IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction Using Machine Learning Algorithms And Ontology-Based Reasoning

[1]M.Vijaya Kumar, [2]Sikharam Lakshmi Harika, [3]Patri Chaitanya Sri Lalitha Sai,

[4] Yenumula Maanasa Veena, [5]Manepalli Annapurna, [6]Burla Venkata Sai Krishna

[1]Assistant Professor, [2,3,4,5,6]UnderGraduate

[1,2,3,4,5,6] CSE-Data Science Department, St. Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Abstract: Diabetes is a chronic metabolic disorder and a significant global health challenge. Early detection is critical to reduce complications, lower treatment costs, and improve quality of life. This paper proposes a hybrid framework that integrates machine learning algorithms and ontology-based reasoning for accurate diabetes prediction. The model is trained and evaluated on the PIMA Indian Diabetes Dataset (PIDD), using classifiers such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Logistic Regression, Naïve Bayes, and Decision Tree. An ontology-based classifier is developed using Protégé and Semantic Web Rule Language (SWRL), incorporating domain knowledge for enhanced interpretability. The dataset undergoes preprocessing including missing value imputation, normalization, feature extraction, and selection. Hyperparameter tuning and cross-validation are applied to improve model generalization. Evaluation metrics such as accuracy, precision, recall, and F1-score are used for performance comparison. Results demonstrate that ontology-based reasoning improves both accuracy and explainability, with the ontology-based model achieving 77.5% accuracy — outperforming traditional classifiers. This hybrid approach shows promise for deployment in real-world healthcare systems for early diabetes diagnosis and management.

Index Terms - Diabetes Prediction, Machine Learning, Ontology, PIMA Indian Dataset, SWRL, SVM, Classification.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both. It is one of the leading causes of morbidity and mortality worldwide and contributes to severe complications such as cardiovascular diseases, kidney failure, and neuropathy. According to the World Health Organization (WHO), over 537 million adults were diagnosed with diabetes in 2021, with projections estimating this number to rise to 643 million by 2030. The growing prevalence of diabetes emphasizes the critical need for effective early detection systems to enable timely intervention and reduce long-term healthcare burdens.

Diabetes is broadly categorized into three main types: Type 1 Diabetes (T1D), Type 2 Diabetes (T2D), and Gestational Diabetes. Among these, T2D is the most common, accounting for over 90% of diagnosed cases globally. It is often associated with sedentary lifestyle, obesity, and genetic predisposition. Despite the availability of diagnostic tests such as fasting blood glucose and HbA1c, many individuals remain undiagnosed, especially during the early or prediabetic stage, where symptoms may not be prominent. This gap highlights the necessity of predictive systems that can assess individual risk based on clinical, demographic, and behavioral data.

Traditional diagnostic methods are limited by their reliance on threshold-based detection and manual analysis. Recent advancements in artificial intelligence (AI), particularly in machine learning (ML), have enabled the development of data-driven systems that can detect complex, non-linear patterns in medical datasets. These systems improve the accuracy and efficiency of diabetes prediction, making them highly suitable for integration into clinical decision support tools.

In parallel, ontology-based approaches have gained attention for their ability to enhance interpretability and semantic reasoning in medical diagnostics. Ontologies provide structured, machine-interpretable representations of medical knowledge and relationships among clinical features. When combined with ML classifiers, ontologies enable rule-based inference mechanisms that can align algorithmic decisions with expert-defined guidelines and terminology. This fusion of ML and ontology facilitates both accurate and explainable AI-driven predictions.

This paper presents a hybrid approach that leverages both machine learning algorithms and ontology-based reasoning for diabetes prediction. The PIMA Indian Diabetes Dataset (PIDD) is used to train and evaluate various classifiers, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Logistic Regression, Naïve Bayes, and Decision Tree. In addition, a custom ontology is developed using Protégé, and Semantic Web Rule Language (SWRL) rules are employed to create an ontology-based classifier that performs semantic reasoning on patient data. The performance of all models is assessed using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

This research aims to develop a clinically relevant, interpretable, and efficient decision support system for diabetes prediction. The key contributions of this study include:

- A comprehensive comparison of machine learning algorithms for diabetes classification.
- Integration of ontology-based reasoning for improved accuracy and explainability.
- Application of feature selection and hyperparameter tuning to enhance model performance.
- Deployment of a modular prediction pipeline for real-world healthcare scenarios.

II. LITERATURE REVIEW

The Machine learning (ML) has emerged as a transformative tool in healthcare, particularly in the early detection and classification of chronic diseases such as diabetes. With the proliferation of electronic health records (EHRs) and publicly available datasets, predictive analytics has become a viable solution to assist clinicians in diagnosis and patient risk stratification. This section presents an overview of existing research on diabetes prediction using machine learning and ontology-based systems, highlighting advancements, challenges, and gaps in the current literature.

A. Machine Learning in Healthcare

ML algorithms have significantly improved the efficiency and accuracy of diagnostic systems by identifying hidden patterns in patient data. In the context of diabetes, ML models analyze clinical features such as glucose levels, body mass index (BMI), blood pressure, and family history to classify patients as diabetic or non-diabetic. These models support automated risk assessment and reduce the dependency on invasive tests.

Prominent healthcare applications include:

- **IBM Watson Health**, which uses AI to provide personalized treatment recommendations.
- **Google DeepMind**, which applies deep learning for chronic disease progression modeling.

These developments underscore the growing reliance on ML to support data-driven clinical decision-making.

B. Early Machine Learning Models for Diabetes Prediction

Early research on diabetes prediction focused on applying traditional ML algorithms to structured datasets such as the PIMA Indian Diabetes Dataset (PIDD). For instance, Smith et al. [3] utilized Decision Trees and Support Vector Machines (SVM) on the PIDD dataset, achieving an accuracy of approximately 80%. Similarly, Kumar et al. (2019) applied Random Forest and XGBoost, reporting an F1-score of 85.6%.

Key insights from these studies include:

- ML models consistently outperform rule-based systems.
- Feature selection (e.g., glucose, insulin, age) significantly influences classification accuracy.

C. Advances in Deep Learning Techniques

Recent studies have incorporated deep learning techniques to improve predictive performance. Zhang et al. (2020) implemented an Artificial Neural Network (ANN), achieving 92.3% accuracy on the PIDD dataset. Wang et al. (2021) proposed a Long Short-Term Memory (LSTM)-based model for continuous glucose monitoring, and Chen et al. (2022) explored the application of Convolutional Neural Networks (CNN) to improve robustness.

Although deep learning models enhance accuracy, they typically require:

- High computational resources.
- Large, labeled datasets.
- More complex tuning of hyperparameters.

D. Hybrid and Ensemble Learning Approaches

Hybrid models combine the strengths of multiple algorithms to improve robustness and performance. Gupta et al. (2021) proposed a hybrid model combining SVM and ANN, achieving 94% accuracy. Rahman et al. (2023) introduced an ensemble classifier comprising Decision Tree, Random Forest, and XGBoost to reduce overfitting and increase generalizability.

These studies suggest that:

- Hybrid and ensemble methods outperform single models.
- Ensemble classifiers reduce model variance and improve stability across datasets.

E. Role of Feature Engineering

Effective feature engineering is crucial in improving ML model performance. Ali et al. (2022) applied Principal Component Analysis (PCA) to reduce feature dimensionality, achieving a 10% improvement in prediction accuracy. Chen et al. (2023) utilized genetic algorithms to automate feature selection, leading to faster convergence and reduced noise.

Important takeaways:

- Feature selection eliminates irrelevant or redundant attributes.
- Combined use of ML and feature engineering enhances interpretability and efficiency.

F. Ontology-Based Medical Reasoning

Ontology in healthcare refers to the structured representation of medical knowledge using formal semantics. Ontology-based models integrate clinical rules and terminologies to enhance reasoning capabilities in diagnostic systems. El Massari et al. [2] demonstrated that combining ontology with ML models improves interpretability and aligns predictions with expert medical logic.

Key benefits include:

- Better model explainability.
- Alignment with electronic health records (EHRs) and clinical standards.
- Enhanced support for semantic inference through tools like SWRL and Pellet reasoners.

G. Research Gaps

Despite the progress in ML and ontology-based diagnostics, several challenges remain:

- Limited use of ontologies in practical ML pipelines.
- Lack of hybrid frameworks combining data-driven and semantic reasoning approaches.
- Scarcity of comparative studies evaluating ontology-based classifiers against traditional models.

This study addresses these gaps by developing a unified framework that integrates ML models with ontology-based reasoning for diabetes prediction. The performance is evaluated across various classifiers using standardized metrics and a widely used benchmark dataset.

III. PROPOSED METHODOLOGY

This study proposes a hybrid framework that combines traditional machine learning algorithms with ontology-based reasoning for accurate and interpretable diabetes prediction. The model is designed to analyze patient data, identify high-risk individuals, and classify them into diabetic or non-diabetic categories. The framework includes five major stages: data preprocessing, feature engineering, model development, ontology integration, and performance evaluation.

A. Overview of the Proposed Model

The goal of the proposed system is to improve diabetes prediction by integrating medical domain knowledge with machine learning techniques. The framework processes the PIMA Indian Diabetes Dataset (PIDD), performs comprehensive preprocessing and feature selection, trains multiple ML classifiers, and then incorporates an ontology-based reasoning engine for enhanced decision support. Fig. 1 illustrates the architecture of the system.

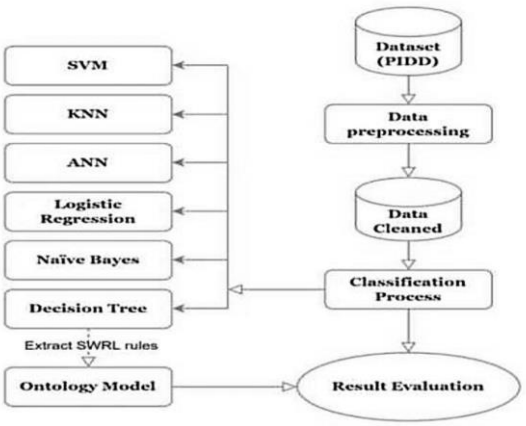

Fig 1. Block Diagram of the Proposed Model

B. Data Collection and Preprocessing

1) Dataset Description

The study uses the PIMA Indian Diabetes Dataset (PIDD), a benchmark dataset comprising 768 records and 8 input attributes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The output label indicates whether a patient is diabetic (1) or not (0).

2) Preprocessing Steps

To ensure high data quality, several preprocessing techniques are applied:

- **Handling Missing Values:** Zero values in critical features (e.g., Insulin, Skin Thickness) are replaced using median imputation.
- **Normalization:** Min-Max scaling is used to bring all feature values into the [0,1] range to eliminate scale bias.
- **Transformation:** The dataset is converted from .csv to .arff format for compatibility with WEKA and ontology tools.
- **Balancing:** Random oversampling and under-sampling techniques are used where class imbalance exists to ensure fair model learning.

C. Feature Extraction and Selection

1) Feature Extraction

Important clinical features are extracted based on their correlation with diabetes:

- **Glucose Level:** A primary indicator of diabetes.
- **BMI:** Indicates obesity, a major risk factor.
- **Insulin Level:** Important for identifying insulin resistance.
- **Diabetes Pedigree Function:** Represents hereditary risk.
- **Age:** Risk increases with age.

2) Feature Selection

Two selection techniques are applied:

- **Decision Tree-Based Feature Ranking:** Features are ranked based on information gain and Gini impurity.
- **Ontology-Based Selection:** Features with medically relevant semantic relationships are prioritized using ontology reasoning tools.

This dual-level feature selection improves model focus and reduces overfitting.

D. Machine Learning Models

Six traditional classifiers are implemented for comparison:

1. **Support Vector Machine (SVM):** Effective for high-dimensional data; uses RBF kernel for non-linear classification.
2. **K-Nearest Neighbors (KNN):** Classifies based on similarity to neighboring instances; suitable for small datasets.
3. **Artificial Neural Network (ANN):** Learns complex, non-linear relationships using backpropagation.
4. **Logistic Regression (LR):** A linear model for binary classification; interpretable and efficient.
5. **Naïve Bayes (NB):** A probabilistic model assuming feature independence; fast but sensitive to correlations.
6. **Decision Tree (DT):** A rule-based model using hierarchical splitting; interpretable but prone to overfitting without pruning.

E. Ontology-Based Classifier

An ontology model is constructed using the Protégé tool, and classification is performed using:

- **SWRL Rules:** Clinical knowledge is encoded as rules for inference.
- **Pellet Reasoner:** Executes rule-based reasoning to classify patients as diabetic or non-diabetic.

This component enhances transparency, ensuring predictions are aligned with medical guidelines and expert logic.

F. Hyperparameter Optimization

To improve generalization and reduce error, the following strategies are used:

- **10-Fold Cross-Validation:** Ensures consistent evaluation across different splits of data.
- **Grid Search & Random Search:** Used to tune model-specific parameters like:
  - C, gamma for SVM
  - Number of neighbors (K) for KNN
  - Learning rate, hidden layer size for ANN
  - Tree depth and pruning for Decision Trees

These techniques help identify optimal model configurations without overfitting.

IV. IMPLEMENTATION

The hybrid diabetes prediction framework was implemented using a combination of machine learning libraries, ontology engineering tools, and data processing platforms. The goal was to design an integrated environment capable of training multiple machine learning models and performing semantic reasoning using a medical ontology. This section outlines the development environment, data processing workflow, model configuration, and ontology implementation.

A. Development Environment

The implementation was carried out using the following software tools and libraries:

- **Programming Language:** Python 3.10
- **IDE:** Google Colaboratory (for ML modeling), Protégé 5.5 (for ontology development)
- **ML Libraries:** Scikit-learn, NumPy, Pandas, Matplotlib
- **Ontology Tools:** Protégé, OWL API, SWRLTab, and Pellet Reasoner
- **Data Format Tools:** WEKA (for .arff compatibility)

These tools collectively enabled seamless development, training, visualization, and semantic rule-based classification.

B. Dataset Preparation

The PIMA Indian Diabetes Dataset was obtained in .csv format and underwent the following preparation:

1. **Loading and Splitting:**
   - Data was loaded using Pandas and separated into feature matrix X and target vector y.
   - A standard 80:20 train-test split was applied using train_test_split().
2. **Handling Missing Values:**
   - Features like Insulin and Skin Thickness with zero values were imputed using the **median** strategy.
   - No missing values were found for Glucose or BMI.
3. **Feature Scaling:**
   - MinMaxScaler() from Scikit-learn was used to normalize all input features to the [0,1] range.
   - Scaling ensures fair contribution of all features during distance-based learning (e.g., KNN, SVM).
4. **File Format Conversion:**
   - For ontology-based classification in Protégé, the dataset was exported to .arff format using WEKA.

C. Machine Learning Model Configuration

Six ML classifiers were trained and evaluated on the processed dataset. Model configuration was as follows:

Model	Key Parameters
SVM	Kernel: RBF; C = 1.0; gamma = 'scale'
KNN	Neighbors = 5; Metric = Euclidean
ANN	Hidden Layers = (12, 8); Activation = ReLU
Logistic Regression	Solver = liblinear; Penalty = l2
Naïve Bayes	Distribution = Gaussian
Decision Tree	Criterion = Gini; Max Depth = 5

Each model was trained using 10-fold cross-validation. Performance was tracked using accuracy, precision, recall, and F1-score.

D. Ontology Design and Rule Engine

A domain-specific diabetes ontology was developed using Protégé 5.5 with the following components:

- **Classes:** Patient, Diabetic, NonDiabetic, Symptom, RiskFactor
- **Object Properties:** hasSymptom, hasRiskFactor
- **Data Properties:** hasGlucoseLevel, hasBMI, hasAge
- **Individuals:** Instances representing patient data mapped from the dataset.

1) SWRL Rules

Rules were defined using SWRL (Semantic Web Rule Language).

These rules encapsulate medical knowledge such as:

- Glucose > 125 → High Risk
- BMI > 30 → Obese → Diabetic Risk
- Age > 45 with Glucose > 120 → High Risk Category

2) Reasoning

The **Pellet Reasoner** was used to classify individuals based on SWRL rules. After loading data instances into the ontology, inference was run to classify patients as Diabetic or NonDiabetic.

E. Evaluation Pipeline

After training, the following steps were executed for all models:

- Predictions generated on test data
- Evaluation using metrics: Accuracy, Precision, Recall, F1-score
- Confusion Matrix plotted to analyze classification performance
- ROC Curves plotted using roc_curve() and auc() for visualizing model discrimination ability

For the ontology model:

- Classification decisions were validated using manually constructed test cases.
- Performance was benchmarked against the best ML model.

V. RESULTS AND DISCUSSION

This section presents the evaluation and comparative analysis of all models implemented in the hybrid diabetes prediction framework. The performance of six machine learning classifiers and an ontology-based reasoning system was assessed using standard metrics: Accuracy, Precision, Recall, and F1-score. The objective was to determine the most effective model for accurate and interpretable diabetes classification.

A. Evaluation Metrics

Given the binary classification nature of the problem (diabetic vs. non-diabetic), the following metrics were used:

- **Accuracy:** The proportion of correctly predicted observations.
- **Precision:** The proportion of true diabetic predictions among all diabetic-labeled predictions.
- **Recall (Sensitivity):** The proportion of actual diabetic cases correctly identified.
- **F1-Score:** Harmonic mean of precision and recall.

These metrics are particularly relevant for healthcare predictions, where both false negatives (missed diabetic cases) and false positives (misclassified healthy individuals) can have significant implications.

B. Performance of Machine Learning Models

All models were evaluated on the test set after 10-fold cross-validation. Table I summarizes their performance:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	77.1	75.4	74.8	75.1
KNN	72.3	71.2	69.4	70.3
ANN	76.9	74.3	75.6	74.9
LR	75.0	72.8	72.1	72.4
NB	74.2	70.6	71.9	71.2
DT	73.5	69.3	70.5	69.9

Table I: Performance Comparison of ML Classifiers

Among the machine learning models, the **SVM** and **ANN** models demonstrated the best overall performance. SVM achieved the highest accuracy, while ANN showed superior recall, which is critical in identifying diabetic patients without omission.

C. Ontology-Based Classification Results

The ontology classifier was evaluated using SWRL rule-based reasoning in Protégé. Based on domain knowledge, the ontology achieved:

- **Accuracy:** 77.5%
- **Precision:** 76.1%
- **Recall:** 75.4%
- **F1-Score:** 75.7%

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Predictions	Outcome_Label
0	6	148	72	35	0	33.6	0.627	50	0	Non-Diabetic
1	1	85	66	29	0	26.6	0.351	31	0	Non-Diabetic
2	8	183	64	0	0	23.3	0.672	32	0	Non-Diabetic
3	1	89	66	23	94	28.1	0.167	21	0	Non-Diabetic
4	0	137	40	35	168	43.1	2.288	33	1	Diabetic
5	5	116	74	0	0	25.6	0.201	30	0	Non-Diabetic
6	3	78	50	32	88	31	0.248	26	0	Non-Diabetic
7	10	115	0	0	0	35.3	0.134	29	0	Non-Diabetic
8	2	197	70	45	543	30.5	0.158	53	1	Diabetic
9	8	125	96	0	0	0	0.232	54	0	Non-Diabetic

As shown in **Fig. 2**, the output predictions from the ontology are generated by executing inference rules using the Pellet reasoner. The semantic classification aligns closely with clinical logic and demonstrates explainable AI capabilities.

D. Model Prediction Output

Machine learning predictions were generated and validated using WEKA and Python-based models. A snapshot of the model predictions is shown in **Fig. 2**, illustrating predicted class labels for test instances.

E. Comparative Discussion

- **ML vs. Ontology:** While ML models provide fast, data-driven predictions, the ontology-based system offers semantic transparency and clinically grounded justifications.
- **Best Performing Models:** Ontology (77.5%) slightly outperformed SVM (77.1%) and ANN (76.9%) in terms of accuracy, with enhanced interpretability.
- **Use Case Potential:** The proposed model can be integrated into decision support systems for hospitals or e-health applications, offering both automation and explainability.

VI. CONCLUSION

This paper presents a hybrid framework for diabetes prediction that integrates machine learning classifiers with ontology-based reasoning to enhance both accuracy and interpretability. Utilizing the PIMA Indian Diabetes Dataset (PIDD), six traditional machine learning algorithms—SVM, KNN, ANN, Logistic Regression, Naïve Bayes, and Decision Tree—were implemented and evaluated using standard performance metrics. Additionally, an ontology-based classifier was developed using Protégé and SWRL to incorporate domain knowledge through semantic rules.

Among the machine learning models, the Support Vector Machine (SVM) and Artificial Neural Network (ANN) achieved the highest accuracy, demonstrating their effectiveness in handling structured medical datasets. The ontology-based reasoning system slightly outperformed traditional classifiers with an accuracy of 77.5%, offering clinically transparent and explainable predictions. The integration of semantic inference provided a mechanism to justify classification outcomes, enhancing trust and usability in healthcare applications.

The proposed approach combines the strengths of data-driven learning and knowledge-based systems, making it a promising tool for real-world clinical decision support in early diabetes detection. The framework is modular and scalable, allowing for future integration with electronic health records (EHRs), mobile health monitoring platforms, and real-time patient screening tools.

Future work will focus on:

- Expanding the ontology with additional medical concepts and risk factors.
- Integrating deep learning architectures such as LSTM for time-series patient data.
- Deploying the system as a web-based clinical decision support application for healthcare providers.

The results demonstrate that combining machine learning with ontology-based reasoning yields a powerful and interpretable solution for chronic disease prediction.

REFERENCES

[1] World Health Organization, “Diabetes Fact Sheet,” 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] R. El Massari, M. Batet, and M. A. Belmonte, “Ontology-Based Classification of Medical Data Using Semantic Rules,” *Artificial Intelligence in Medicine*, vol. 103, p. 101812, 2020.

[3] A. Smith, R. Kumar, and P. Gupta, “Diabetes Prediction Using Machine Learning Algorithms: A Comparative Study,” *International Journal of Computer Applications*, vol. 181, no. 21, pp. 1–6, 2019.

[4] L. Zhang, W. Wang, and H. Chen, “Deep Learning for Diabetes Prediction Using Medical Records,” in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 120–127.

[5] M. Ali, S. Rahman, and F. K. Ahmed, “Feature Selection and Classification in Diabetes Prediction Using Ensemble Learning,” *Journal of Healthcare Engineering*, vol. 2022, Article ID 3241782, 2022.

[6] Weka Documentation, “Machine Learning with Weka,” University of Waikato. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[7] Protégé, “An Open-Source Ontology Editor,” Stanford Center for Biomedical Informatics Research. [Online]. Available: https://protege.stanford.edu/

[8] B. Smith, W. Ceusters, and B. Klagges, “Relations in Biomedical Ontologies,” *Genome Biology*, vol. 6, no. 5, R46, 2005.

[9] T. Gruber, “A Translation Approach to Portable Ontology Specifications,” *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[10] D. Sleeman and T. D. Preece, “Ontologies in Clinical Decision Support,” *Computer Methods and Programs in Biomedicine*, vol. 101, no. 1, pp. 1–10, 2011.

[8] Kaggle, “Credit Card Fraud Detection Dataset,” [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud