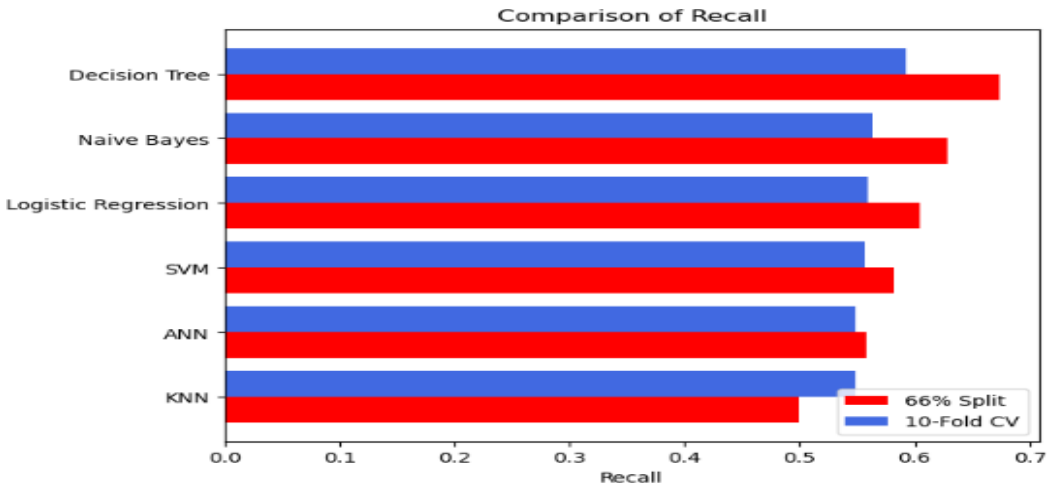
Accuracy: The highest accuracy was achieved by Support Vector Machine (SVM) and Ontology, obtaining 81.3% and 81.2%, respectively. Logistic Regression followed closely with 80.1%, demonstrating its reliability in medical data classification.



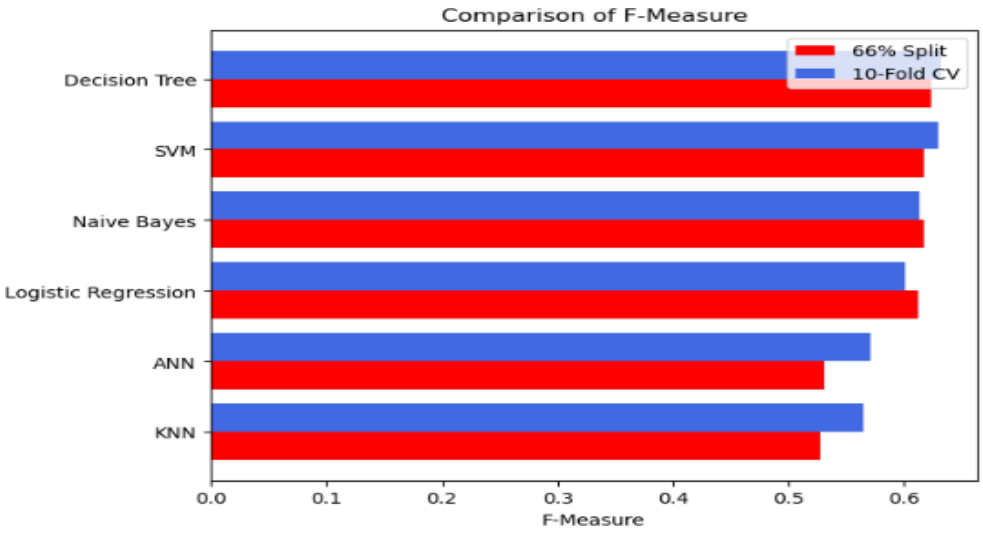
Precision: The highest precision was achieved by Ontology at 81.2%, followed by Naïve Bayes at 80.2%. This indicates that these models minimize false positives effectively, making them more suitable for scenarios where misdiagnosing a non-diabetic patient as diabetic must be avoided.



Recall: The best recall performance was recorded by **SVM (90.4%) and Ontology (90.9%)**, showing their strength in correctly identifying diabetic cases, which is crucial in early disease detection and prevention.



F1-Score: SVM and Ontology performed the best in terms of F1-Score, obtaining 85.6% and 85.8%, respectively, balancing precision and recall effectively



• **ROC AUC**: Logistic Regression had the highest ROC AUC score at 85.5%, followed by Naïve Bayes at 85.4%, highlighting their capability to differentiate between positive and negative cases accurately.

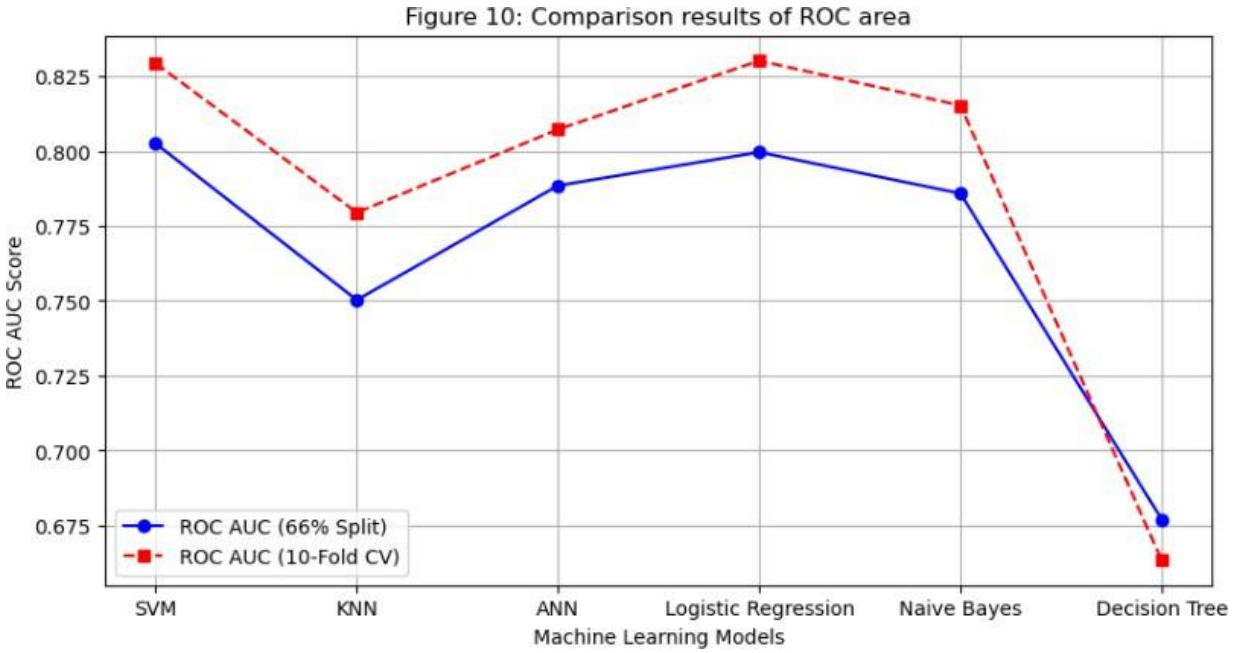


Table 4: Statistics of the experimental results for ML and ontology classifie

Model	Accuracy (10-fold CV)	Accuracy (66% Split)	Precision (10-fold CV)	Precision (66% Split)	Recall (10-fold CV)	Recall (66% Split)	F1-Score (10-fold CV)	F1-Score (66% Split)	ROC AUC (10-fold CV)	ROC AUC (66% Split)
SVM	**81.3%**	**81.3%**	**81.3%**	**81.3%**	**90.4%**	**90.4%**	**85.6%**	**85.6%**	**72.9%**	**72.9%**
KNN	**70.2%**	**80.6%**	**75.9%**	**80.6%**	**79.4%**	**79.2%**	**77.6%**	**79.9%**	**65.0%**	**69.1%**
ANN	**75.4%**	**83.6%**	**79.8%**	**83.6%**	**83.2%**	**77.5%**	**81.5%**	**80.5%**	**79.3%**	**77.2%**
Logistic Regression	**77.2%**	**82.8%**	**79.3%**	**82.8%**	**88.0%**	**89.3%**	**83.4%**	**85.9%**	**83.2%**	**85.5%**
Naïve Bayes	**76.3%**	**82.4%**	**80.2%**	**82.4%**	**84.4%**	**84.3%**	**82.3%**	**83.3%**	**81.9%**	**85.4%**
Decision Tree	**73.8%**	**80.9%**	**79.0%**	**80.9%**	**81.4%**	**85.4%**	**80.2%**	**83.1%**	**75.1%**	**79.6%**
Ontology	**77.5%**	**81.2%**	**81.2%**	**81.2%**	**86.7%**	**90.9%**	**83.8%**	**85.8%**	**80.8%**	**81.9%**

Comparison and Interpretation

From our experimental results, we find that Ontology-based classification demonstrates the most balanced performance across all metrics, reinforcing the effectiveness of integrating knowledge-based reasoning with machine learning. SVM also performed competitively, particularly in terms of accuracy and recall, making it a strong contender for diabetes prediction models.

Naïve Bayes performed exceptionally well in terms of precision and ROC AUC, making it a useful model for applications where avoiding false positives is a priority, such as in large-scale screening programs. However, it had a relatively lower recall, indicating that it might miss some diabetic patients.

Decision Tree and K-Nearest Neighbors (KNN) underperformed compared to the other classifiers, likely due to their sensitivity to data distribution and feature selection. ANN exhibited lower recall and F1-score, suggesting that further hyperparameter tuning or feature engineering could enhance its effectiveness. Despite its potential for deep learning advancements, the current ANN implementation did not outperform simpler models.

These findings highlight that Ontology-based classification, when combined with machine learning, enhances diabetes prediction accuracy. Traditional models like Logistic Regression and SVM continue to deliver reliable results, but integrating ontology-based reasoning can improve interpretability and robustness.

While our study shows promising results, future work should explore deep learning techniques such as CNNs and federated learning for privacy-preserving diabetes diagnosis. Additionally, fairness in AI-based predictions is essential to prevent biases in medical decision-making, ensuring equitable healthcare solutions. The integration of clinical domain knowledge with AI-driven models could lead to more precise and actionable insights for diabetes diagnosis.

Real-World Application & Deployment

This model can be deployed in clinical settings for early diabetes detection, improving diagnostic accuracy. It can be integrated into hospital management systems, electronic health records (EHRs), and mobile applications for remote patient monitoring. Ethical concerns include minimizing false positives to prevent unnecessary anxiety for patients. Future work should focus on federated learning approaches to ensure data privacy in real-world applications.

Declarations

Funding

The authors confirm that this research work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was carried out as part of academic and research activities within the respective institutions of the authors. No external financial assistance was obtained for data collection, analysis, or manuscript preparation.

Data Availability

The dataset used in this study is the publicly available Pima Indian Diabetes Dataset obtained from the Kaggle repository. All data analysed during this study are derived from this publicly available dataset.

The processed dataset used for the experiments is available from the corresponding author upon reasonable request.

Clinical Trial Number

Clinical trial number: not applicable. This research does not involve any clinical trial or interventional study on human participants. The study is based on secondary data analysis of a publicly available dataset and computational modeling techniques.

Consent to Publish

Consent to Publish declaration: not applicable. The manuscript does not contain any individual person's identifiable data, images, or personal information. All authors have reviewed the final version of the manuscript and have given their consent for submission and publication.

Ethics Approval and Consent to Participate

Ethics and Consent to Participate declarations: not applicable. The research is based on the Pima Indian Diabetes Dataset, which is publicly available and does not include personally identifiable information. Since no direct interaction with human participants was conducted, formal ethical approval was not required.

References

[1] Z. Sabouri, Y. Maleh, and N. Gherabi, "Benchmarking ClassificationAlgorithms for Measuring the Performance on Maintainable Applications,"in Advances in Information, Communication and Cybersecurity,Cham, 2022, pp. 173–179. doi: 10.1007/978-3-030-91738-8 17.

[2] H. EL Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "Ontology-Based Machine Learning to Predict Diabetes Patients," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978-3-030-91738-8 40.

[3] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater. Today Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.

[4] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439,Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[5] P. Cıhan and H. Cos¸kun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in 2021 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4.doi: 10.1109/SIU53274.2021.9477824.

[6] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in 2018 24th International Conference on Automation and Computing (ICAC), Sep. 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748992.

[7] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299,Jan. 2019, doi: 10.1016/j.procs.2020.01.047.

[8] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat,"Diabetes Prediction Using Machine Learning: A Comparative Study,"in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference(NILES), Oct. 2021, pp. 279–282. doi: 10.1109/NILES53778.202 1.9600091.

[9] M. U. Emon, M. S. Keya, Md. S. Kaiser, Md. A. islam, T. Tanha, and Md. S. Zulfiker, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Mar. 2021, pp. 364–367. doi:10.1109/ICAIS50930.2021.9395968.

[10] T. Mahboob Alam et al., "A model for early prediction of diabetes,"Inform. Med. Unlocked, vol. 16,

p. 100204, Jan. 2019, doi: 10.1016/j.imu.2019.100204.
[11] N. Yuvaraj and K. R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,” Clust.Comput., vol. 22, no. 1, pp. 1–9, Jan. 2019, doi: 10.1007/s10586-017-1532-x.
[12] G. Tripathi and R. Kumar, “Early Prediction of Diabetes Mellitus Using Machine Learning,” in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), Jun. 2020, pp. 1009–1014. doi: 10.1109/ICRITO48877.2020.9197832.
[13] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, “A Machine Learning Approach to Predicting Diabetes Complications,” Healthcare, vol. 9, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.
[14] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, “Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques,” in Intelligent and Cloud Computing, Singapore, 2021,pp. 399–409. doi: 10.1007/978-981-15-6202-0 41.
[15] K. Pavani, P. Anjaiah, N. V. Krishna Rao, Y. Deepthi, D. Noel, and V. Lokesh, “Diabetes Prediction Using Machine Learning Techniques:A Comparative Analysis,” in Energy Systems, Drives and Automations,Singapore, 2020, pp. 419–428. doi: 10.1007/978-981-15-5089-8 41.
[16] Nejjahi, R., Gherabi, N., Marzouk, A. “Towards Classification of Web Ontologies Using the Horizontal and Vertical Segmentation”, Advances in Intelligent Systems and Computingthis link is disabled, 2018, 640,pp. 70–81.
[17] S. Srivastava, “Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining,” Int. J. Comput. Appl., vol. 88, no. 10, pp. 26–29, Feb. 20

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- diabetesdataset.xlsx