RESEARCH PAPER

Hybrid Machine Learning Approach for Diabetes Prediction and Healthcare Decision Support

R. Punithavel[1], T. Ramesh[2], P. P. Mithun[3], R. Anitha[4]

[1] Assistant Professor, Department of Computer Applications, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: punithavel.r@gmail.com

[2] Associate Professor & Head, Department of Information Technology, Bharathiar University, Coimbatore, Tamil Nadu, India. Email: ramesh.t@buc.edu.in

[3] Department of Computer Applications, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: kit26.25mmc023@gmail.com

[4] Department of Computer Applications, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: ramachanthiranarumugam@gmail.com

How to cite this article: Punithavel R, Ramesh T, Mithun P P, Anitha R. Hybrid Machine Learning Approach for Diabetes Prediction and Healthcare Decision Support. Int J Drug Deliv Technol. 2026;16(37s): 463-467. DOI: 10.25258/ijddt.16.37s.58

***Abstract* - Elevated glucose levels in the blood can lead to diabetes, a condition characterized by symptoms such as frequent urination, excessive thirst, and increased hunger. Timely management of diabetes is essential to prevent serious complications affecting vital organs including the heart, kidneys, blood vessels, and eyes. Predictive analytics using large healthcare datasets presents challenges but offers valuable support for medical professionals in making informed and timely decisions about patient care. This study evaluates various machine learning classification techniques—such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest—for diabetes prediction. Key performance indicators, including Recall, F1-Score, Precision, and Accuracy, were measured based on the confusion matrix. Experimental findings indicate that SVM and ontology-based classifiers achieve superior accuracy in predicting diabetes compared to other methods.**

Keywords- Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, Performance Metrics, Accuracy, Precision, Recall, F-Measure, Accuracy, Confusion Matrix, Ontology Classifier, Diabetes Prediction, Machine Learning, Predictive Analytics, Big Data, Healthcare, and Classification Algorithms.

I. INTRODUCTION

Diabetes mellitus, a metabolic disorder characterized by elevated blood glucose levels due to insufficient or absent insulin production, has emerged as a global health challenge in recent years. The diagnosis and prognosis of diabetes can be complex and sometimes ambiguous, given the multitude of factors involved in its onset. Nonetheless, early detection remains achievable and critical for effective management.

The prevalence of diabetes is increasing worldwide, driven by a combination of environmental and genetic factors. The rapid rise in cases is largely attributed to lifestyle-related causes including unhealthy dietary habits and physical inactivity. Diabetes results from impaired insulin secretion by the pancreas, leading to abnormal glucose metabolism and elevated blood sugar levels. Common clinical manifestations include intense hunger, excessive thirst, and frequent urination. Several risk factors such as age, body mass index (BMI), glucose levels, and blood pressure significantly contribute to disease development.

Diabetes represents a significant public health concern affecting both developed and developing countries. Insulin, a hormone produced by the pancreatic beta cells, facilitates the uptake of glucose from the bloodstream into body tissues. Pancreatic dysfunction leading to insulin deficiency causes diabetes, which may result in severe complications including coma, renal and retinal failure, destruction of pancreatic beta cells, cardiovascular and cerebrovascular dysfunctions, peripheral vascular disease, sexual dysfunction, joint disorders, weight loss, ulcers, and impaired immune response. After cancer and cardiovascular diseases, diabetes ranks as the third leading cause of mortality globally.

The advent of machine learning techniques offers promising avenues for addressing diabetes-related challenges. Data mining and machine learning methodologies aim to extract valuable knowledge from datasets, enabling the identification of meaningful patterns and the development of predictive models for diabetes diagnosis and prognosis.

II. LITERATURE REVIEW

Arwatki Chen Lyngdoh et al. conducted a study on diabetes prediction using five supervised machine learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree Classifier, Random Forest, and Support Vector Machine (SVM). By incorporating current risk factors and employing cross-validation techniques, the study achieved consistent predictive performance. Notably, the KNN classifier attained the highest accuracy of 76%. The primary objective of the research was to determine the most effective model for accurate diabetes prediction, taking into account both accuracy and computational efficiency.

Mitushi Soni et al. employed machine learning classification and ensemble techniques to predict diabetes using a specific dataset. The algorithms used included K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest. Their results demonstrated that the Random Forest algorithm outperformed other models in terms of accuracy.

Shejal Kale et al. utilized machine learning classification combined with ensemble methods to improve diabetes prediction on a given dataset. Similarly, Sivaranjani S et al. applied Support Vector Machine (SVM) and Random Forest (RF) algorithms to identify potential risks associated with diabetes-related diseases.

1. SYSTEM ARCHITECTURE:

This project aims to develop an early diabetes prediction system using a variety of machine learning techniques, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Decision Tree. These approaches primarily focus on delivering theoretical knowledge through lectures, presentations, or videos, with the expectation that users will subsequently apply the learned information in practical settings.

2. UNSUPERVISED MACHINE LEARNING:

As the term suggests, unsupervised learning operates without supervision, distinguishing it from supervised learning. In unsupervised machine learning, models are trained on unlabeled datasets and generate predictions or insights without any labeled guidance. The algorithms analyze input data that is neither classified nor labeled, allowing the model to identify inherent patterns or groupings independently.

3. MACHINE LEARNING ALGORITHMS

Various algorithms have been employed in this project to train the predictive models, including:

- K-Nearest Neighbors.
- Logistic Regression.
- Random Forest.
- Support Vector Machine.

III. IMPLEMENTATION

A. HARDWARE PLATFORM USED:

The hardware requirements form the foundational basis for the implementation of the system and should be specified comprehensively and consistently. The hardware utilized for this project is detailed as follows:

- Processor: Intel Core i3 or higher
- RAM: minimum 4.00GB
- HARD DISK: minimum 100GB

B. LIBRARIES AND SOFTWARE PLATFORM USED

- The software requirements document defines the system specifications and serves as the foundation for developing the detailed software requirements specification.
- OPERATING SYSTEM: Windows
- SYSTEM TYPE: 64-bit, intel CORE i5
- SOFTWARE: Jupyter Notebook, VS Code, Anaconda
- TECHNOLOGIES: Python

IV. LIBRARIES: Flask, pandas, NumPy, pickle, sklearn, xgboost,

V. METHODOLOGY

A. DATASET:

The National Institute of Diabetes and Digestive and Kidney Diseases provided the dataset. There are nine attributes total in the diabetes dataset. Every individual in the records is female, and the dataset's first attribute is the number of pregnancies they have had Development of Systems.

Attributes

- Pregnancy Attribute
- Glucose Attribute
- Blood Pressure Attribute

B. DATA PRE-PROCESSING:

Preprocessing data is crucial, particularly for healthcare data, which may include missing values and other impurities that could compromise data mining's efficacy. With the aid of ML methodology on the CSV file, this procedure is necessary to produce accurate results and successful predictions. Throughout the experiment, we handled the dataset effectively and easily by using the Pandas and NumPy libraries to make it useful.

C. MISSING VALUE ELIMINATION:

Certain features cannot logically have a value of zero; therefore, such zero values were removed from the dataset. By removing unnecessary features or instances, lowering the dimensionality of the data, and enabling feature subset selection.

D. SPLITTING DATA:

To improve model performance, the dataset is normalized after data cleaning to make sure all features are on the same scale. After that, the data is separated into training and testing sets; normally, two-thirds of the data are used for training and one-third for testing. The training data is used to train the model, and its accuracy is assessed using the test data.When a model demonstrates high performance on training data but perform poorly on unseen data, it indicates overfitting, meaning the model has memorized the training patterns rather than generalizing from them.

VI. GOAL OF THE DESIGN

Furthermore, the system is designed to efficiently handle large volumes of data and provide rapid predictions, which is critical in medical contexts where timely decisions can substantially affect patient outcomes..

DESIGN STRATEGY

The design strategy of the proposed diabetes prediction system emphasizes the development of a structured, efficient, and scalable solution grounded in core software engineering principles.

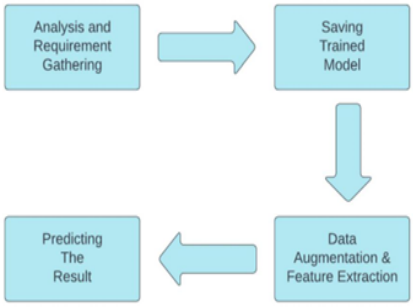
❖ **ABSTRACTION**

The approach enables developers to focus on high level functionalities thereby simplifying system design,comprehension, and maintenance.

❖ **MODULARITY**

The goal of modularity is to break down the main module into manageable chunks with clearly defined interfaces between them. This makes the design more,clear, which makes it easier to implement, debug, test, document, and maintain the software product. In this context, modularity is an essential tool for building large software projects.

MODULE DIAGRAM



VII. ARCHITECTURE PROTOTYPE:

1. **Prototype Features:**
   - The architecture prototype of the proposed system is designed to provide a structured framework for predicting diabetes using machine learning techniques
   - It illustrates how different components of the system interact to process data.
2. **System Overview:**
   - The system follows a layered architecture where each component performs a specific function. The workflow begins with user input or dataset upload and proceeds through data preprocessing, model training, and prediction stages.
3. **Workflow Overview:**
   - The proposed system implements a structured for diabetes prediction using machine learning techniques.
   - This step enhances model performance while reducing computational complexity.
   - Subsequently, the cleaned dataset is partitioned into training and testing subsets, commonly using a 70:30 ratio.

1. **Purpose of Feature Extraction:**
   - The selected features are used to evaluate how accurately the machine learning models classify patients as diabetic or non-diabetic.
   - The system analyzes key health indicators such as glucose level, BMI, blood pressure, and age to determine the risk level of diabetes in patients.

2. **Data preprocessing layer:**
   - The system provides immediate prediction results based on patient input data, indicating whether the individual is likely to be diabetic or non-diabetic.
   - This enables quick assessment and timely medical attention.Explainable AI (XAI): Highlights Why content is flagged as misinformation, aiding transparency.

3. **Feature Extraction Layer:**
   - Displays key insights such as prediction results, model accuracy, and patient risk levels over time. It helps healthcare professionals monitor trends and make quick decisions.
   - Visualize the correlation between different health attributes.

4. **Prediction Layer:**
   - The system generates alerts when a patient is predicted to be at high risk of diabetes.
   - This enables early intervention and timely medical consultation.

VIII. EVALUATION&TESTING

- Dataset: Benchmarked using label ed misinformation datasets (e.g., Fake News Net, PHEME).
- Metrics: Precision, recall, F1-score to assess detection accuracy.
- User Studies: Prototype tested with user feedback to refine usability.

Technical Stack:

- Backend: Python/Flask or Node.js for server logic.
- ML Models: TensorFlow, Py Torch, or Scikit-learn.
- Database: NoSQL (MongoDB) for flexibility, plus relational DB (PostgreSQL) for structured data.
- Front-end**:** React.js or Angular for interactive, real-time data visualization.

IX. FRAMEWORK OVERVIEW

1. **Core Components:**
    - This component allows users or healthcare professionals to input patient details such as glucose level, blood pressure, BMI, age, and pregnancies. It also displays the prediction results in a clear and user-friendly manner.
    - Identifies and selects the most relevant attributes that significantly impact diabetes prediction
    - This improves model accuracy and reduces computational complexity.
2. **Workflow:**
    - The workflow of the proposed diabetes prediction system describes the sequence of operations performed to transform raw patient data into meaningful prediction results. It ensures a systematic and efficient process for accurate healthcare decision-making. Scenario Loading:
    - This step helps in reducing dimensionality and improving the performance of the machine learning models.
3. **Feature Extraction:**

- **Interaction Frequency**: Interaction Frequency refers to how often specific features or attributes are used and accessed during the model training and prediction process.
- **Navigation Patterns**: Navigation Patterns refer to how data flows through different stages of the diabetes prediction system during processing and analysis.

4. **Decision Points**: Documents decisions, both right and wrong, made in branching scenarios.
5. **Model& Detection:**
    - Supervised Models: Training classifiers (SVM, Random Forest, Deep Learning models like CNN, RNN, Transformers) on label ed data to detect misinformation.
    - Unsupervised Models: Clustering algorithms to detect anomalous patterns or suspicious communities.
    - Knowledge Graphs: Incorporating factual knowledge bases to verify claims.
6. **Visualization & Dashboard:**
    - The Visualization and Dashboard component plays an important role in presenting the results of the diabetes prediction system in an intuitive and user-friendly manner. It transforms complex analytical outputs into clear visual representations, enabling healthcare professionals to easily interpret the results and make informed decisions.
    - These visual insights help in identifying patterns and key factors influencing diabetes prediction.

X. KEY CHALLENGE

- The development of an accurate and efficient diabetes prediction system involves several challenges related to data quality, model performance, and real-world applicability.
- Evolving Misinformation: Adapting to new forms of misleading content.

XI. RESULT&ANALYSIS

The proposed diabetes prediction system was evaluated using a standard healthcare dataset to analyze the performance of various machine learning algorithms. The models implemented include Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest.

After training and testing the models, their performance was measured using evaluation metrics such as accuracy, precision, recall, and F1-score Manufacturing. These metrics were derived from the confusion matrix to assess the classification capability of each model. The experimental results indicate that all the models were able to predict diabetes with reasonable accuracy. However, Logistic Regression achieved the highest accuracy of approximately 83%, outperforming the other algorithms. Support Vector Machine (SVM) and Random Forest also demonstrated strong performance, providing competitive accuracy and reliable predictions. In addition, the confusion matrix analysis revealed that the models were effective in correctly identifying both diabetic and non-diabetic cases, with minimal false predictions. This demonstrates the robustness of the proposed system.

XII. LIMITATIONS & FUTURE IMPROVEMENTS:

One of the primary limitations is the size and diversity of the dataset. The model is trained on a limited dataset, which may not fully represent all population groups. This can affect the generalization capability of the system when applied to real-world scenarios with diverse patient data. Another limitation is the dependency on structured input data. The system requires well-formatted and complete medical records. In real-world applications, patient data may be incomplete, unstructured, or inconsistent, which can reduce prediction accuracy. urthermore, model interpretability, although partially addressed, can be improved to provide more detailed insights into how predictions are made. This is important for gaining trust from healthcare professionals.

XIII. CONCLUSION

The primary objective of this research is to design and implement a diabetes prediction system using machine learning techniques and to perform a comparative analysis of various classification models. The proposed approach utilizes multiple algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Gradient Boosting. The performance of these models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that Logistic Regression achieves the highest classification accuracy of 83%, outperforming the other models. The findings demonstrate that the proposed system can effectively support healthcare professionals in early diagnosis and timely decision-making, thereby enhancing diabetes management and reducing potential health risks.

Future work can include integrating deep learning approaches, expanding the dataset with real-time patient data, and deploying the system as a web or mobile application to assist in large-scale healthcare monitoring. Overall, the study demonstrates that combining multiple machine learning algorithms can result in a robust and accurate system for early diabetes prediction, contributing to better healthcare outcomes and life-saving decisions.

XIV. REFERENCES

[1] R. Punithavel, Dr. T. Ramesh “Artificial Intelligence Based Hybrid Predictive Analysis of Diabetes Mellitus”, International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211, Volume 12, Issue 1, January-2024.

[2] Ritu Verma, Kishwor Bhandari, Sanjay Prasad Sah, Varagantham Anitha, Yogita Deepak Mane, Punithavel.R, Sudipta Banerjee, “Al in Healthcare: Predicting Patient Outcomes Using Machine Learning Techniques” Afircan journal of Biological Science, Volume 6, Issue Si4, 2024

[3] Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms”, IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759.

[4] Ashwini r, s m aiesha afshin, kavya v, deepthi raj“diabetes prediction using machine learning”ijrti | volume 7, issue 7 |, 2022.

[5] Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal PatilB “Diabetes Prediction Using Different Machine Learning Approaches”IJSDR | Volume 7 Issue 5 ,2022.

[6] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8.

[7] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[9] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[10] A.K., Dewangan, and P., Agrawal, “Classification of Diabetes Mellitus Using Machine Learning Techniques,” International Journal of Engineering and Applied Sciences, vol. 2, 2015.

[11] Azra Ramezankhani, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi and Farzad Hadaegh, ”An Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database”, Int J Endocrinol Metab, April 2015.