International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting Hospital Readmission for Diabetes Patients

Hajar Hussein AL Qahtani [1], Abdulmohsen Algarni [2]

Submitted: 16/02/2025 **Revised:** 17/03/2025 **Accepted:** 26/03/2025

Abstract: Predicting hospital readmission among diabetes patients is essential for improving patient outcomes, reducing healthcare costs, and optimizing the use of medical resources. However, this task is complex due to the intricate nature of healthcare data, high feature dimensionality, class imbalance issues, and the necessity of integrating both demographic and clinical variables. To address these challenges, a variety of machine learning models were developed and assessed, including traditional classifiers such as Decision Trees, Logistic Regression, and Random Forests, as well as more advanced approaches like XGBoost and Deep Neural Networks. To enhance model performance, we applied preprocessing techniques such as feature transformation, data balancing, and categorical encoding. Experiments were conducted on clinical datasets to predict patient readmission within 30 days, after 30 days, or not at all. Performance metrics included classification accuracy and the AUC-ROC score. Results showed that the Random Forest model achieved the highest performance in binary classification, with an accuracy of 94% and an AUC-ROC of 0.97, while a proposed Multi-Stage Classifier excelled in the multi-class task with 80% accuracy and an AUC-ROC of 0.89. Overall, the study highlights the potential of machine learning, particularly when coupled with effective preprocessing, to accurately predict hospital readmissions in diabetes care, thereby aiding clinical decisions and improving healthcare efficiency.

Keywords: *Diabetes Readmission Prediction, Machine Learning (ML), Deep Learning (DL), Classification*

1. Introduction

Diabetes is a highly prevalent chronic condition which affects millions of people across the world and significantly worsens healthcare issues. The population with diabetes is anticipated to rise to around 643 million by the year 2030, while up to 783 million can be expected to reach by the year 2045 [1]. Patients with diabetes experience repeated hospitalizations as a result of complications from poor disease management thereby resulting in increased costs, overloading the already struggling health care system, and resulting in poor patient outcomes [2], lacked the establishment of cost-effective programs for prevention. Maximizing care and avoiding unnecessary hospitalization in diabetic patients may constitute the process of predicting the readmission of diabetic patients. When clinicians identify the risk of readmission early, it can help reduce resource use, prevent the readmission itself, improve patient outcomes, and lower overall healthcare costs.

Prediction of readmission of patients may have significant positive implications especially in patients with chronic condition like diabetes as it can contribute to better patient care and reduce unnecessary hospital admission. Early detection of likely readmissions by health providers facilitates the targeted distribution of resources and minimizes the potential of unnecessary readmissions, and consequently, leads to the enhancement of patient outcomes and reduction of costs in health care [3].

But, predicting hospital readmission, especially for a complicated disease like diabetes is fraught with significant challenges. Patients' own clinical history, demographics, as well as lab tests and treatment interventions generate vast amounts of data [4]. In addition, some of the factors that complicate this task is the imbalance of classes, missing data, and irrelevant features. Good predictive modelling relies on combining different types of data such as quantitative and qualitative variables and understanding and addressing the key drivers of readmission.

In this paper, both machine learning (ML) and deep learning (DL) models were trained to predict diabetic patient's readmission rates. Many models (Decision Trees, Logistic Regression, Random Forest) and advanced techniques (XGBoost, Deep Neural Networks) were used for binary and multiclass classification tasks. Preprocessing steps involved feature transformation, data normalization, and handling class imbalance. The results imply the advantage of classical machine learning models over deep learning models for this task, due partly to effective feature selection and preprocessing techniques.

This paper aims to improve the prediction of hospital readmissions in diabetic patients through several key contributions. It presents the development and evaluation of both machine learning (ML) and deep learning (DL) models specifically tailored to this task. The study addresses data preprocessing challenges such as handling missing values and correcting class imbalances. It involves training and testing various classifiers—both binary and multi-class—to predict early readmission (within 30 days), late readmission (after 30 days), or no readmission at all. Model performance is rigorously evaluated using metrics including

[1] *Department of Data Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia,*
Email ID: hhalqahtani95@gmail.com
ORCID ID: 0009-0004-5760-8794
[2] *Department of Informatics and Computer Systems, College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia,*
Email ID: a.algarni@kku.edu.sa
ORCID ID: 0000-0002-7556-958X

accuracy, precision, recall, and F1-score, providing a comprehensive assessment across both classification tasks. Finally, the results are benchmarked against existing studies, demonstrating improved performance and underscoring the effectiveness of the proposed methodology.

The paper is outlined into six sections. Section 2 presents an analysis of the related work, highlighting earlier research efforts that utilized machine learning methodologies. Section 3 presents the proposed methodology, including preprocessing of the dataset and application of the model. Section 4 presents the results, highlighting notable performance values. Section 5 compares the findings to current research and analyzes the results. Section 6 represents a conclusion for the whole paper and a suggestion for future improvements.

2. Related Work

Several studies have explored predicting hospital readmissions for diabetic patients using ML techniques. These studies generally focus on various classification models and preprocessing methods, aiming to improve predictive accuracy. Research has demonstrated that algorithms like Random Forest (RF), XGBoost, and DL methods are successful in identifying patients at high risk, pointing to the need to handle data imbalances and choose appropriate features to enhance predictive capability.

Masoomeh Zeinalnezhad and Saman Shishehchi proposed an extensive methodology that combined meta-heuristic methods with data mining techniques to predict the likelihood of early readmission for diabetic patients within 30 days post-discharge. [5]. The study leveraged the "UC Irvine Machine Learning Repository" dataset that contained 101,765 samples along with 50 relevant features of hospital and patient outcomes. A Chi-square analysis was conducted to identify the key factors associated with readmission risk, and various classification methods—including RF, Support Vector Machines (SVM), and Neural Networks—were evaluated and compared. In addition, Genetic Algorithm was utilized to optimize the hyperparameters of SVM and thus increase its accuracy. Results showed that RF performed better than all examined approaches, having an accuracy of 74.04%, while GA-SVM improved SVM by 1.12%. The research emphasized the capability of such techniques in minimizing hospital readmission, specifically in patients with diabetes, and demands additional studies on hybrid models, as well as other meta-heuristic approaches, to improve predictive accuracy.

Mahmoud et al. proposed a research study with the purpose of predicting short-term and long-term readmissions of hospital patients who have uncontrolled diabetes based on ML techniques [6]. The objective of this study was to assist healthcare providers in enhancing patient care and minimizing readmission rates, which are a key indicator of hospital quality. On the basis of information gathered from the “Diabetes 130-US hospitals” dataset, authors applied different ML algorithms, i.e., RF, Naive Bayes (NB), SVM, etc., for predicting readmissions. The results indicated that SVM achieved the best accuracy (64.5%) in the prediction of all readmission episodes, while RF outperformed other models in the prediction of short-term readmission (86.38%). The study indicated the potential of ML for risk factor identification and facilitating timely intervention, thereby enhancing disease management and health cost reduction.

Liu et al. presented a comparative study of different ML models for predicting 30-day hospital readmission rates for patients with diabetes, using the Grey Wolf Optimizer (GWO) for feature optimization [7]. A total of 11 different ML algorithms, including XGBoost, Decision Trees (DT), and SVM, were compared with DL approaches, specifically Long Short-Term Memory (LSTM). Based on data obtained from more than 100,000 patient visits to 130 hospitals across the US, the results indicated that RF achieved the highest accuracy, F1 score, and precision. While XGBoost showed good performance, the DL models failed to outperform the traditional ML models in the framework of this research. These results highlight the effectiveness of using ML algorithms, specifically RF and XGBoost, and auxiliary feature optimization techniques like GWO, to enhance predictive accuracy for hospital readmissions among diabetic patients.

Shang et al. developed 30-day diabetic patient hospital readmission risk prediction models using ML classifiers [8]. The test was carried on a "Health Facts Database" dataset containing over 100,000 diabetic patient records highlighting 23 risk factors such as age, sex, admission type, and drug use. The authors implemented some of the ML techniques, such as RF, NB, and DT ensemble techniques, to predict the patient readmission probability. The RF algorithm outperformed the other algorithms with the highest AUC, thereby proving effective for the prediction of short-term readmission. The research revealed significant readmission risk factors of prior hospitalization, age, and count of emergency admissions that were of great value to healthcare professionals in managing high-risk diabetic patients.

Sathyavathi D. and Mary Sowjanya A. proposed a diabetic patient readmission prediction model in hospitals using ML to cut down healthcare expenses and enhance patients' quality of care [9]. The researchers utilized decision trees, random forests, CATBoost, and XGBoost ML algorithms to develop a prediction model that yielded superior results compared to other algorithms with real-time data used to test it. The model was created to predict high-risk patients for readmission within a 30-day period so that healthcare institutions can offer more care and prevent excessive readmissions. Data engineering methodologies, including feature transformation and feature selection, were also part of the study; the study additionally addressed problems related to imbalanced datasets and replicated records. A summary of the works mentioned earlier is presented in Table 1.

Previous studies on diabetes-related readmission have made important advances, yet each leaves key gaps that our work will bridge. Zeinalnezhad and Shishehchi combine meta-heuristics with classical classifiers, but their best model still achieves only mid-70 percent accuracy and they explore early (≤30-day) readmission alone, leaving later readmissions unmodelled and class imbalance largely untreated [5]. Mahmoud et al. narrow their focus to “uncontrolled” diabetes and split the task into two single-stage scenarios; their pipeline depends on manual feature selection and their results fluctuate between short- and long-term settings, suggesting instability when classes overlap [6]. Liu et al. introduce Grey-Wolf-Optimizer feature selection, yet their evaluation remains a binary 30-day task and relies on SMOTE alone to temper skewed data [7]. Shang et al. likewise restrict prediction to 30-day readmission; despite using down/over-sampling, their RF model’s AUC reveals room for improvement, and the study does not disentangle post-30-day risk factors [8]. Finally, Sathyavathi and Sowjanya demonstrate several tree-based learners, but they treat categorical encoding superficially and report no strategy for multiclass imbalance-limitations that hamper generalisation [9].

Our work will overcome these constraints by (i) tackling the *full* three-way outcome (early, late and no readmission) through a Multi-Stage classifier that decomposes the problem into sequential binary decisions; (ii) applying class-specific balancing schemes (SMOTENC for binary, targeted under-sampling for multiclass) to

Table 1. Summary of Related Work

Authors	*Title*	*Year*	*Model*	*Dataset*	*Accuracy*
Zeinalnezhad and Shishehchi [5]	"An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients"	2024	RF, GA-SVM, SVM, NN	"Diabetes 130-US Hospitals for Years 1999–2008"	RF: 74.04 %, GA-SVM: 73.52 %, SVM: 72.40 %, NN: 70.44 %
Mahmoud et al. [6]	"Short-Term and Long-Term Readmission Prediction in Uncontrolled Diabetic Patients using Machine Learning Techniques"	2023	RF, NB, SVM, AdaBoost, KNN, NN	"Diabetes 130-US Hospitals for Years 1999–2008"	RF: 63.03%, NB: 61.7%, SVM: 64.2%, AdaBoost: 62.3%, KNN: 57.7%, NN: 59.5%
Liu et al. [7]	"Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes"	2024	MLP, XGBoost, DT, RF, LR, SVM linear kernel, SVM RBF kernel, KNN, NB, AdaBoost, LSTM	"Diabetes 130-US Hospitals for Years 1999–2008"	MLP: 0.77, XGBoost: 0.88, DT: 0.79, RF: 0.88, LR: 0.64, SVM linear kernel: 0.63, SVM RBF kernel: 0.69, KNN: 0.63, NB: 0.12, AdaBoost: 0.86, LSTM: 0.77
Shang et al. [8]	"The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers"	2020	RF, NB, TE	"Diabetes 130-US Hospitals for Years 1999–2008"	N/A
Sathyavathi and Sowjanya [9]	"Predicting Hospital Readmission for Diabetes Patients Using Machine Learning"	2020	DT, RF, CATBoost, XGBoost	"Diabetes 130-US Hospitals for Years 1999–2008"	RF: 84.59%

curb skew without synthetic noise; (iii) standardising numerical features and one-hot-encoding categorical variables within a unified preprocessing pipeline rather than ad-hoc transformations; and (iv) benchmarking both traditional and deep architectures under identical folds, with comprehensive metrics beyond accuracy alone. By addressing imbalance, feature heterogeneity and outcome granularity in a single, integrated framework, our study will deliver a more stable and clinically actionable predictor than the earlier single-stage, binary-only or lightly-balanced approaches.

3. Methods

Predicting hospital readmissions among diabetic patients is a complex task due to data imbalance, heterogeneous feature types, and varying patient outcomes. The dataset is characterized by a high-class imbalance and a mix of numerical and categorical features, which required careful preprocessing. To handle the imbalance, Random Under-Sampling (RUS) was applied for multi-class classification and SMOTENC for binary classification. Then a range of ML and DL models were trained and evaluated, including Logistic Regression, Decision Trees, RF, XGBoost, DNN, and a custom Multi-Stage Classifier, to assess their predictive performance.

Fig. 1 illustrates the proposed methodology for predicting the readmission of diabetic patients in hospitals based on machine learning models. The process begins with the preprocessing of the dataset, covering data cleaning and transformation for quality and consistency. Data balancing is then applied to correct the class imbalance problem inherent in the dataset.

Then followed by the application of different ML algorithms on two distinct classification tasks: binary and multi-class classification. The algorithms implemented on both these tasks are Logistic Regression, DT, RF, XGBoost, and DNN. Multi-Stage Classifier is also employed for the multi-class classification task to handle challenging cases by breaking down into numerous stages. The multi-class classification situation predicts three possible results depending on the readmission time: readmission within 30 days (30>), after 30 days (30<), or no readmission. The binary classification problem is aimed at predicting if a patient will have early readmission within 30 days (Yes) or not (No).

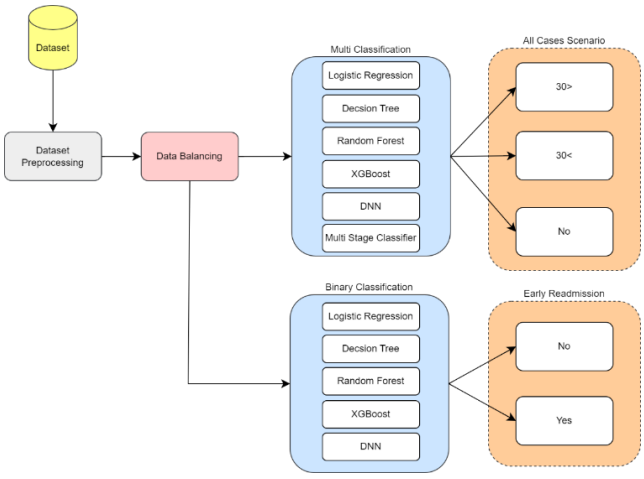

Fig. 1. Proposed Method

3.1. Handling Imbalanced Data:

Data imbalance happens when the class distribution of a training dataset for a prediction model is imbalanced. Typically, one class of samples (e.g., the positive class) is significantly smaller compared to the other classes. The reason behind this imbalance is that disease-related samples usually only account for a small percentage in the whole population, which results in an imbalanced

class distribution. Modeling on imbalanced data is an issue since most of the algorithms are frequency-biased, paying more attention to the majority class instances [10]. As a result, class imbalance may cause the classifier to predict most of the instances as normal in a bid to reduce classification error and meet the objective function [11]. In this paper, Random Under Sampler was used for multi-classification and SMOTENC for binary classification. Random Under Sampler works by randomly reducing the number of samples in the majority class to balance the dataset, which can prevent overfitting towards the majority class [12]. After applying under sampling, the dataset became 8,853 for every class in multi-classification. On the other hand, SMOTENC is an extension of SMOTE that generates synthetic samples for the minority class, specifically handling datasets with categorical features [13]. After applying SMOTENC, the dataset became 86,986 for every class in binary classification.

3.2. Model Development:

3.2.1. Decision Tree:

DT is widely used for tasks such as classification, regression, and feature selection. It is a tree-like structure that has a root, leaves, and branches. It starts at the root node and goes all the way down to the leaves. Attribute selection in Decision Tree is usually done using entropy and information gain. Entropy calculates the randomness or uncertainty of data, and information gain calculates how much reduction in entropy is achieved by choosing a specific attribute. By doing this, it generates high-quality decisions by putting the attribute with highest information gain at the root of the tree, which leads to a smaller model [14] [15].
Equation (1) illustrates how to calculate entropy, where H(S) is the entropy:

$$H(S) = \sum_{x \in X} p(\mathrm{x}) \log_2 p(\mathrm{x}) \quad (1)$$

Where $p(\mathrm{x})$ represent the probability of x. where X denotes the attributes in the dataset, including the target class. Information gain is determined using the following formula, where $\mathrm{IG(S, A)}$ represents the information gain:

$$IG(S, A) = H(S) - \sum_{i=0}^{n} p(\mathrm{x}) * \mathrm{H}(\mathrm{S}, \mathrm{X}) \quad (2)$$

Here, H(S) is the target attribute entropy, and H (S, X) is the attribute entropy conditional on the target class.

3.2.2. Random Forest

RF is a supervised learning method that can be applied for classification or regression. Random Forest chooses the best features from a random subset of features in splitting nodes, unlike conventional decision trees that seek the most effective features in splitting nodes. In comparison to the DT algorithm, RF picks observations randomly, constructs a number of trees with diverse subsets of features, and averages the result [16]. One of the significant advantages of RF is that it has the capability to prevent overfitting through the formation of random sub-trees and developing smaller trees, which are further combined together in the final model [17].

3.2.3. XGBoost

Extreme Gradient Boosting (XGBoost) is a sparsity-aware when faced with sparse data and applies weighted quantile sketch to perform approximated tree learning. XGBoost is optimized for performance and speed, but is a regular gradient-boosted decision tree implementation. Boosting works on the idea of assigning each data observation a weight. The models are constructed iteratively and weights are boosted whenever an observation gets mis-classified by a model. The ultimate ensemble model is created by integrating those decision trees [18] [19].

3.2.4. Logistic Regression:

Logistic Regression is both a classification and regression statistical method on the basis of the maximum-likelihood estimation method. Logistic regression, also known as the logit model, makes use of the sigmoid function. Logistic regression predicts the probability of the event occurring based on a linear combination of the observed values of the explanatory variables. Logistic regression has the benefit of having a clear probabilistic interpretation in the case of classification. However, one drawback of linear regression models is that they cannot solve nonlinear problems. Training the model comprises choosing parameters that specify the function to maximize the posterior likelihood function [20] [21].
For instance, if C is the number of classes as $C \in \{1,2, \ldots, C\}$ is the feature vector of dimension n. The following equation is the probability that X is in one of the C classes. The vectors $\beta_1, \beta_2, \ldots, \beta_k$ are the parameter vectors that specify the regression coefficients, and $\langle \beta_k, X \rangle$ is the inner product of the vectors.

$$\mathrm{P}(Y = k \mid x) = \frac{e^{\langle \beta_k, X \rangle}}{\sum_{i=1}^{K} e^{\langle \beta_i, X \rangle}} \text{ for } k = 1,2, \ldots, k \quad (3)$$

Where p is the probability of success, K is the class label, and k | x indicates that x is in the k th class label. βk coefficients are learned during training. Equation (4) below will be used to predict the result for the feature vector X.

$$\begin{aligned} k^* &\in \operatorname{argmax} \Pr(Y = k \mid X), k \in \{1,2, \ldots, K\} \\ k^* &\in \operatorname{argmax} \langle \beta_x, X \rangle, k \in \{1,2, \ldots, K\} \end{aligned} \quad (4)$$

3.2.5. Deep Neural Network (DNN)

An Artificial Neural Network (ANN) is designed to mimic the structure and functioning of the human brain. Since neural networks (NN) are powerful nonlinear discriminators in the event of problems in classification, because they are able to describe any decision boundary in the feature space. In recent years, Deep Neural Networks (DNNs) gained significant interest in medical research and evolved from Shallow Neural Networks (SNNs). The feature abstraction ability in DNNs and the ability to represent highly complex patterns make them extremely useful in applications in DL. Because of their ability to represent data in a good way, DNNs are in high demand in order to design efficient and robust solutions [22] [23].
The results are produced in a DNN based on the connection weights and activation functions in the neurons. The DNN is composed of multiple processing layers, and every layer contributes to decision-making and feature extraction. Several hyperparameters dictate the operation of a DNN and are to be determined in advance, including the number of units, number of layers, weights and bias initializers, activation function, regularizes coefficient, learning rate, and the optimizer. In this DNN model, ReLU activation is applied in the input layer and in every hidden layer. The ReLU function is a piecewise linear function and returns the same input in the situation where the input is a positive number and a value of zero in the situation where the input is a negative number. The neurons activated by this function are also rectified linear activation units [24].

$$ReLU(x) = \max(0, x) \quad (5)$$

3.2.6. Multi-Stage Classifier:

In this paper, a multi-stage ML classifier was used to address the complexity of the classification task by breaking it down into sequential decision-making steps, thereby improving the ability of model to differentiate between the classes effectively. Random Forest was used as a backbone for Multi-Stage Classifier

3.2.6.1. First Stage:

The first stage of the classifier focuses on separating the majority class (Class 0) from the remaining classes (Class 1 and Class 2). This phase approaches the problem as a binary classification task, aiming to distinguish and identify instances that fall under the majority class. A ML model is trained to distinguish Class 0 from the other classes. Instances that are confidently classified as Class 0 in this stage are directly assigned to that class, while the remaining instances proceed to the second stage for further classification.

3.2.6.2. Second Stage:

The second stage is engaged for cases not falling under Class 0 as determined by the first stage. Those cases are then forwarded to a second machine learning model, specifically designed to distinguish between the two remaining classes (Class 1 and Class 2). In this stage, the problem is addressed as another binary classification problem, but this time with the sole target of discriminating between the two minority classes. Now, the model is trained on a specific subset of data from these classes, thereby making it learn the specific patterns and features for this discrimination.

3.2.6.3. Final Predictions:

The final predictions are made by taking a combination of results from both stages. Examples classified as Class 0 in the initial stage are strictly classified into that class. For examples that proceed to the subsequent stage, the classification decisions made by the second model decide their classification into Class 1 or Class 2. This multi-stage process is responsible for the model first solving the easier binary classification problem before dealing with the harder differentiation between the other classes.

3.3. Evaluation Metrics:

3.3.1. Accuracy:

Evaluating the overall model's performance in predicting readmission rates.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (6)$$

3.3.2. F1-score:

Balancing precision and recall to calculate the model's performance in identifying true readmissions.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

3.3.3. Recall:

The model's prediction of which patients are likely to be readmitted and which should not be admitted.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

3.3.4. Precision:

Evaluating the model's prediction for readmissions without incorrectly classifying non-readmitted cases.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (9)$$

3.3.5. Confusion Matrix:

Providing a detailed breakdown of correct and incorrect predictions regarding readmissions.

3.3.6. AUC-ROC:

Assessing the model's performance in differentiating between the patients who would be readmitted and those who would not be readmitted.

4. Result:

In this paper, the evaluation of the performance of multiple ML and DL models in predicting hospital readmissions among diabetic patients, covering both binary and multi-class classification tasks. For binary classification, the aim was to distinguish between patients' readmission within ≤30 days and other classes. In the multi-class setting, readmissions were categorized into three classes: "no readmission", "readmission within 30 days", and "readmission after 30 days". To address class imbalance, Random Under-Sampling (RUS) was employed for multi-class classification and SMOTENC for binary classification.

4.1. Dataset:

The dataset was obtained from the "UCI Machine Learning Repository" [25]. It was collected from a voluntary program named Health Facts with the objective of creating a database for those institutions that utilize the "Cerner Electronic Health Record System" [26]. The dataset comprises extensive details about patients from the hospitals involved, including emergency, outpatient, and inpatient. Data gathered comprises the patient ID, demographics, diagnosis, length of hospital stay, laboratory tests, test results, etc. The 1999-2008 Diabetes 130-US Hospitals dataset has 101,766 observations and comprises 50 features, of which 13 are numerical and 37 are categorical. Features include patients' administrative data, medication type and count, diagnosis, and laboratory tests for diagnosis. The data also contains the "Readmitted" column, showing if the patient has readmitted within 30 days, more than 30 days, or not at all.

Fig. 2 is a pie chart breaking down the distribution of readmission classes in the dataset. The biggest slice, constituting 53.9% of the data, is for non-readmitted patients. The second biggest slice, constituting 34.9%, is for patients readmitted more than 30 days later. The smallest slice, constituting 11.2%, is for patients readmitted within 30 days of initial discharge.

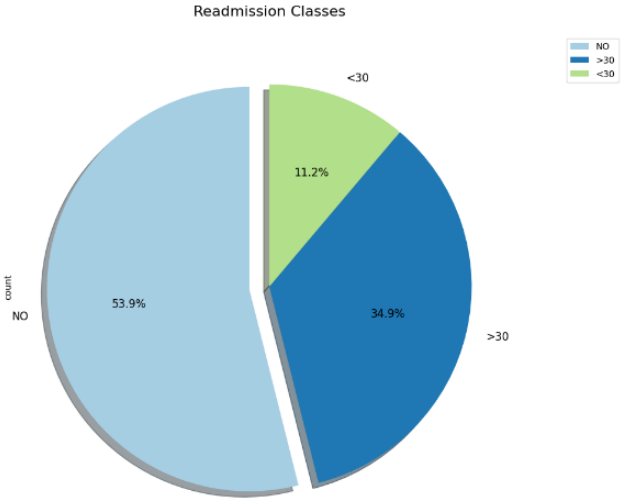

Fig. 2. Readmission Classes

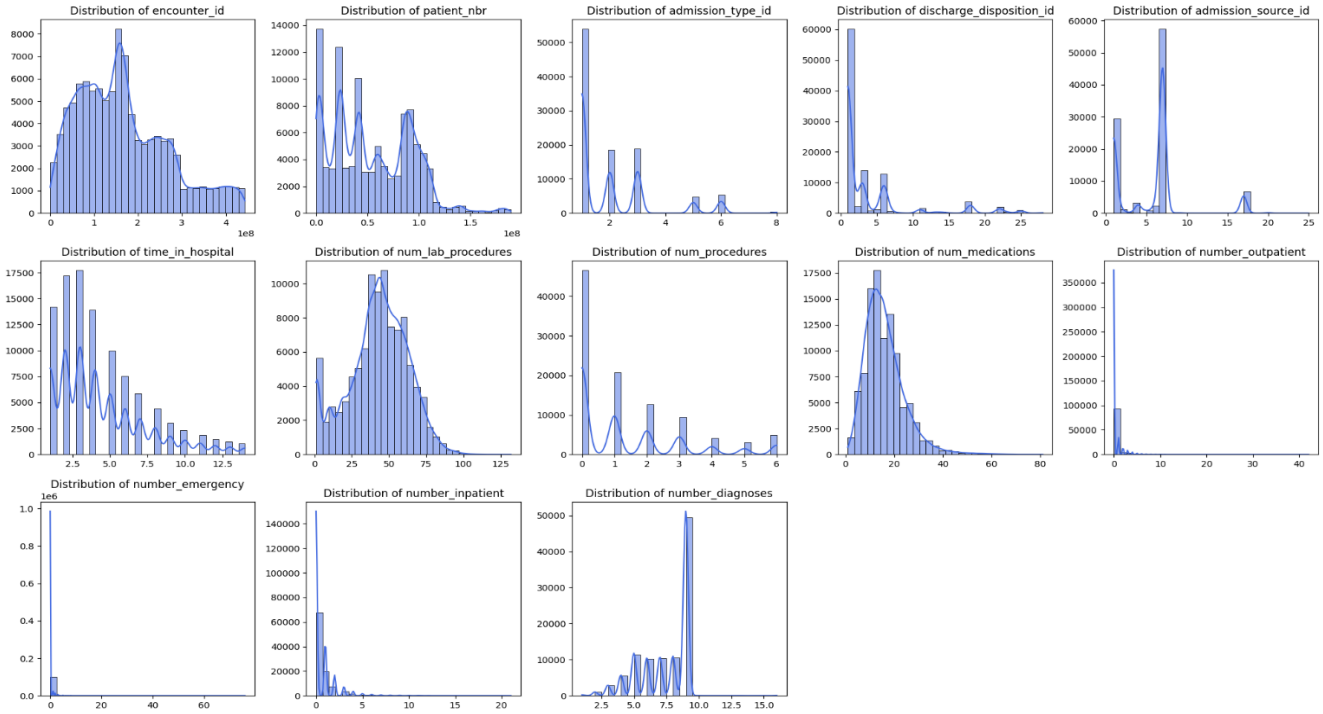

Fig. 3. Histograms that visualize the distribution of various numerical variables from a dataset

Fig. 3 presents a series of histograms that visualize the distribution of various numerical variables from a dataset. Each histogram is accompanied by a Kernel Density Estimate (KDE) curve, providing a smoothed display of the data distribution. The figure is organized into a grid with 3 rows and 5 columns, resulting in 15 subplots. From the chart, it can be observed that there are no extreme or sharp outliers in the data. The distributions appear to be relatively consistent, with no significant skewness or abrupt deviations that would indicate the presence of unusual or extreme values.

In Fig. 4 below, after observing this chart, it becomes clear that there is no significant difference between males and females in terms of their impact on readmission rates. The proportions of readmissions for both genders are approximately equal, indicating that gender does not play a notable role in influencing whether a patient is readmitted or not.

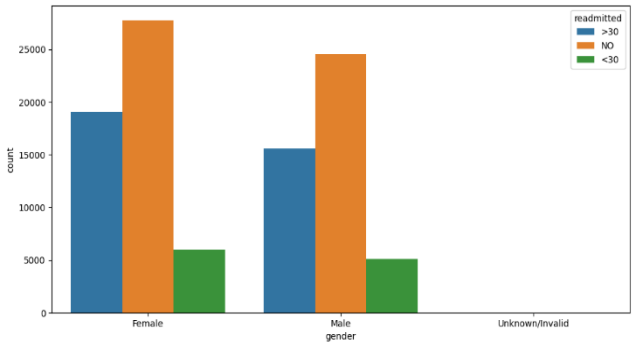

Fig. 4. Distribution of readmission rates by gender.

Fig. 5 represents a chart that analyzes the relationship between age and readmission rates. It demonstrates that age has a moderate influence on readmission, as the likelihood of readmission tends to increase with advancing age. This means the older a person is, the more likely he or she will be readmitted to the hospital compared to younger patients. The trend highlights the importance of considering age while designing healthcare interventions aimed at cutting down readmissions.

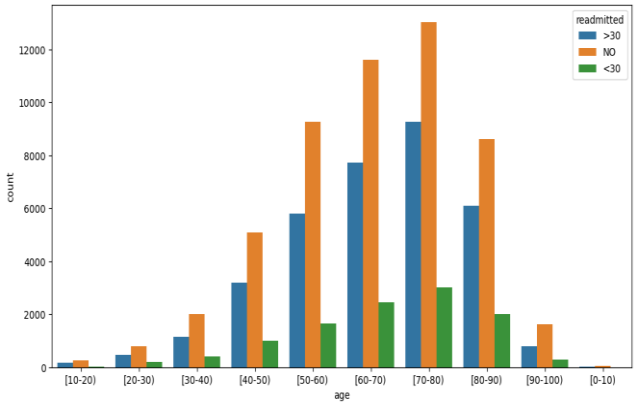

Fig. 5. Distribution of readmission rates across different age groups

4.2. Data Preprocessing:

4.2.1. Dataset Cleaning:

For the quality and relevance of the dataset, columns that were either not predictive or had numerous null values were dropped. These were columns like "examide", "citoglipton", "weight", "payer_code", and many others which pertained to medication or test results that were missing or not relevant for prediction purposes. Rows with null values were also dropped for a clean and consistent dataset. This was done to remove noise and have quality data that was used for modeling and analysis.

4.2.2. Feature Transformation:

The diagnosis columns (diag1, diag2, diag3) were converted to ICD-9 codes so that their representation would be standardized. This made the data consistent and relevant because ICD-9 codes are utilized everywhere in healthcare for classifying diagnoses. With this standardization of columns, the data was made readable and analysis-ready.

4.2.3. Column Removal:

Uninformative columns like "encounter_id", "patient_nbr", "admission_source_id", "discharge_disposition_id", "admission_type_id", and "medical_specialty" were deleted. These columns consisted of either unique identifiers (i.e., patient and encounter IDs) or categorical variables that were uninformative for the predictive model process. Deletion of these columns down-sized the dataset and eliminated unnecessary complexity.Their removal streamlined the dataset from 50 to 27 key features, reducing noise and computational complexity. The retained features (e.g., race, gender, time_in_hospital, diagnosis codes [diag_1-3], medication histories, and lab/visit counts) were selected for their clinical relevance to readmission prediction. Post-cleaning, the dataset contained 98,053 entries, all non-null, ensuring robustness for subsequent modeling.

4.2.4. Numerical Features:

Numerical features were standardized using the StandardScaler to normalize their distributions. The scaling standardizes data to a mean of 0 and standard deviation of 1, which is needed by most ML algorithms having an assumption of normal distribution in data. Equation (10) represents the formula for standardization:

$$z = \frac{x-\mu}{\sigma} \quad (10)$$

In this formula, z represents the standardized value, the original value is represented by x, μ denotes the mean of the feature, and σ represents the standard deviation of the feature.
This process normalizes numerical features to a consistent scale, ensuring that features with larger magnitudes do not disproportionately influence the model.

4.2.5. Categorical variables

were transformed with OneHotEncoder in order to be able to express them in binary form (0 or 1), which machine learning models are okay with. One-hot encoding creates new binary columns for each category in a categorical feature so that the model can deal with these features appropriately. For example, a categorical feature "gender" with categories "Male" and "Female" would be one-hot encoded into two binary features: "gender_Male" and "gender_Female."

4.3. Hyperparameter Settings

To ensure optimal model performance, hyperparameters were carefully selected and tuned for each machine learning algorithm. For Logistic Regression, max_iter=1000 was set to guarantee model convergence during training. The Decision Tree classifier was configured with max_depth=100 to maintain sufficient complexity while avoiding underfitting. In this paper, the Random Forest implementation utilized 1000 estimators (n_estimators=1000) with a maximum depth of 40 (max_depth=40) to create a robust ensemble model capable of handling complex patterns in the data. The XGBoost model was similarly configured with 1000 estimators and a maximum depth of 40, with the additional specification of eval_metric='logloss' to optimize for binary classification tasks. For the Deep Neural Network architecture, a sequential model was implemented featuring multiple hidden layers with ReLU activation functions and strategically placed dropout layers (with rates between 0.1 and 0.3) to prevent overfitting. The output layer employed softmax activation for multi-class prediction tasks. A key innovation in the approach of this paper was the development of a Multi-Stage Classifier using Random Forest as the base algorithm. This hierarchical classification system operates in two distinct phases: First, it performs binary classification to separate non-readmitted patients (Class 0) from those requiring readmission (Classes 1 and 2). Cases identified as potential readmissions then proceed to a second classification stage where they are further categorized as either early (≤30 days) or late (>30 days) readmissions. This staged approach allows for more precise classification by progressively addressing increasingly subtle distinctions between patient groups. The selected hyperparameters for each model are presented in Table 2.

Table 2. Hyperparameters

Model	*Key Hyperparameters*
Logistic Regression	max_iter=1000
DT	max_depth=100
RF	n_estimators=1000, max_depth=40
XGBoost	n_estimators=1000, max_depth=40
DNN	5 layers (1024–128 neurons), dropout=0.1–0.3

4.4. Performance of Best Models

4.4.1. Binary Classification (SMOTENC-balanced data):

RF had the best performance with 94% accuracy and 97% AUC-ROC, demonstrating its superiority in distinguishing early readmissions. XGBoost followed closely with 93% accuracy and 95% AUC-ROC, while DNN yielded 89% accuracy as shown in Table 3 and Fig. 6.

Table 3. Binary Classification (SMOTENC)

Model	*Accuracy*	*F1-Macro*	*AUC-ROC*
Logistic Regression	86%	0.86	0.92
DT	87%	0.87	0.87
RF	94%	0.94	0.97
XGBoost	93%	0.93	0.95
DNN	89%	0.89	0.94

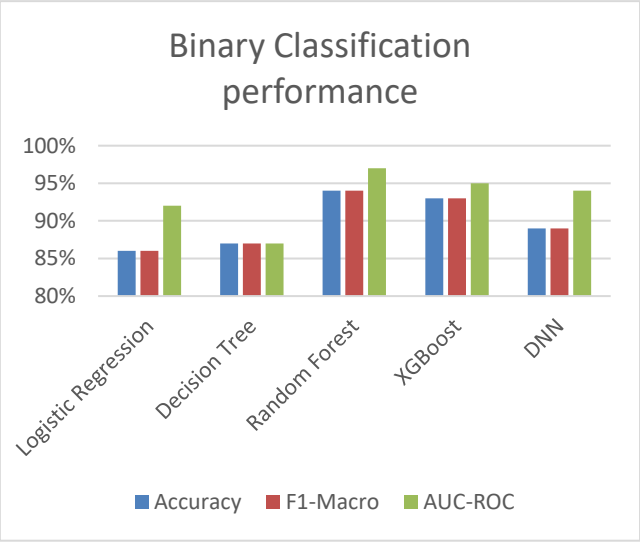

Fig. 6. Binary classification results showing the superiority of the RF algorithm.

The ROC curve for the RF model demonstrates near-perfect

classification performance, with an AUC of 0.97 as shown in Fig. 7. This high AUC value reflects the model's strong ability to rank patients at risk of readmission within 30 days above those without readmission. For context, an AUC of 1.0 signifies flawless prediction, while 0.5 implies no better than chance.

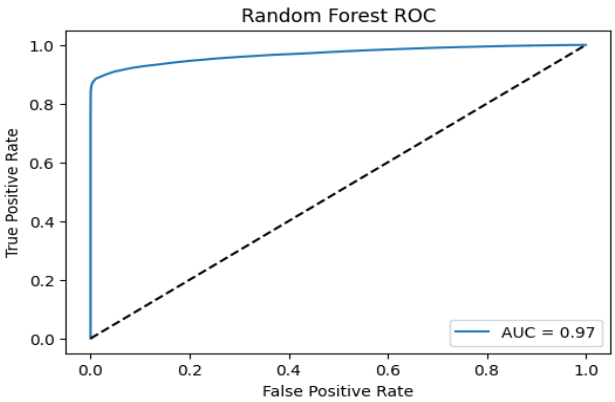

Fig. 7. ROC curve for the Random Forest model in binary classification of hospital readmissions (⩽30 days vs. others).

4.4.2. Multi-Class Classification (RUS-balanced data):

The Multi-Stage Classifier (RF-RF) significantly outperformed single-stage models, achieving 80% accuracy, and 89% AUC-ROC. In contrast, standard RF and XGBoost attained 45% and 42% accuracy, respectively, highlighting the challenge of multi-class prediction. The DNN performed comparably to RF (45% accuracy) but lagged behind the multi-stage approach as. shown in Table 4 and Fig. 8.

Table 4. Multi-Class Classification (RUS)

Model	*Accuracy*	*F1-Macro*	*AUC-ROC*
Logistic Regression	49%	0.41	0.64
DT	37%	0.34	0.52
RF	45%	0.40	0.63
XGBoost	42%	0.38	0.59
DNN	45%	0.40	0.50
Multi-Stage (RF-RF)	80%	0.76	0.89

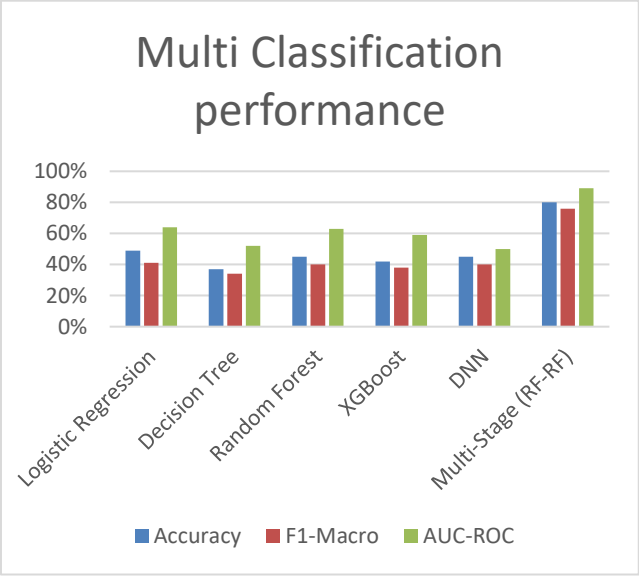

Fig. 8. Multi-classification results showing the superiority of the multi-stage algorithm.

Fig. 9 represents a confusion matrix that shows the ability of the multi-stage model to differentiate between the categories in the data. The biggest difficulty that the model faces is in distinguishing between no readmission and readmission after 30 days.

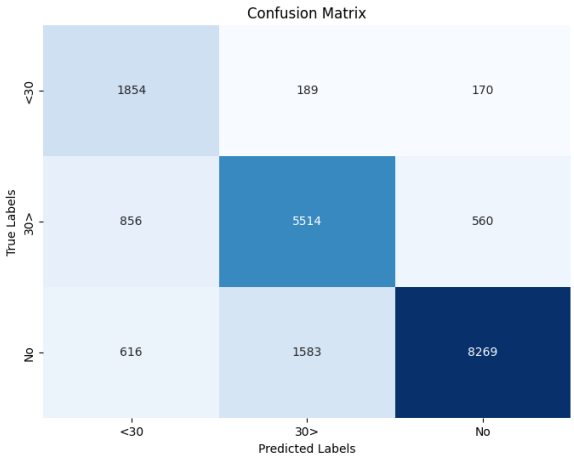

Fig. 9. Multi Stage Model Confusion Matrix

5. Discussion:

This paper significantly advances the predictive modeling of diabetic patient readmissions by addressing critical limitations in both binary and multi-class classification tasks, building upon the works of Mahmoud et al. and Shang et al. [6] [8]. While Shang et al. focused exclusively on binary classification (30-day readmission), and Mahmoud et al. examined both binary and multi-class scenarios, the methodological innovations—particularly the Multi-Stage Classifier and advanced imbalance-handling techniques—yielded superior performance across all prediction tasks. As shown in Table 5, the RF model achieved 94% accuracy and 0.97 AUC-ROC in binary classification, surpassing both Shang et al.'s RF AUC: 0.64 with over-sampling, Mahmoud et al.'s RF (86.38% accuracy, 0.63 AUC) For multi-class prediction, the Multi-Stage RF (80% accuracy, 0.89 AUC) outperformed Mahmoud et al.'s single-stage models, which struggled with class overlap (e.g., their SVM: 64.5% accuracy for "all readmissions").

Table 5. Multi-Class Classification (RUS)

Study	*Task*	*Model*	*Accuracy*	*AUC-ROC*
Shang et al. (2021) [8]	Binary Early Readmission	RF (Over-sampled)	—	0.64
Mahmoud et al. (2023) [6]	Binary Early Readmission	RF	86.38%	0.63
Mahmoud et al. (2023) [6]	Multi-Class	SVM	64.5%	0.60
This paper	Binary Early Readmission	RF	94.0%	0.97
This paper	Multi-Class	Multi-Stage RF	80.0%	0.89

6. Conclusion

This paper developed and evaluated multiple ML and DL models to predict hospital readmissions among diabetic patients, utilizing the "UCI Diabetes 130-US Hospitals" dataset. Key challenges such as class imbalance and heterogeneous data types were addressed through data balancing techniques like SMOTENC and Random Under-Sampling, along with appropriate preprocessing

steps. Traditional models, particularly Random Forest and XGBoost, demonstrated superior performance, achieving up to 94% accuracy and 0.97 AUC-ROC in binary classification. In the multi-class setting, the Multi-Stage Classifier significantly improved predictive accuracy, reaching 80% accuracy and 0.89 AUC-ROC. These results emphasize the effectiveness of classical machine learning methods when combined with proper preprocessing strategies, particularly in complex healthcare prediction tasks. Future improvements may include exploring hybrid ensembles, incorporating temporal features, and applying interpretability methods to support clinical decision-making.

Acknowledgements

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] C. J. Ejiyi, Z. Qin, J. Amos, M. B. Ejiyi, A. Nnani, T. U. Ejiyi, V. K. Agbesi, C. Diokpo and C. Okpara, "A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," Healthcare Analytics, vol. 3, p. 100166, 2023.

[2] B. Sly, A. W. Russell and C. Sullivan, "Digital interventions to improve safety and quality of inpatient diabetes management: A systematic review," International Journal of Medical Informatics, vol. 157, p. 104596, 2022 .

[3] S. Davis, J. Zhang, I. Lee, M. Rezaei, R. Greiner, F. A. McAlister and R. Padwal, "Effective hospital readmission prediction models using machine-learned features," BMC health services research, vol. 22, no. 1, p. 1415, 2022.

[4] Y. Xue, D. Klabjan and Y. Luo, "Predicting ICU readmission using grouped physiological and medication trends," Artificial Intelligence in Medicine, vol. 95, pp. 27-37, 2019.

[5] M. Zeinalnezhad and S. Shishehchi, "An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients," Healthcare Analytics, vol. 5, p. 100292, 2024.

[6] M. Mahmoud, M. Bader and J. McNicholas, "Short-Term and Long-Term Readmission Prediction in Uncontrolled Diabetic Patients using Machine Learning Techniques," in 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023), 2023.

[7] V. B. Liu, L. Y. Sue and Y. Wu, "Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes," Medical Artificial Intelligence, vol. 7, p. 23, 2024.

[8] Y. Shang, K. Jiang, L. Wang, Z. Zhang, S. Zhou, Y. Liu, J. Dong and H. Wu, "The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers," BMC Medical Informatics and Decision Making, vol. 21, no. 2, p. 57, 2021.

[9] D. Sathyavathi and A. Mary Sowjanya, "Predicting Hospital Readmission for Diabetes Patients Using Machine Learning," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 7, no. 11, pp. 1006-1012, 2020.

[10] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[11] P. Branco, L. Torgo and R. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions.," arXiv, 2015.

[12] T. Elhassan, A. M, A.-M. F and M. Shoukri, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," Global Journal of Technology and Optimization, vol. 1, 2016.

[13] A. P. Ratnasari, "Performance of Random Oversampling, Random Undersampling, and SMOTE-NC Methods in Handling Imbalanced Class in Classification Models," International Journal of Scientific Research and Management (IJSRM), vol. 12, no. 4, pp. 494-501, 2024.

[14] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," IEEE Access PP():1-1, vol. 12, pp. 86716 - 86727, 2024.

[15] Y. Izza, A. Ignatiev and J. Marques-Silva, "On Explaining Decision Trees," arXiv, 2020.

[16] M. Fratello and R. Tagliaferri, "Decision Trees and Random Forests," in Reference Module in Life Sciences, 2018.

[17] X. Zhou, P. Lu, Z. Zheng, D. Tolliver and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," Reliability Engineering & System Safety, vol. 200, p. 106931, 2020.

[18] C. Bentéjac, A. Csörgő and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," arXiv, 2019.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," arXiv, 2016.

[20] M. Maalouf, "Logistic regression in data analysis: An overview," International Journal of Data Analysis Techniques and Strategies, vol. 3, no. 3, pp. 281-299, 2011.

[21] M. K. Chung, "Introduction to logistic regression," arXiv, 2020.

[22] E. Jawad, "THE DEEP NEURAL NETWORK-A REVIEW," IJRDO -JOURNAL OF MATHEMATICS, vol. 9, no. 9, pp. 1-5, 2023.

[23] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," arXiv, 2020.

[24] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," arXiv, 2018 .

[25] B. Strack, J. P. DeShazo, K. J. Cios and J. N. Clore, "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 March 2025].

[26] 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, vol. 2014, 2014.