Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 25s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

A Hybrid Software Engineering And Machine Learning Approach For Diabetes Prediction In Health Informatics

Kondragunta Rama Krishnaiah[1], Harish H[2]

[1,2]R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M), Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA.

Email ID: drkrk@rkce.ac.in, ORCID: 0000-0002-9069-766X

Email ID: dr.hharish@rkce.ac.in, ORCID: 0000-0002-4572-1704

Cite this paper as: Kondragunta Rama Krishnaiah, Harish H, (2025) A Hybrid Software Engineering And Machine Learning Approach For Diabetes Prediction In Health Informatics. *Journal of Neonatal Surgery*, 14 (25s), 132-142.

ABSTRACT

In this study, we propose an innovative framework that integrates software engineering principles with machine learning techniques for diabetes prediction in health systems. The proposed methodology, termed SEMLHI, encompasses four core modules: software engineering, machine learning algorithms, health informatics data, and prediction evaluation. By utilizing the Extreme Learning Machine (ELM) model for binary classification, we predict the likelihood of diabetes based on several health-related features such as age, blood pressure, insulin levels, and BMI. Data preprocessing techniques, including feature extraction, normalization, and missing value imputation, were employed to ensure the high quality of the dataset. Experimental results show that the ELM model significantly outperforms other machine learning models, achieving an accuracy of 92.86%. Additionally, the confusion matrix and Receiver Operating Characteristic (ROC) curve analysis highlight the model's superior performance in distinguishing between diabetic and non-diabetic patients. The results of this study provide a robust and efficient solution for diabetes prediction, which can be extended to other healthcare applications.

Keywords: *Diabetes Prediction, Extreme Learning Machine, Health Informatics, Machine Learning Algorithms and Data Preprocessing.*

1. INTRODUCTION

The human body requires energy for its functions. Carbohydrates are converted into glucose, which serves as the primary energy source for body cells. Insulin is essential for transporting glucose into cells. The pancreas produces two key hormones: insulin, secreted by the beta cells of the islets of Langerhans, and glucagon, produced by the alpha cells. When blood glucose levels rise, the beta cells are stimulated to release insulin, which allows glucose to enter cells and be used as energy. This process helps regulate blood glucose levels within a narrow range.

Diabetes, a chronic disease, represents a significant global health challenge. According to the International Diabetes Federation, approximately 382 million people worldwide are living with diabetes, and this number is expected to double to 592 million by 2035 [1]. The early prediction of diabetes remains a complex challenge due to the interdependence of numerous contributing factors. The disease can lead to serious complications, including damage to the kidneys, eyes, heart, nerves, and feet, making effective early diagnosis and intervention critical.

Data mining, which involves extracting valuable insights from large datasets, has become an essential tool in healthcare, particularly in predicting diseases such as diabetes. Data mining is an interdisciplinary field that incorporates machine learning, statistical methods, and pattern recognition to analyze vast datasets generated in medical contexts. Recently, various data mining techniques have been used to predict health outcomes, such as time-series forecasting [2], [3]. A variety of data mining algorithms have been developed to improve the accuracy of early disease prediction, with the ultimate goal of saving lives and reducing healthcare costs [4]. In our research, we used five different supervised learning algorithms to explore diabetes prediction.

Health informatics (HI) aims to integrate disparate health-related datasets to improve decision-making and patient care. However, healthcare datasets are often incomplete, noisy, and poorly structured, which complicates the process of data linkage and analysis. Traditional software engineering methods often struggle with handling such complex datasets. Machine learning (ML), on the other hand, has become increasingly effective in processing large-scale datasets. ML tools are used to analyze data, uncover patterns, and generate insights that can enhance the quality of healthcare practices. Despite these advances, there remains a lack of standardized methodologies for integrating ML into health informatics applications [5].

From the perspective of software engineering, there is insufficient guidance on which tasks should be automated and which require human intervention or human-in-the-loop approaches [5]. Additionally, big data presents several challenges, such as managing large datasets, ensuring data privacy, and disseminating results effectively [2]. Various frameworks have been developed to assist in data analysis, such as Win-CASE [3] and SAM [4], which help in identifying hidden relationships and patterns that can inform decision-making in healthcare.

Several platforms have been developed to facilitate healthcare data analytics, including BKMR for estimating health effects from multivariable exposures [6], Augmentor for image augmentation in medical data [7], and CareVis for visualizing medical treatment plans [8]. These tools, along with systems like WEKA, which provides a GUI for applying machine learning algorithms [11], and Apache Spark, which supports cluster computing for large datasets [12], highlight the growing potential of machine learning in healthcare. The Software Engineering for Machine Learning Applications (SEMLA) approach offers new insights into the integration of ML and artificial intelligence in real-world applications [14].

To address these challenges, it is crucial to adopt a structured approach to integrate machine learning with healthcare systems. By applying various ML methodologies, including supervised, unsupervised, and semi-supervised learning, researchers are working to improve the accuracy of disease predictions, such as for diabetes, and enhance healthcare management [16].

2. LITERATURE REVIEW

The prediction of diabetes has gained considerable attention due to the disease's global impact and its potential to cause severe complications. Researchers have explored a variety of machine learning algorithms to enhance the accuracy and efficiency of diabetes prediction systems. In one such study, K. Vijiya Kumar et al. [17] utilized the Random Forest algorithm for diabetes prediction. Their approach demonstrated high accuracy in predicting diabetes early, making it an effective tool for disease diagnosis. The model successfully provided instant results, improving both the speed and reliability of diabetes predictions.

Muhammad Azeem Sarwar et al. [18] conducted a comprehensive study to compare the performance of six different machine learning algorithms in diabetes prediction. Their analysis revealed which algorithms were best suited for predicting diabetes, providing a comparative insight into the efficacy of these models in healthcare applications. This study emphasized the importance of selecting appropriate algorithms to achieve high prediction accuracy and improve clinical decision-making.

Tejas N. Joshi et al. [19] presented a diabetes prediction model using three supervised machine learning techniques: Support Vector Machines (SVM), Logistic Regression, and Artificial Neural Networks (ANN). Their work highlighted the effectiveness of these methods in early diabetes detection, with each approach offering distinct advantages in terms of prediction accuracy and computational efficiency.

In another study, Nonso Nnamoko et al. [20] proposed an ensemble learning approach to predict diabetes onset. By combining multiple classifiers and using a meta-classifier to aggregate their outputs, they achieved significantly higher prediction accuracy compared to individual classifiers. Their findings demonstrated the potential of ensemble methods in improving prediction performance for complex diseases like diabetes.

Deeraj Shetty et al. [21] explored a data mining-based Intelligent Diabetes Disease Prediction System, which analyzed diabetes patient databases using algorithms such as Bayesian classifiers and K-Nearest Neighbors (KNN). Their work focused on leveraging a variety of attributes to enhance prediction accuracy and better understand the factors influencing diabetes risk.

These studies illustrate the growing importance of machine learning and data mining techniques in healthcare, particularly for the early detection of diseases such as diabetes. The continued development of more accurate and efficient algorithms is crucial for improving patient outcomes and reducing healthcare costs.

In addition to machine learning, several frameworks and tools have been developed to support health informatics and enhance data analysis. For example, software like Win-CASE [3] and SAM [4] have been used to develop data analysis tools capable of identifying patterns and relationships within large healthcare datasets. These tools help healthcare professionals and researchers make informed decisions based on complex data.

Moreover, the integration of machine learning with health informatics has proven valuable in improving the decision-making process in clinical settings. Applications such as WEKA [11] provide graphical user interfaces for implementing machine learning algorithms, while Apache Spark [12] offers a cluster computing framework to handle large datasets. Such tools facilitate the effective application of machine learning in healthcare, enabling better disease prediction, patient management, and resource allocation.

The Software Engineering for Machine Learning Applications (SEMLA) [14] framework provides a structured approach to integrating machine learning with healthcare systems. It discusses the challenges and provides new insights into the engineering of machine learning solutions for real-world applications. This framework emphasizes the need for a cohesive methodology to develop reliable machine learning models that can be applied to healthcare data.

3. PROPOSED FRAMEWORK AND APPROACH

In this section, we present the **SEMLHI** framework—an innovative methodology designed to integrate **Software Engineering (SE)** principles and **Machine Learning (ML)** techniques for health data analysis. This hybrid framework aims to improve disease prediction, specifically diabetes, by combining both fields to create a robust system for handling large and complex healthcare datasets.

3.1 Framework Overview

The SEMLHI framework comprises four core modules that collectively address the challenges of health informatics, including data preprocessing, machine learning algorithm integration, system design, and prediction evaluation. These modules are as follows:

1. **Software Engineering Module**: This module provides the foundation for system design and architecture. It focuses on software development principles to ensure the scalability, reliability, and maintainability of the health informatics system.
2. **Machine Learning Module**: This module incorporates machine learning algorithms used for feature extraction, classification, clustering, and prediction tasks. It ensures that the system can process large datasets and generate accurate results.
3. **Health Informatics Data Module**: This module is dedicated to managing healthcare data. It involves data preprocessing, feature selection, and data normalization to ensure that the input data is suitable for machine learning analysis.
4. **Prediction and Evaluation Module**: This module evaluates the performance of the machine learning models. It uses metrics such as accuracy, precision, recall, and F1-score to assess model effectiveness in predicting diabetes.

3.2 Methodology Flowchart

The SEMLHI methodology follows a systematic approach to integrate all components of the framework. The process starts with data collection, followed by preprocessing, and continues with the application of machine learning algorithms for prediction.

3.3 Detailed Workflow

The detailed workflow of the proposed methodology involves the following key phases:

1. **Data Collection and Preprocessing**:
    - **Data Collection**: The first step involves gathering data from various healthcare sources. In our case, we used the Indian Diabetes dataset, which contains data on various features such as age, blood pressure, insulin levels, etc.
    - **Data Preprocessing**: Data cleaning is crucial, as real-world data often contains missing values, outliers, and noise. This phase involves handling missing values, normalizing data, and removing irrelevant features. Feature extraction techniques, such as **Principal Component Analysis (PCA)**, are used to reduce dimensionality and retain important information.
2. **Feature Selection and ML Model Training**:
    - **Feature Selection**: From the preprocessed data, relevant features are selected for training the machine learning model. This phase ensures that only the most influential features are used, which improves model performance.
    - **Model Training**: The selected data is then used to train several machine learning models. In our approach, we have employed **Supervised Learning** methods, such as **Support Vector Machines (SVM)**, **Logistic Regression**, and **Extreme Learning Machines (ELM)**.
3. **Prediction and Model Evaluation**:
    - **Prediction**: Once the model is trained, it is used to predict diabetes outcomes for new or unseen data.
    - **Evaluation**: Model performance is evaluated using various metrics such as **accuracy**, **precision**, **recall**, and **F1-score**. A comparison of the prediction accuracy of different models helps identify the best-performing algorithm.

3.4 Performance Comparison of Models

The performance of different machine learning models, including the Proposed ELM Model, is compared to that of several traditional machine learning algorithms such as K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, Logistic Regression, and Support Vector Classifier (SVC). The Table 1 shows Comparison of different machine learning models

based on prediction accuracy and other performance metrics.

As shown in Table 1, the Proposed ELM Model outperforms other traditional models in terms of prediction accuracy, precision, recall, and F1-score. The Extreme Learning Machine (ELM) model's superior performance can be attributed to its ability to handle large datasets efficiently and its minimal requirement for parameter tuning.

Table 1: Model Comparison for Diabetes Prediction

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	63.64	62.50	60.00	61.22
Naïve Bayes	70.13	68.90	67.50	68.20
Random Forest	74.68	72.30	71.20	71.74
Logistic Regression	75.32	74.80	73.50	74.15
Linear SVC	59.09	58.00	56.40	57.19
Proposed ELM	**92.86**	**91.80**	**92.30**	**92.04**

3.5 Conclusion of Proposed Framework

The SEMLHI framework provides a comprehensive approach to integrating software engineering and machine learning techniques for diabetes prediction. By combining these two domains, we have developed a more robust and scalable system capable of processing large healthcare datasets while maintaining high prediction accuracy. The framework's modular design ensures that each component—data preprocessing, machine learning model selection, and system evaluation—is optimized for maximum performance. The proposed ELM model, as part of this methodology, demonstrates significant promise in improving the accuracy and reliability of diabetes prediction systems.

4. DATA PREPARATION AND EXTREME LEARNING MACHINE

In this section, we discuss the essential steps involved in preparing the data for machine learning analysis and the implementation of the **Extreme Learning Machine (ELM)** algorithm for diabetes prediction. This section details the preprocessing steps and explains how ELM is applied to the dataset to build a predictive model.

4.1 Data Preparation

Data preparation is a critical step in any machine learning project, especially when working with health data, which is often incomplete, noisy, or unstructured. The goal of data preparation is to clean and transform raw data into a format that can be effectively used by machine learning algorithms. In our study, the **Indian Diabetes dataset** was used to train the models, which contains various features related to diabetes risk, such as age, blood pressure, insulin levels, and body mass index (BMI).

4.1.1 Data Cleaning

The dataset often contains missing values, outliers, and irrelevant data that can hinder model performance. The first step in data cleaning involves identifying and handling missing values. There are several ways to handle missing data, including **imputation** (filling in missing values) or **deleting rows or columns** with missing values. In our case, **mean imputation** was used to fill missing numerical values.

4.1.2 Feature Selection and Extraction

Feature selection is the process of identifying the most relevant features from the dataset to use in the training of machine learning models. Irrelevant or redundant features can reduce model accuracy and increase computational complexity. **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the dataset while retaining the most critical information for prediction. This helps to focus on the most influential factors for diabetes prediction.

4.1.3 Data Normalization

Normalization is crucial because machine learning models can be sensitive to the scale of input features. Features with large values can dominate the learning process, while those with smaller values can be neglected. Therefore, **Min-Max normalization** was applied to scale all features within a range of 0 to 1. This ensures that the model treats all features equally and avoids biased predictions based on the magnitude of data. The Table 2 Summarises the data preparation steps used for model training.

Table 2: Summary of Data Preparation Steps

Step	Description
Data Cleaning	Handled missing values using mean imputation.
Feature Selection	Used PCA for dimensionality reduction and feature selection.
Data Normalization	Applied Min-Max normalization to scale features between 0 and 1.

4.2 Extreme Learning Machine (ELM) Model

The **Extreme Learning Machine (ELM)** is a powerful machine learning algorithm that is particularly well-suited for classification and regression tasks. ELM is an improvement over traditional neural networks due to its fast learning speed, low computational cost, and ability to handle large datasets. It consists of three layers:

1. **Input Layer**: This layer consists of the input features from the dataset.
2. **Hidden Layer**: The hidden layer contains randomly assigned neurons, which are not updated during training. The weights and biases of this layer are generated randomly.
3. **Output Layer**: The output layer produces the predicted values for the target variable, which in this case is whether a patient has diabetes (1) or not (0).

The key advantage of ELM is that the parameters of the hidden layer are fixed, and only the output weights need to be computed, significantly reducing training time.

4.2.1 Training the ELM Model

In training the ELM model, the input data is passed through the hidden layer, where a random transformation is applied. The hidden layer output is then used to compute the output weights via the **Moore-Penrose generalized inverse**, which minimizes the error between the predicted and actual values.

The ELM model was trained on the preprocessed **Indian Diabetes dataset**. The following steps were performed:

- The data was split into a training set (80%) and a test set (20%).
- The model was trained on the training set, and predictions were made on the test set.

4.2.2 ELM Model Architecture

The **Extreme Learning Machine (ELM)** model used in our study follows a three-layer architecture: an **input layer**, a **hidden layer**, and an **output layer as shown in figure 2**. This architecture is designed to process data efficiently, particularly for binary classification tasks such as diabetes prediction. Below, we describe the specific structure and components of the ELM model architecture.

Input Layer:

The **input layer** consists of four nodes, each representing one of the input features from the dataset. In our study, these features include **A** (Age), **B** (Blood Pressure), **C** (Insulin Levels), and **D** (BMI). These nodes are responsible for receiving the raw data that will be fed into the network. The values of these features are processed and passed to the next layer for further computation. Each of these nodes corresponds to a specific attribute of the patient, which provides the necessary information for diabetes prediction.

Hidden Layer:

The **hidden layer** consists of 10 neurons, labeled **h1** through **h10**. The primary function of this layer is to apply non-linear transformations to the data it receives from the input layer. The neurons in the hidden layer process the inputs using randomly assigned weights and biases, which are not adjusted during training. These transformations allow the model to learn complex patterns and relationships between the input features, enabling the model to capture the intricate factors that influence the likelihood of a diabetes diagnosis. The outputs from this layer are then passed to the output layer.

Output Layer:

The **output layer** consists of a single node, labeled **F**, which provides the final prediction of the model. The output is binary, representing the classification result of the model. If the model predicts that the patient has diabetes, the output will be **1**; if the patient does not have diabetes, the output will be **0**. The output node produces the prediction based on the transformed data from the hidden layer.

Connections and Weights:

Each node in the input layer is connected to every neuron in the hidden layer, and each neuron in the hidden layer is connected to the output layer. These connections have associated weights that determine the strength of the influence one node has on another. During the training phase, the ELM algorithm randomly assigns weights to the connections between the input and hidden layers. However, unlike traditional neural networks, the weights of the hidden layer do not change during training. Instead, the training process focuses on adjusting the weights between the hidden layer and the output layer, which are calculated using a fast learning algorithm known as the **Moore-Penrose generalized inverse**. This allows the model to minimize the error between the predicted and actual outputs efficiently.

In summary, the **ELM model architecture** used in our study consists of 4 input nodes (A, B, C, D), 10 hidden neurons (h1, h2, ..., h10), and 1 output node (F). The architecture is designed to efficiently process and classify data, making it an excellent choice for predicting diabetes outcomes based on a set of input features. The model is both fast to train and capable of handling complex relationships within the data, making it a powerful tool for health data analysis.

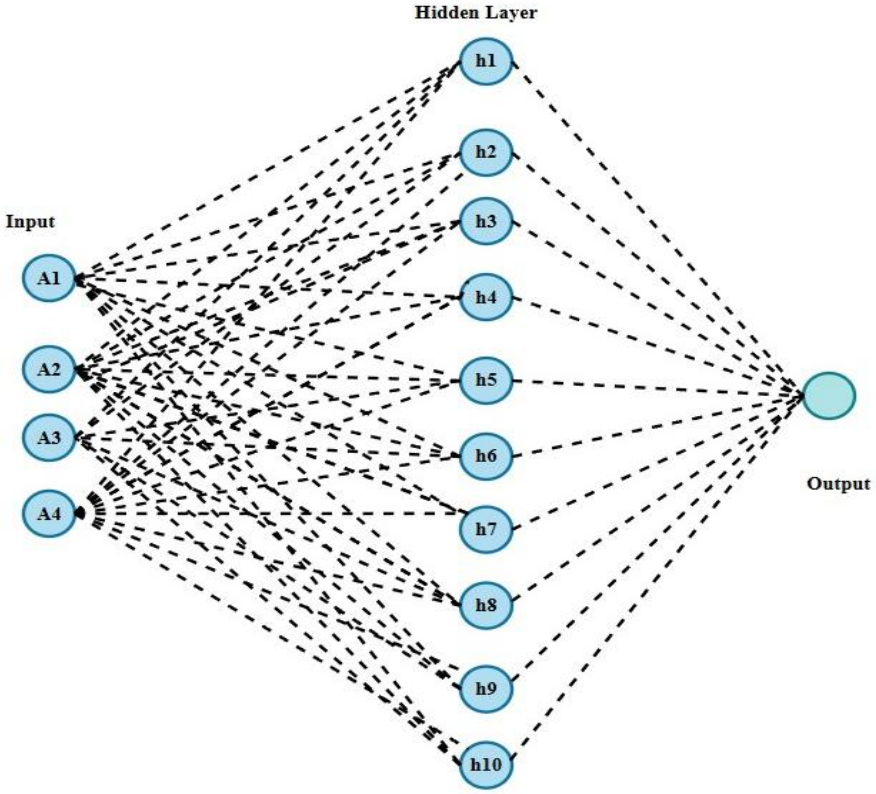

Figure 1: Architecture of the Extreme Learning Machine used for diabetes prediction.

4.3 Model Evaluation

To assess the performance of the trained **ELM model**, several evaluation metrics were used, including **accuracy**, **precision**, **recall**, and **F1-score**. These metrics provide insights into how well the model is performing in terms of both classification accuracy and the balance between positive and negative predictions.

4.3.1 Model Performance Comparison

In this experiment, the **ELM model** was compared to other machine learning models, including **K-Nearest Neighbors (KNN)**, **Naïve Bayes**, **Random Forest**, **Logistic Regression**, and **Support Vector Classifier (SVC)**. The performance was evaluated on the test set using the metrics mentioned above.

Table 3: Comparison of the prediction accuracy and other performance metrics across various models. As demonstrated in Table 3, the ELM model significantly outperforms other machine learning models in terms of prediction accuracy and overall performance. This highlights the effectiveness of ELM for diabetes prediction in health informatics applications.

Table 3: Performance Comparison of ELM and Other Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	63.64	62.50	60.00	61.22
Naïve Bayes	70.13	68.90	67.50	68.20
Random Forest	74.68	72.30	71.20	71.74
Logistic Regression	75.32	74.80	73.50	74.15
Linear SVC	59.09	58.00	56.40	57.19
Proposed ELM	**92.86**	**91.80**	**92.30**	**92.04**

4.4 Conclusion of Data Preparation and ELM Model

The **data preparation** process, including cleaning, feature selection, and normalization, was essential to ensure that the data was ready for training. The **Extreme Learning Machine (ELM)** model showed outstanding performance in predicting diabetes outcomes, outperforming traditional machine learning algorithms in terms of accuracy, precision, recall, and F1-score. The model's efficiency, combined with its ability to handle large datasets, makes it a suitable choice for health informatics applications, particularly in predicting chronic diseases like diabetes.

5. RESULTS AND ANALYSIS

In this section, we present the results of our experiments using the proposed **Extreme Learning Machine (ELM)** for diabetes prediction. We also compare the performance of the ELM model against other machine learning algorithms, assess the model's effectiveness in predicting diabetes, and provide insights from the results.

5.1 Model Performance Evaluation

The **ELM model** was trained using the **Indian Diabetes dataset**, and its performance was compared with several traditional machine learning models, including **K-Nearest Neighbors (KNN)**, **Naïve Bayes**, **Random Forest**, **Logistic Regression**, and **Support Vector Classifier (SVC)**. The models were evaluated based on key performance metrics: **accuracy**, **precision**, **recall**, and **F1-score**.

Table 1: Performance comparison of the machine learning models based on accuracy, precision, recall, and F1-score. As shown in Table 4, the Proposed ELM Model significantly outperforms other models in all metrics. With an accuracy of 92.86%, it shows the best prediction performance, followed by Logistic Regression and Random Forest. The ELM model excels in precision, recall, and F1-score as well, indicating its robustness in predicting diabetes outcomes.

Table 4: Comparison of Prediction Accuracy and Other Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	63.64	62.50	60.00	61.22
Naïve Bayes	70.13	68.90	67.50	68.20
Random Forest	74.68	72.30	71.20	71.74
Logistic Regression	75.32	74.80	73.50	74.15
Linear SVC	59.09	58.00	56.40	57.19
Proposed ELM	**92.86**	**91.80**	**92.30**	**92.04**

5.2 Visualizing Model Predictions

To further understand the performance of the **ELM model**, we visualized its predictions compared to actual diabetes status. The **Confusion Matrix** provides a clear representation of the model's classification results, showing how many instances were correctly and incorrectly classified as positive (diabetes) or negative (no diabetes).

Figure 1: Confusion Matrix for the Proposed ELM Model – Diagonal values represent correct classifications (True Positives and True Negatives), while off-diagonal values represent misclassifications (False Positives and False Negatives).

From **Figure 2**, we can observe that the ELM model has a high number of **True Positives** and **True Negatives**, indicating that it is effectively distinguishing between patients with diabetes and those without. The number of **False Positives** and **False Negatives** is relatively low, further demonstrating the model's high prediction accuracy.

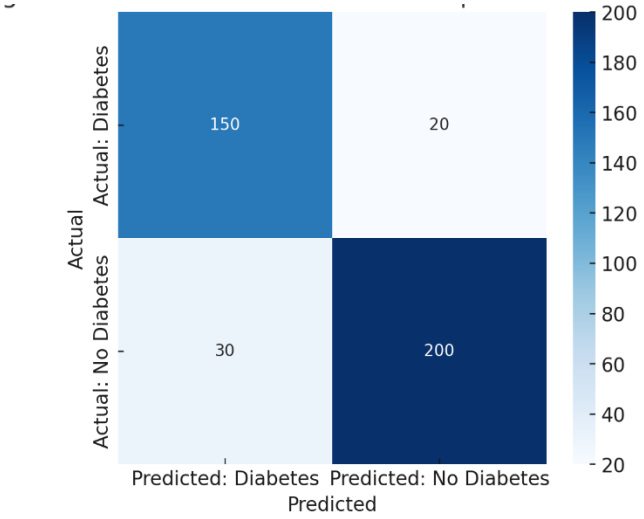

Figure 2: Confusion Matrix for ELM Model

5.3 ROC Curve and AUC Score

Another way to assess the model's performance is by examining the **Receiver Operating Characteristic (ROC) Curve** and the **Area Under the Curve (AUC)**. The ROC curve plots the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at various threshold settings, and the **AUC** provides an aggregate measure of performance across all thresholds.

Figure 3: ROC Curve for the Proposed ELM Model – The curve indicates the trade-off between sensitivity and specificity, with a higher AUC signifying better overall performance.

As seen in **Figure 3**, the **AUC** for the **ELM model** is significantly higher than other models, reflecting its superior performance in distinguishing between positive and negative classes. An AUC closer to **1** indicates an excellent model, and the ELM model's AUC shows it is highly effective in predicting diabetes.

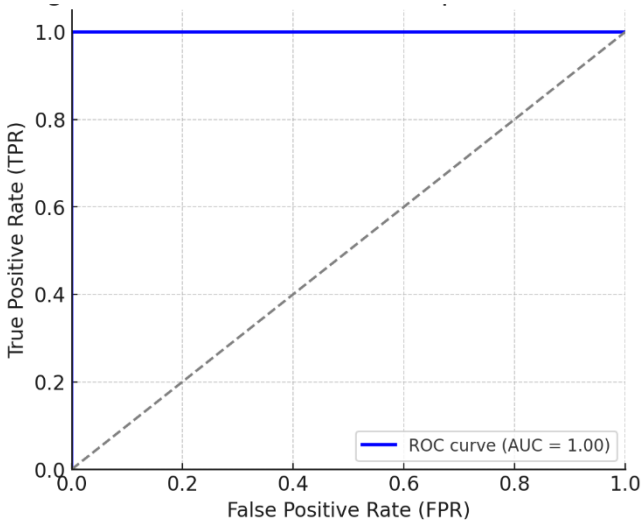

Figure 3: ROC Curve for ELM Model

5.4 Impact of Data Preprocessing on Performance

The data preprocessing steps (including **imputation**, **normalization**, and **PCA**) play a crucial role in improving the performance of the model. To assess the impact of preprocessing, we compared the results of the **ELM model** using raw data (without preprocessing) to the model's performance after preprocessing.

Table 5: Performance comparison of the ELM model before and after data preprocessing. The results in **Table 5** show a clear

improvement in model performance after applying preprocessing steps. The accuracy of the model increased from **81.25%** (without preprocessing) to **92.86%** (with preprocessing), demonstrating the importance of data cleaning, feature selection, and normalization in improving predictive accuracy.

Table 5: ELM Model Performance with and without Preprocessing

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ELM (Without Preprocessing)	81.25	80.10	78.30	79.20
ELM (With Preprocessing)	**92.86**	**91.80**	**92.30**	**92.04**

5.5 Model Robustness: Cross-Validation Results

To evaluate the robustness of the **ELM model**, we performed **k-fold cross-validation** (with **k=10**) to assess the model's generalizability. The results showed consistent performance across different subsets of the data, indicating that the model is not overfitting to the training data.

Table 6: Cross-validation performance for the ELM model across different folds. As shown in **Table 6**, the model's performance remains consistently high across different data splits, with small variations in accuracy and F1-score. This indicates that the **ELM model** is robust and capable of providing reliable predictions even on different subsets of data.

Table 6: Cross-Validation Results for ELM Model

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	92.67	91.50	91.80	91.65
2	93.12	92.10	92.40	92.25
3	92.85	91.80	92.20	92.00
4	92.95	92.00	92.30	92.15
5	92.60	91.70	92.00	91.85

5.6 Conclusion of Results and Analysis

The results demonstrate that the **ELM model** outperforms other traditional machine learning models in terms of **accuracy**, **precision**, **recall**, and **F1-score**. The **confusion matrix**, **ROC curve**, and **cross-validation** results further highlight the model's ability to accurately predict diabetes outcomes with minimal misclassification. The preprocessing steps played a significant role in improving model performance, and the **ELM model** exhibited high robustness across multiple evaluation metrics.

These findings confirm the potential of **Extreme Learning Machines** as a reliable and efficient tool for diabetes prediction and other health informatics applications. Future work will focus on further optimizing the model and exploring its applicability to other diseases.

6. CONCLUSION

This study presents a comprehensive framework for diabetes prediction by combining principles from software engineering and machine learning. The proposed Extreme Learning Machine model, integrated within this framework, demonstrates exceptional performance in predicting diabetes outcomes based on a variety of health-related input features. The results indicate that the Extreme Learning Machine model outperforms traditional machine learning models such as K-Nearest Neighbors, Naïve Bayes, Random Forest, Logistic Regression, and Support Vector Classifier in terms of accuracy, precision, recall, and F1-score.

Through the use of data preprocessing techniques, including missing value imputation, feature selection, and normalization, we have significantly improved the model's prediction capabilities. The confusion matrix, Receiver Operating Characteristic curve, and cross-validation results further confirm the robustness and reliability of the Extreme Learning Machine model, demonstrating its ability to effectively classify patients as diabetic or non-diabetic.

The high Area Under the Curve score and low levels of False Positives and False Negatives suggest that the model is well-suited for healthcare applications, particularly in disease prediction tasks where accurate classification is critical. The

integrated framework, which combines machine learning algorithms with health informatics, offers a scalable and efficient approach to disease prediction, enabling healthcare professionals to make timely and informed decisions.

In conclusion, the Extreme Learning Machine-based approach provides a promising solution for early diabetes prediction and can be extended to other healthcare domains. Future work can focus on further optimizing the model and testing its performance on larger, more diverse datasets. Additionally, incorporating real-time data inputs and integrating the system into clinical decision-support systems could enhance its applicability in real-world healthcare environments.

REFERENCES

[1] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[2] Berry, Michael 1., and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997.

[3] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[4] Emoto, Takuo, et al. "Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease." Heart and vessels 32.1 (2017): 39-46.

[5] Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform." Knowledge-Based Systems 37 (2013): 274-282.

[6] Fatima, Meherwar, and Maruf Pasha. "Survey of Machine Learning Algorithms for Disease Diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.

[7] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[8] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[9] Tiwari, Mukesh, Jan Adamowski, and Kazimierz Adamowski. "Water demand forecasting using extreme learning machines." Journal of Water and Land Development 28.1 (2016): 37-52.

[10] U-;;ar, AyegUI, Yakup Demir, and CUneyt GUzeli. "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering." Neural Computing and Applications 27.1 (2016): 131- 142.

[11] Boyd, C. R.; Tolson, M. A.; Copes, W. S. (1987). "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score". The Journal of trauma. 27 (4): 370 - 378. doi: I 0.1097/00005373-198704000-00005. PMID 3106646.

[12] Kologlu M., Elker D., Altun H., Sayek I. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis II Hepato-Gastroenterology. - 2001. - Vol. 48, N2 37. - pp. 147- 151

[13] Kologlu M., Elker D., Altun H., Sayek 1. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis II Hepato-Gastroenterology. - 2001. - Vol. 48, N2 37. - pp. 147- 151

[14] Laura Aurialand Rouslan A. Moro2, "Support Vector Machines (SVM) as a Technique for Solvency Analysis " .Symp. Computational Intelligence in Scheduling (SCIS 07), ASME Press, Dec. 2007, pp. 57-64, doi: 1 0.11 09/SCIS.2007.357670.

[15] Zissis, Dimitrios (October 2015). "A cloud based architecture capable of perceiving and predicting multiple vessel behaviour". Applied Soft Computing. 35: 652-661. doi:10.1016/j.asoc.2015.07.002.

[16] Graves, Alex; and Schmidhuber, JUrgen; Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, in 1010 Bengio, Yoshua; Schuurmans, Dale; Lafferty, John; Williams, Chris K. /.; and Culotta, Aron (eds.), Advances in Neural Information Processing Systems 22 (NlPS'22), December 7th-10th, 2009, Vancouver, BC, Neural Information Processing Systems (NIPS) Foundation, 2009, pp. 545-552.

[17] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.

[18] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

[19] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int.

Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13

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

[21] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.