**Journal of Science and Technology – Smart Systems and Devices**
*Volume 36, Issue 3, September 2026, 000–000*

# Comparative Analysis of Machine Learning Models for Diabetes Readmission Classification

***Hang T. Dang[1], Quyen T. N. Vo[1], Quang Tran Ngoc[1], Dinh Do Van[2,*]***
*[1] Le Quy Don Technical University, Ha Noi, Vietnam*
*[2]Sao Do University, Hai Phong, Viet Nam*
*[*] Corresponding author email: dinh.dv@saodo.edu.vn*

**Abstract**

*This study presents a comparative analysis of various machine learning and deep learning models to predict the 30-day readmission risk for diabetic patients using Electronic Health Record (EHR) data. Utilizing the Diabetes 130-US Hospitals dataset, which comprises 101,766 records and 50 clinical features, the research aims to enhance classification accuracy for a highly unbalanced medical dataset. The proposed technical pipeline incorporates comprehensive preprocessing steps, specifically employing SMOTEENN for data balancing, Z-score normalization, and Principal Component Analysis (PCA) to reduce dimensionality while retaining 80% of the total variance (26 principal components). Seven classification models were benchmarked: Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest (RF), Support Vector Machine, Multilayer Perceptron, and a OptimizedDeep Multilayer Perceptron. Experimental results demonstrate that the Random Forest model significantly outperforms others, achieving an accuracy of 86.90%, a peak F1-score of 89.94%, and a remarkably high recall of 92.67%. The Random Forest (RF) model achieved a Recall of 92.67% on the balanced training set. However, when evaluated on the imbalanced held-out test set to simulate real-world performance, it maintained a Recall of 38% and an Average Precision (AP) of 0.1361, significantly surpassing the calculated random baseline of 0.0954. This represents a substantial improvement of 78.37% in recall compared to existing baseline studies. Furthermore, the optimized configuration for clinical EHR within the Deep MLP framework showed a 2.77% improvement in recall over standard MLP models, highlighting its effectiveness in capturing complex clinical correlations. These findings suggest that the optimized Random Forest model possesses high potential for integration into early warning systems. By identifying high-risk patients promptly, healthcare providers can implement timely interventions, thereby reducing readmission rates and optimizing medical resources.*

Keywords: Diabetes readmission, healthcare prediction, random forest.



## 1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by persistently elevated blood glucose levels resulting from insufficient insulin production by the pancreas or the body's ineffective use of insulin. This condition leads to the accumulation of sugar in the bloodstream, causing long-term damage to blood vessels, nerves, eyes, kidneys, and the cardiovascular system. Diagnosis typically involves analyzing medical history and performing clinical tests such as Fasting Plasma Glucose (FPG) and Hemoglobin A1c (HbA1c). Currently, artificial intelligence and data science are revolutionizing healthcare through the analysis of Electronic Health Records (EHR). A critical application is predicting 30-day hospital readmission, which is a key factor influencing high medical costsand patient mortality rates, especially when glycemic control is suboptimal.

Previous studies using the Diabetes 130-US Hospitals (UCI) dataset have achieved significant milestones but maintain certain technical limitations. Strack et al. (2014) primarily utilized multivariate primarily utilized multivariate Logistic Regression models, focusing more on statistical interpretation than on optimizing predictive performance [1]. Liu et al. (2024) achieved a high F1- score (0.83) but reported a relatively low Area Under the Receiver Operating Characteristic (AUROC) curve (~0.64) [2]. More recently, Emi-Johnson et al. (2025) reported AUC-ROC values ranging from 0.58 to 0.67; however, the actual predictive performance for the positive class (readmission <30 days) remained low, with a sensitivity (Recall) of only 0.143 for Deep Neural Networks (DNN) [3].

This research proposes a comprehensive technical pipeline involving SMOTEENN for data balancing, Z-score normalization, and Principal Component Analysis(PCA) for dimensionality reduction (selecting 26 components to retain 80% variance). We benchmark seven models ranging from traditional algorithms to a proposed Optimized Deep MLP for tabular EHR. The core contribution of this study lies in maintaining sensitive features (e.g., race, age) while optimizing classification performance for high-risk readmission. Based on 101,766 clinical records from 1999–2008, the

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.192.ssad.2026.36.3.x
Received: Feb 28, 2026; Revised: Apr 5, 2026;
Accepted: Apr 7, 2026; Online: June 13, 2026

experimental results demonstrate that the Random Forest model achieves a peak accuracy of 86.90% and an F1-score of 89.94%, confirming its significantpotential for clinical early warning systems.

## 1. Methodology

### *1.1.* Data and Preprocessing

#### *1.1.1. Dataset description*

The study utilizes the Diabetes 130-US Hospitals (1999-2008) dataset from the UCI Machine Learning Repository. This multivariate dataset comprises 101,766 records of diabetic patient encounters from130 US hospitals, featuring 50 clinical attributes [4]. The dataset includes sensitive demographic information such as race, age, and gender, alongside clinical metrics like HbA1c test results and medication changes. The classification objective is to predict early readmission within 30 days of discharge—a critical factor for reducing healthcare costs and improving patient outcomes. The target variable is binary-coded: 1 for high readmission risk (readmitted <30 days) and 0 for low risk (readmitted >30 days or no readmission). Detailed descriptions of the attributes are presented in Table 1.

Table 1. Description of clinical attributes and features

| Attribute | Description |
|---|---|
| Encounter ID | Unique identifier of an encounter. |
| Patient number | Unique identifier of a patient. |
| Race | Caucasian, Asian, African American, Hispanic, and others. |
| Gender | Male, female, and unknown/invalid. |
| Age | Grouped in 10-year intervals: [0, 10), ..., [90, 100). |
| Weight | Weight in pounds (Removed due to 97% missing values). |
| Admission type | Emergency, urgent, elective, newborn, etc. |
| Time in hospital | Number of days between admission and discharge (1–14 days). |
| Diagnosis 1-3 | Primary, secondary, and additional diagnoses (ICD9 codes). |
| Glucose/A1c test | Results of serum glucose and HbA1c tests. |
| 23 Medications | Generic names for specialized diabetic medications. |

#### *2.1.2. Data cleaning and feature selection*

To ensure the integrity of the predictive models and avoid data leakage, we applied an encounter-level data split. Several redundant or highly incomplete features were removed, including Encounter ID, Patient number, Weight (97% missing), and Payer code. Duplicate records and clinical outliers were filtered to refine the dataset. Categorical variables were transformed using appropriate encoding techniques, and string-based features were replaced with numerical representations to facilitate algorithmic processing.

Regarding categorical features with cumulative missingness (e.g., Medical Specialty), we treated '?' as a distinct 'Missing' category rather than applying simple imputation. This captures the clinical context where absent information often correlates with emergency admissions. Furthermore, a 'Rare Filtering' technique was applied to reclassify codes appearing fewer than 10times as 'Rare' [User Info]. These refined features were processed via Target Encoding before being fed intothe PCA pipeline, ensuring statistical robustness.

#### *2.1.3. Data balancing using SMOTEENN*

The original dataset is highly imbalanced: 53.91% no-readmission, 34.93% readmission >30 days, and only 11.16% high-risk readmission (<30 days). To address this, we implemented SMOTEENN, a hybrid technique combining oversampling (SMOTE) to boost minority classes and undersampling (Edited Nearest Neighbors) to clean overlapping instances. This approach significantly enhances model performance by creating a balanced training environment.

#### *2.1.4. Z-score normalization*

To standardize features with varying scales and handle large value fluctuations, Z-score normalization was applied. This process transforms data to have a mean of 0 and a standard deviation of 1, as defined in Equation (1):

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

where $x$ is the original value, μ is the mean, and σis the standard deviation of the feature.

#### *2.1.5. Dimensionality reduction via PCA*

Principal Component Analysis (PCA) was utilized to reduce data dimensionality, thereby shortening processing time and mitigating overfitting. The algorithm was configured to retain at least 80% of the original variance, resulting in 26 principal components.

PCA effectively addresses multicollinearity among features (such as medication sequences and ICD-9 codes), allowing machine learning algorithms to operate more efficiently on the condensed feature space.

### *2.2. Logistic Regression*

Logistic Regression is a linear classification model that employs the logistic function (sigmoid) to predict the probability of a specific class. In this study, the model addresses binary classification by finding an optimal separating hyperplane through the minimization of the log-loss (cross-entropy) function. After training via gradient descent or the Newton method, the test samples are mapped through the following sigmoid function:

$$f(z) = \frac{1}{1+e^{-z}} \quad (2)$$

where $z$ represents the linear combination of input features. A sample is classified as high-risk (class 1) if $f(z) \geq 0.5$, and low risk (class 0) otherwise [5].

### *2.3. Decision Tree*

The Decision Tree model builds a branching structure based on optimal splitting criteria to handle heterogeneous clinical data. To determine the best split at each node, we utilize Gini Impurity (Gini Index), which measures the "impurity" of the dataset. A lower Gini Index indicates a higher degree of purity within a node. To mitigate overfitting, we constrain the model using the max_depth parameter. This algorithm is particularly effective for medical datasets as it can segment the population into homogeneous groups based on the most significant clinical variables [6].

### *2.4. K-Nearest Neighbors (KNN)*

KNN is a distance-based classification algorithm that identifies the nearest instances to a query point within the Euclidean space. The model makes local decisions based on a majority vote from the closest neighbors. While a small value may be sensitive to noise in EHR data, a larger might overlook local clinical patterns. In this study, the model calculates the distance to all training samples to assign the most probable risk class [7].

### *2.5. Random Forest*

Random Forest is an ensemble learning method consisting of multiple independent decision trees. It enhances predictive stability and accuracy by building trees on bootstrap samples and selecting random feature subsets for each node split. This "forest" approach is superior to individual trees as it evaluates feature importance and remains robust against clinical outliers and skewed data distributions. The final classification is determined through majority voting across all trees, leading to high generalization performance [8].

### *2.6. Support Vector Machine (SVM)*

SVM aims to identify an optimal hyperplane in a high-dimensional space that maximizes the margin between different classes. The points closest to this hyperplane, known as support vectors, are critical for defining the decision boundary. To handle non-linear relationships in diabetic data, we employ the "kernel trick"—specifically the Radial Basis Function (RBF) Kernel. This technique maps the input features into a higher-dimensional space, enabling the model to establish complex non-linear classification boundaries while avoiding overfitting [9].

### *2.7. Multilayer Perceptron (MLP)*

MLP is an artificial neural network consisting of an input layer, at least one hidden layer, and an output layer. Each neuron processes inputs via an activation function and transmits the result to the subsequent layer, allowing the model to learn complex non-linear correlations. MLP is highly adaptable, as the number of neurons and layers can be tuned to suit the complexity of the EHR tabular data.

### *2.8. Optimized Deep Multilayer Perceptron (Optimized Deep MLP)*

In this study, we optimized the deep neural network architecture with a hierarchical structure through progressively decreasing hidden layers (128, 64, 32) instead of using the conventional two-hidden-layer configuration (100, 50). This architecture enables the model to automatically learn feature representations at multiple levels: from raw clinical features at the 128-neuron layer to more condensed information representations in subsequent layers. The value pursued by this study lies in designing a deep neural network architecture that aligns with the hierarchical and nonlinear nature of clinical data. A summary comparison of the technical characteristics between the traditional two-layer MLP model and the three-layer hierarchical Optimized Deep MLP architecture is presented in Table 2.

Specifically, the Optimized Deep MLP model is initialized with a hidden_layer_sizes = (128, 64, 32) configuration. This structure allows the neural network to perform a more sophisticated feature representation learning process, where the first hidden layer learns basic clinical patterns (low-level patterns), while the following layers (64 and 32 neurons) perform information compression to synthesize abstract high- level feature representations.

The implementation of the ReLU activation function combined with the Adam optimizer enhances the ability to learn complex non-linear relationships and ensures stable gradient flow. To address the risk of overfitting, the research applies an L2 regularization coefficient and an early stopping mechanism based on a 10% validation split. This balanced approach achieves optimal performance between accuracy and sensitivity for

identifying high-risk patients. Furthermore, the model not only ensures stable learning on complex EHR data but also achieves an improvement in speed: the training time is only 63.52 seconds, nearly twice as fast as Random Forest at 132.02 seconds.

The detailed hyperparameters of the Optimized Deep MLP model are summarized in Table 3.

Table 2. Comparison of technical characteristics between MLP and Optimized Deep MLP

| Criterion | MLP | Optimized Deep MLP |
|---|---|---|
| Main structure | Fully connected network, processing only vectorized or flat tabular data. | Employs hierarchical architecture through multiple information-compressing hidden layers. |
| Suitable data type | Tabular data, flat data, or vectorized inputs. | Tabular data with high non-linear correlations. |
| Feature learning capability | Global learning, lacks locality, susceptible to "noise" in large datasets. | Capable of self-learning hierarchical clinical patterns through low-to-high level feature representation learning. |
| Training speed | Faster due to a simpler network architecture and fewer parameters. | Medium, depending on the depth of network layers and the number of neurons (128-64-32). |
| Scalability | Limited when processing datasets with high complexity. | Highly effective in processing Electronic Health Records (EHR) with complex non-linear correlations. |
| Criterion | MLP | Optimized Deep MLP |
| Main structure | Fully connected network, processing only vectorized or flat tabular data. | Employs hierarchical architecture through multiple information-compressing hidden layers. |
| Suitable data type | Tabular data, flat data, or vectorized inputs. | Tabular data with high non-linear correlations. |
| Feature learning capability | Global learning lacks locality, susceptible to "noise" in large datasets. | Capable of self-learning hierarchical clinical patterns through low-to-high level feature representation learning. |
| Training speed | Faster due to a simpler network architecture and fewer parameters. | Medium, depending on the depth of network layers and the number of neurons (128-64-32). |
| Scalability | Limited when processing datasets with high complexity. | Highly effective in processing Electronic Health Records (EHR) with complex non-linear correlations. |

Table 3. Hyperparameter values of the Optimized Deep MLP model

| Hyperparameter | Optimized Deep MLP Value |
|---|---|
| Number of hidden layers | 3 Dense layers (128, 64, 32) |
| Information compression | Progressively decreasing neurons |
| Hidden layer activation | ReLU |
| Output layer activation | Sigmoid |
| Batch size | 10 |
| Optimizer | Adam |
| Learning rate | Constant |
| L2 Regularization | 0.001 |
| Early stopping | True |
| Validation split | 0.1 |

## 3. Experimental Results

### *3.1. Evaluation Methods*

The performance of the classification models is evaluated using a Confusion Matrix, which provides four fundamental metrics for each class: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In the context of diabetes readmission, a "Positive" result indicates a high risk of readmission (<30 days), while a "Negative" result indicates a low risk (>30 days or no readmission).

To assess the models comprehensively, four primary metrics are calculated using the following equations:

• **Accuracy:** The overall proportion of correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

• **Precision:** The ratio of correctly predicted high-risk instances to the total predicted high-risk instances.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

• **Recall (Sensitivity):** The ability of the model to identify all actual high-risk cases.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

• **F1-score:** The harmonic mean of Precision and Recall, providing a balanced measure of performance.

$$\text{Fl} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

The research dataset exhibits a significant class imbalance, with readmitted patients accounting for only 11.16% of the total samples. To ensure a comprehensive and objective evaluation, particularly regarding the accurate identification of patients at risk of readmission—which is the clinical focus of this study—I additionally employ the Precision-Recall (PR) Curve and Average Precision (AP). AP is calculated according to Equation (7):

$$AP \approx \int_0^1 P\,(r)\,dr \quad (7)$$

AP provides a generalized measure of model performance across all classification thresholds, where a higher value reflects superior capability in accurately predicting positive cases. This enables the early warning system to operate effectively without generating an excessive number of false alarms that could overwhelm clinicians.

### 3.2. Results and Discussion

The dataset was randomly split into a training set (80%) and a testing set (20%). We benchmarked seven algorithms, ranging from traditional Logistic Regression to the Optimized Deep MLP. The comparative results are summarized in Table IV.

As shown in Table 4, Random Forest achieved the best overall performance with an Accuracy of 86.90% and the highest F1-score of 89.94%. Notably, the high Recall (92.67%) demonstrates the model's effectiveness in identifying the vast majority of patients at high risk of readmission, which is the primary objective in healthcare optimization.

Table 4. Performance evaluation of classification models (Unit: %)

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Random Forest (RF) | **86.90** | **87.36** | **92.67** | **89.94** |
| Optimized Deep MLP | 85.68 | 88.80 | 88.50 | 88.65 |
| KNN | 85.13 | 84.16 | **94.02** | 88.07 |
| MLP | 84.89 | 89.87 | 85.73 | 87.75 |
| SVM | 80.74 | 90.27 | 77.90 | 83.63 |
| Decision Tree | 77.08 | 81.70 | 82.09 | 81.90 |
| Logistic Regression | 70.41 | 78.54 | 73.12 | 75.74 |

When compared to the baseline study by Emi - Johnson et al. (2025), our proposed pipeline (SMOTEENN + PCA + RF) showed a substantial leap in performance, as detailed in Table 5. This improvement confirms that combining SMOTEENN for data balancing with PCA effectively addresses the challenge of unbalanced medical data. Fig. 1 presents the Precision-Recall (PR) curve on the held- out test set (n = 13,363) with its inherent imbalanced distribution.

Table 5. Performance improvement compared to the reference study

| Metric | DNN/XGBoost | Unbalanced RF Model | Proposed RF Model | Improvement |
|---|---|---|---|---|
| Recall | 14.30% | 37.60% | 92.67% | **+78.37%** |
| Precision | 18.60% | 14,39% | 87.36% | **+68.76%** |

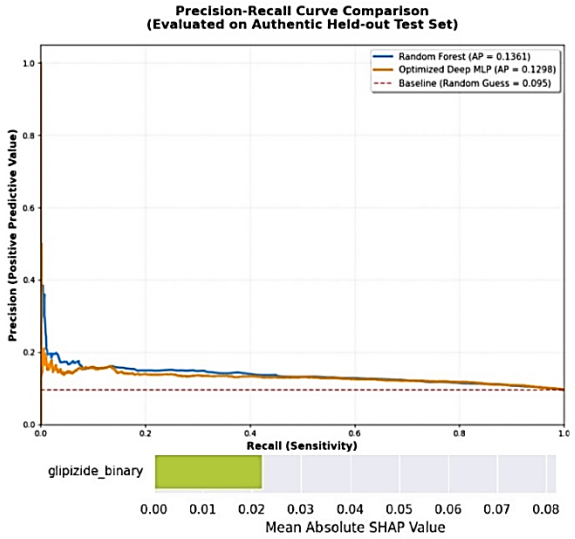


Fig. 1. PR curve on the imbalanced test set

The results show that the Random Forest model achieved an AP of 0.1361, significantly exceeding the random baseline of 0.0954 (the random baseline was derived from the proportion of the positive class in the actual test set after splitting), confirming the model's effective discriminative ability on the minority class.

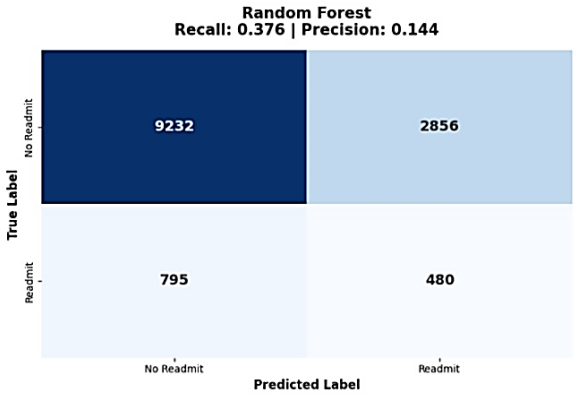


Fig. 2. Confusion matrix of the RF model on the imbalanced test set

Achieving a Recall of 38%, representing an approximately threefold improvement over previous studies, is a significant advancement in the early detection of patients at risk of readmission, even in the presence of severe data imbalance.

Furthermore, the Optimized Deep MLP model outperformed the standard MLP, particularly in terms of Recall, as shown in Table 6.

Table 6. Comparison between Optimized Deep MLP and standard MLP models

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Optimized Deep MLP | 85.68 (+0.79%) | 88.80 (-1.07%) | 88.50 (+2.77%) | 88.65 (+0.90%) |
| MLP (100, 50) | 84.89 | 89.87 | 85.73 | 87.75 |

The 2.77% increase in Recall in the Optimized Deep MLP highlights the advantage of hierarchical architecture in capturing complex, non-linear clinical correlations within Electronic Health Records (EHR). Although Precision decreased slightly, the higher F1- score indicates a superior balance, which is vital for clinical early warning systems. It should be noted that the Optimized Deep MLP is not intended to replace Random Forest in all aspects. Rather, it serves as an alternative solution specifically optimized for real-time clinical monitoring. This role is demonstrated by its high computational efficiency, with a training time nearly twice as fast as Random Forest, along with its inherent ability to capture complex nonlinear clinical correlations within EHR data that traditional decision tree-based models may overlook.

By employing SHAP for post-hoc interpretation, I provide transparency regarding how these biological factors influence risk assessment, thereby addressing concerns about algorithmic bias and ensuring fairness in prediction outcomes.

The PCA technique was used to address multicollinearity among clinical indicators with large numbers of values, combined with SHAP to quantify the contribution of each feature. This two-tiered approach maintains optimal predictive performance through PCA while preserving clinical interpretability.

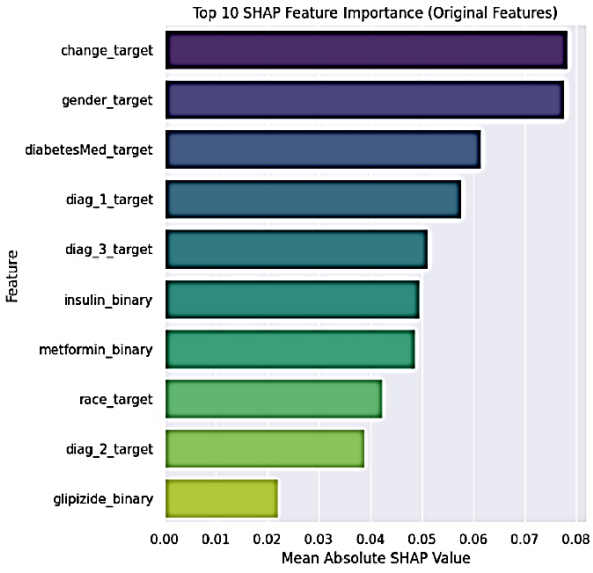


Fig. 3. Top 10 most important original features ranked by SHAP

Table 7. Differences in importance according to preprocessing procedure (PCA) and clinical significance (SHAP) Feature SHAP

| Feature | SHAP Rank | PCA Rank | Rank Difference |
|---|---|---|---|
| change | 1 | 38 | 37 |
| gender | 2 | 14 | 12 |
| diabetesMed | 3 | 36 | 33 |
| race | 8 | 6 | 2 |

This demonstrates that clinically significant factors-such as medication change (change), gender (gender), and diabetes medication usage (diabetesMed) - exhibit low variance within the data. PCA should be employed to reduce dimensionality and address multicollinearity, followed by SHAP to restore clinical interpretability, thereby ensuring both predictive performance and transparency. This approach thoroughly resolves the trade-off between accuracy and interpretability in medical applications.

## 4. Conclusion

In this study, the authors implemented an encounter-level data split (80/20 ratio), ensuring a consistent objective distribution (63.1%) across datasets. This large-scale approach allowed the proposed "golden pipeline"—combining SMOTEENN for data balancing and Principal Component Analysis (PCA) with 26 principal components—to effectively address the high-risk readmission imbalance (11.16%).A key contribution of this work is the retention of sensitive demographic features (race, age) while achieving high classification performance.

The experimental results demonstrate that the Random Forest model is the most robust solution for clinical early warning, achieving an F1-score of 89.94% and a sensitivity (Recall) of 92.67%. The Random Forest model achieved a Recall of 92.67% on the balanced training set. However, when evaluated on the imbalanced held-out test set to simulate real-world performance, it maintained a Recall of 38% and an Average Precision (AP) of 0.1361, significantly surpassing the calculated random baseline of 0.0954. Furthermore, the Optimized Deep MLP architecture (detailed in Tables II and III) successfully captured complex non-linear clinical correlations, yielding a 2.77% improvement in Recall compared to standard flat neural networks.

Future research will explore advanced balancing techniques beyond SMOTEENN and evaluate sophisticated ensemble models such as XGBoost and LightGBM. The ultimate goal remains the practical integration of the Random Forest model into hospital Electronic Health Record (EHR) systems to support real-time physician decision-making, thereby reducing 30-day readmission rates and optimizing healthcare resources.

## References

[1] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records, BioMed Research International., Apr. 2014, Art. no. 781670. https://doi.org/10.1155/2014/781670

[2] V. B. Liu, L. Y. Sue, and Y. Wu, Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes, Journal of Medical Artificial Intelligence., vol. 7, Sep. 2024.

[3] O. G. Emi-Johnson and K. J. Nkrumah, Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data, Cureus,, vol. 17, iss. 4, Apr. 2025, Art. no. e82437. https://doi.org/10.7759/cureus.82437

[4] UC Irvine, Diabetes 130-US Hospitals for Years 1999-2008, UCI Machine Learning Repository, 2014. [Online] Available:https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008. Accessed on: Feb. 28, 2026

[5] C. M. Bishop, Pattern Recognition and Machine Learning, New York, USA: Springer, 2006, ch. 4, sec. 4.3.2.

[6] L. Breiman, J. Friedman, R. A. Olshen, and C. Stone, Classification and Regression Trees, Belmont, CA, USA: Wadsworth, 1984.

[7] T. M. Cover and P. E. Hart, Nearest neighbor pattern classification, IEEE Transaction Information Theory, vol. 13, iss. 1, Jan.1967. https://doi.org/10.1109/TIT.1967.1053964

[8] L. Breiman, Random Forests, University of California, Berkeley, CA, USA, Jan. 2001.

[9] S. Keerthi and C.-J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Comput., vol. 15, iss. 7, pp. 1667–1689, Jul. 2003. https://doi.org/10.1162/089976603321891855