International Journal on
Advanced Science
Engineering
Information Technology

Vol.11 (2021) No. 5
ISSN: 2088-5334

Predicting Diabetic Patient Hospital Readmission Using Optimized Random Forest and Firefly Evolutionary Algorithm

Nida Aslam [a,*1], Irfan Ullah Khan [a,*2], Samar Alkhalifah [a], Sarah Abbas AL-Sadiq [a], Shahad Wael Bughararah [a], Meznah Abdullah AL-Otabi [a], Zainab Mohammed AL-Odinie [a]

[a] *College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia*
*Corresponding author: [*1]naslam@iau.edu.sa, [*2]iurab@iau.edu.sa*

Abstract— **Diabetes is one of the most prevailing diseases worldwide. The number of hospitalized patients with diabetes is usually huge. Readmission in the hospital is expensive, and early prediction of diabetes patient's hospital readmission can reduce the cost and help healthcare professionals evaluate the quality of healthcare services at the hospital. The proposed study aimed to develop an early prediction model for diabetes readmission and identify the significant factors that lead to readmission of diabetes patients. The early prediction will reduce the risk of hospital readmission. Several machine learning classifiers, such as Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), were applied. Firefly bio-inspired technique was used for feature selection and model optimization. Synthetic Minority Oversampling Technique (SMOTE) was applied to alleviate the data imbalance problem. The performance of the classifiers was compared using different feature sets. Experiments showed that RF outperformed the other models using reduced features selected by the Firefly algorithm. The study achieved the highest accuracy, precision, recall, and Area Under Curve (AUC) of 0.99, 0.99, 0.94, and 0.98, respectively. The results show the significance of the proposed model in diabetes readmission prediction. As a result, it is suggested that other system models and multiple data sets be investigated in order to obtain better results and identify significant features for early readmission prediction in diabetic patients.**

Keywords—**Diabetes patient's hospital readmission; optimized random forest; firefly technique; bio-inspired; SMOTE.**

Manuscript received 12 Jan. 2021; revised 2 Jan. 2021; accepted 3 Mar. 2021. Date of publication 31 Oct. 2021.

I. Introduction

Diabetes is one of the commonly occurring diseases worldwide and contains high glucose levels in the blood due to insulin production problems. Uncontrolled diabetes leads to serious complications to the cardiac, kidney, and eyes, respectively. Over the last several years, the number of diabetic patients is increasing at an exponential rate. According to World Health Organization (WHO) report, approximately 422 million people are affected by diabetes worldwide. Furthermore, 1.6 million people died due to diabetes in low and middle-income countries [1].

Readmissions are usually expensive to the patient and indicate inadequacies in the healthcare system. The hospital readmission rate indicates the overall quality of the healthcare services in the hospital. Readmission is a critical healthcare quality measure for reducing costs. A financial penalization is set whenever a hospital exceeds the permitted rate of readmission, i.e., 30-days. Identifying patients at high risk of readmissions not only improve healthcare but also reduces expenses for the patient. Machine learning has been widely used to predict the readmission in the hospital for various categories of patients such as pediatric care, esophagectomy, oral cancer, and diabetes, respectively [2]–[6]. Several studies have been conducted for the diabetes patient's readmission using machine learning. Some of them are discussed below.

Several factors like inpatient visit, deposition, and admission type were recognized as the significant factors for diabetic patient readmission. Moreover, laboratory tests and disposal discharges can be used to predict readmission of the patient being admitted after a brief period of discharge less than 30 days or longer [7]. They used several classification algorithms such as Naïve Bayes (NB), Random Forest (RF), Ada Boost (AB), and Multi-Layer perceptron (MLP). Furthermore, the Associative Rule Mining (ARM) technique was used to identify the important features. Similarly, another study by Duggal *et al.* [8] used NB, Decision Tree (DT), and Logistic Regression (LR) for exploring the impact of feature selection for early prediction of diabetes patients hospital readmission. The study used the data of patients with the age of 18 and above. All the selected models have substantially outperformed simple approaches of pretreatment techniques. However, DT outperformed the other classifiers with the range [0.56-0.68] to [0.83-0.86] over-absorption strategy. Furthermore, Duggal *et al.* [9] made another study using RF classifier, and risk factors such as readmission department, patients history, Length of stay(LOS) are the key features for prediction and achieved Area Under precision-recall Curve (ROC) of 0.296.

A hybrid feature selection technique was proposed to identify the significant features for diabetes readmission using 19 features. Improved Support Vector Machine (SVM) for classification and genetic algorithms were used for optimization. The study achieved an accuracy of 0.81 for identifying at-risk hospital readmission of diabetes patients [10]. Furthermore, the imbalance was alleviated using SMOTE technique. The study proved the significant impact of SMOTE technique on the overall prediction performance. Furthermore, Ghazo [11] developed a hybrid ensemble model for the diabetic readmission. Several classifiers such as SVM, MLP, DT, K-Nearest Neighbor (KNN), LR, GNB (Gaussian Naïve Bayes) and Probabilistic Neural Network (PNN) was used in ensemble using grid search optimization and genetic algorithms feature selection technique. The model achieved the highest AUC of 0.81 with 10-fold cross-validation. Similarly, another study [12] used several classifiers such as LDA, RF, KNN, and NB, respectively. The study found that women are at high risk of readmission. NB outperformed the other classifier with the accuracy of 0.56 and AUC of 0.64.

Additionally, another study used boruta feature selection algorithm and generic ensemble model for developing a risk prediction model for patients' readmissions [5]. The study includes patient participation, yet the anonymized data set study consists of 55 attributes and a sample size of 100 K instances for 10 years data obtained from 130 hospitals in United States. Classifiers such as SVM, recursive partitioning and regression tree, Gradient Boosting Method (GBM) and General Linear Model (GLM) was used to identify patients more likely to be readmitted and achieved an accuracy of 0.97 to classify patients who are vulnerable to readmission.

Conventional Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) models comparative analysis was made for predicting hospital readmission of diabetic patients [13]. They found that CNN outperformed the traditional ANN using 70,000 diabetic patient's data from 130 hospitals. SMOTE was used for handling data imbalance. The model achieved an accuracy: 0.92 and AUC of 0.95. Recently, a study made by Ramirez *et al.* [14] used machine learning model using the same dataset of 130 hospitals. Some of the features in the data set was normalized and reduced from 100 to 45. Four classifiers were used, such as Single Layer Perceptron (SLP), MLP, LR, and RF with 10-fold cross validation. The study converted the class label to binary class combining no readmission and >30 and labeled as no readmission. The study outperformed the previous studies by achieving 0.99 accuracy using 45 features for the binary class. Recently, a study for diabetes hospital readmission used deep learning model and achieved an accuracy of 0.95 and AUC of 0.97 for the binary class [15]. The study found that change in the medication can increase the chance of readmission. Furthermore, another study [16] also used RF and achieved an AUC of 0.84, accuracy of 0.83. The study aimed to identify the optimal medication for reducing the risk of readmission.

Despite of several studies have been made to predict diabetic patient's readmission but there is still a room for the improvement. The aim of the studies was to develop a prediction model for reducing financial costs, discomfort for patients, and maintaining positive hospital reputation. Similarly, the proposed study investigates the effectiveness of the metaheuristic evolutionary Firefly technique in diabetes patients' readmission. The firefly technique was used for feature selection and model optimization. Three classification techniques will be used namely, LR, DT, and optimized RF. To alleviate the data imbalance, synthetic data was generated using SMOTE. Several data preprocessing techniques was applied to clean the data

II. Material and Method

A. *Dataset* Description

The dataset used in the study contains health records of 100k patients from the year 1999 to 2008 (10 years) of 130 hospitals and clinical care in the US [17]. It consists of 50 patients' potential risk factors and a "readmitted" target feature, indicating if the patient is readmitted within or after 30 days or never readmitted after being discharged from the hospital. The data set suffers from a huge imbalance between the never readmission class label and less than 30 days readmission class label. Figure 1 presents number of records per class label. The data set contains several categories of features such as demographic, hospital visit details, diagnosis, drugs, and medication dosage.

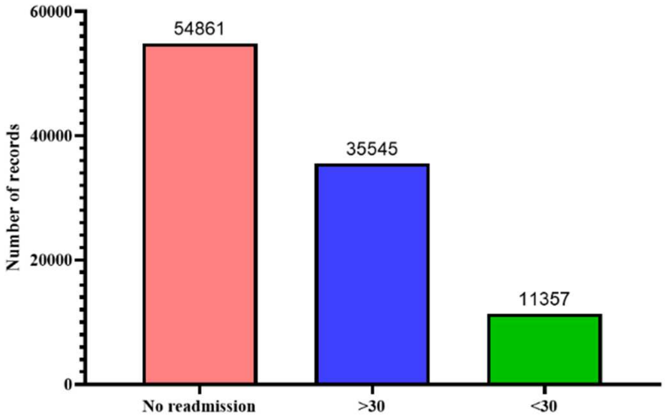

Fig. 1 Number of records per class category in the data set

The data set contains some demographic features such as age, gender, weight, and race. Figure 2 shows the age range of the patients for three categories in the data set. Most of the patient's age in the dataset is in the range of [60-80). Similarly, Figure 3 shows diabetes patient's race per category. Most of the patients that were readmitted before 30 days were African American. However, the highest number of patients admitted and admitted after 30 days category were Caucasian. Figure 4 shows the gender-wise distribution of the data. Many of the patients were female and in the age range of 70 to 80.

Nearly all the features in the data set contains categorical data. Table 1 contains the statistical distribution of the numerical features in the data set. For the numeric attributes mean and standard deviation was used. As per the below table, the time in hospital and the number of diagnosis attribute has similar mean for all the three categories of the dataset. However, the number of emergencies is higher for before 30 days of readmission compared with the other two class categories.

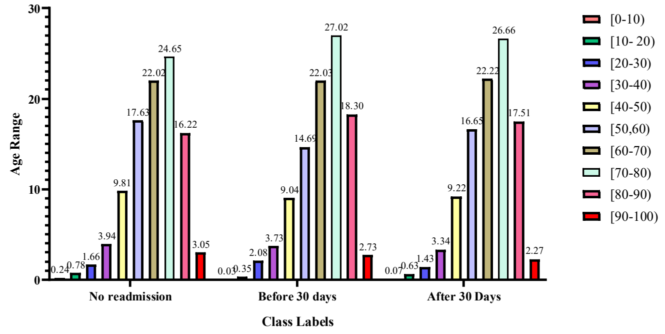

Fig. 2 Age range per class category in the data set

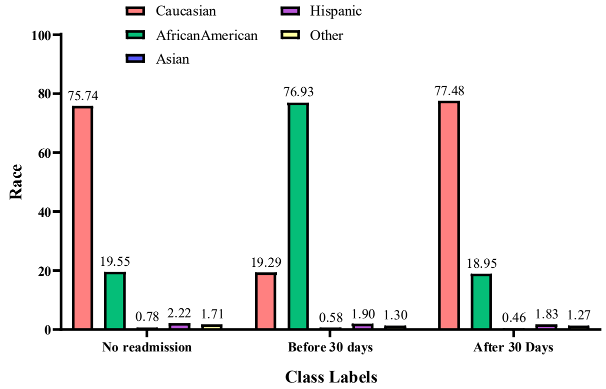

Fig. 3 Patient's race per class category in the data set

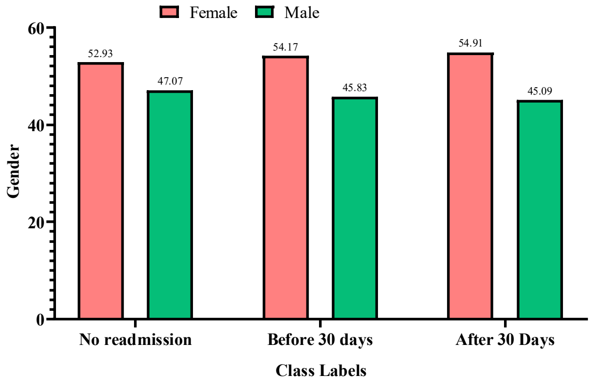

Fig. 4 Patient's gender per class category in the data set.

TABLE I
STATISTICAL ANALYSIS OF THE NUMERIC FEATURES IN THE DATA SET.

Features	After 30 days	Before 30 days	Not Admitted
	μ ± σ	μ ± σ	μ ± σ
Time in hospital	4.49 ± 2.99	4.76 ± 3.02	4.25 ± 2.96
Num_lab procedure	43.84 ± 19.57	44.2 ± 19.27	42.4 ± 19.8
Num_procedure	1.25 ± 1.67	1.28 ± 1.63	1.41 ± 1.74
Num_medications	16.28 ± 7.62	16.9 ± 8.09	15.67 ± 8.42
Num_outpatients	0.49 ± 1.54	0.43 ± 1.30	0.27 ± 1.03
Num_emergency	0.28 ± 1.19	0.36 ± 1.37	0.11 ± 0.52
Num_inpatients	0.84 ± 1.39	1.22 ± 1.96	0.38 ± 0.86
Num_Diagnosis	7.65 ± 1.81	7.7 ± 1.77	7.22 ± 2.02

B. *Data Preprocessing*

Data preprocessing is one of key step in machine learning. It focuses on cleaning the data set, eliminating the features that have huge number of missing values. Figure 5 represents the number of missing values in the data set. Moreover, removing the weakly relevant features. During this phase 17 features were removed due to the below-mentioned reasons.

- In the data set there were some features with the huge number of missing values such as weight, payer code and medical specialty. The missing values percentage were 97%, 40% and 49% respectively, these features were removed from the data set.
- Additionally, features that contain only one distinct value for all the samples were removed, such as citoglipton medication and examide medication having only one value "No".
- Some of the features in the data set contain additional information of another feature, such as diagnosis features, containing three features, where the first one is considered primary, and the other two are considered secondary and extra information. Therefore, diag_2 and diag_3 was removed.

Some of the weakly relevant features were also removed, such as 10 medicine, because most of the patients either show no changes after taking them or have not used them. The features are repaglinide, nateglinide, chlorpropamide, acarbose, tolazamide, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone and metformin-pioglitazone, respectively.

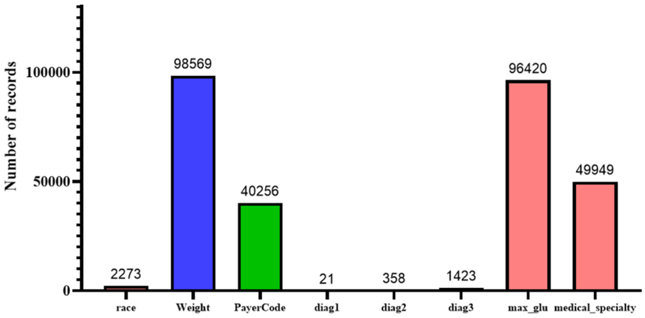

Fig. 5 Missing values per attribute in the data set

Two features, namely diag_1 and race, are the significant therefore data imputation technique was not applied. So, the samples with the missing diag_1 and race were eliminated in the study. Diag_1 indicates the first diagnosis. Similarly, the samples in which the gender contains "Unknown/Invalid were also removed. Moreover, for the feature discharge_disposition_id, we eliminate the samples that have values equal to 11, 13, 14, 19, 20, 21 because they are related to death or hospice.

Furthermore, data encoding scheme was applied to some of the features such as admission type, discharge type, admission source and diag_1 features into small number of categories. The age was converted into 10 categories, every 10 years is categorized in one number. Below mentioned steps were applied for data encoding.

- Converted the medicine feature values from four categories to binary category i.e. "No" is 0 while "Up", "Down" and "Steady" is replaced by 1.

- Replaced the change feature values such as “Ch” to 1 and “No” to 0.
- Replaced the gender feature, where “Male” is replaced by 1 and “Female” is replaced by 0.
- Replaced the diabetesMed feature whose values are “Yes” to 1 and “No” to 0.
- Replaced the A1Cresult feature whose values are “>7” or “>8” to 1 and “Norm” to 0 and “None” to -99.
- Replaced the max_glu_serum feature whose values are “>200” or “>300” to 1 and “Norm” to 0 and “None” to -99.
- Replaced the readmitted feature whose values are “>30” to 2 and “<30” to 1 and “No” to 0.
- Replaced the level1_diag1 and level2_diag1 features whose values contain “E” or “V” to 0 and “?” to -1.
- Replaced the level1_diag1 features values with a number from 0 to 8, based on some conditions.
- Replaced the level2_diag1 features values with a number from 0 to 22, based on some conditions.

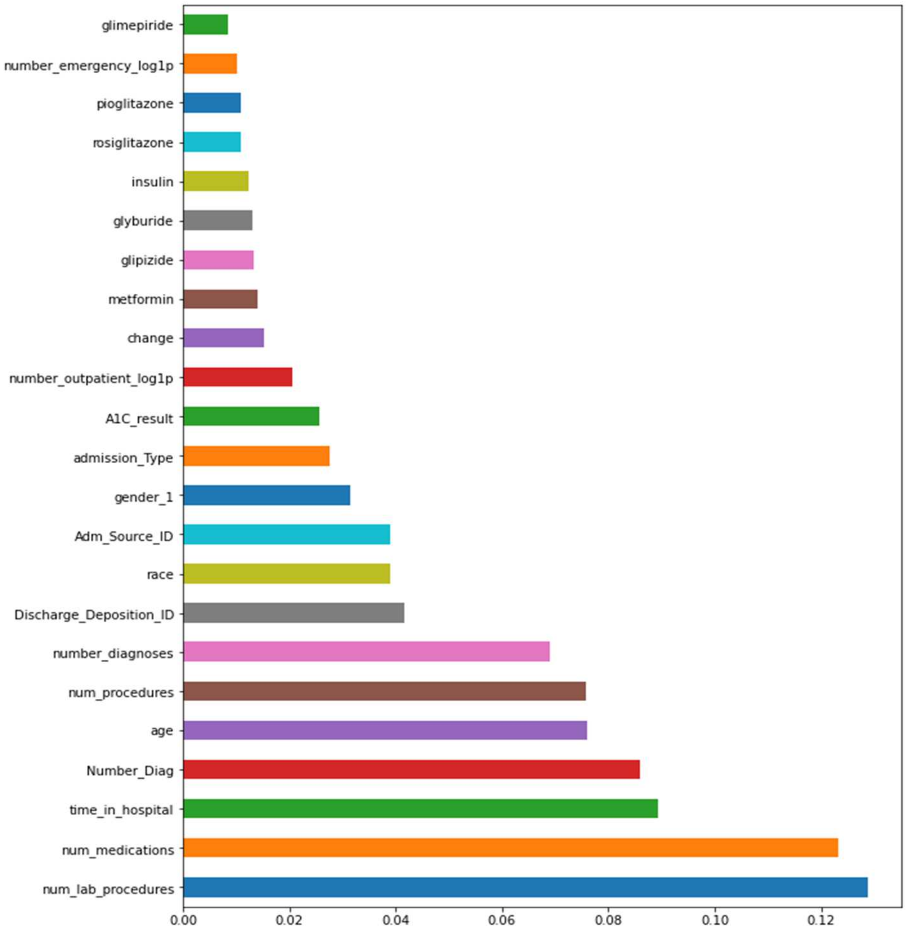

Fig. 6 Feature ranking using feature importance function for top 20 features selected by Firefly feature selection.

Furthermore, we converted features of nominal type to object type and categorical features to numerical. In this dataset, age is presented in categorical values for example 20 to 30. However, in the conversion process from categorical to numerical, we set the age values as the midpoint of the categorical values.

In the dataset, numerical features such as number of inpatient admissions and outpatient visits has high skew and high kurtosis. Log transformation can be used to deal with the skewed data [18]. Therefore, log transformation was performed to remove skewness and kurtosis that are above a threshold value of 2. And since log(0) is not defined, rules were performed to avoid bulk-removing records by computing log(feature) only if the 0’s in the feature are less than 2%. After performing log transformation on the numerical features, we used the standardization function to further qualify the data.

Subsequently, we used bioinspired Firefly feature selection to identify the highly significant features. Figure 6 contains ranking of 20 selected features using Firefly technique. The result showed that 10 medicine eliminated in the previous step has the lowest relevancy and was not selected using the feature selection. The ranking of the selected features was performed using the feature importance function in python.

C. *Firefly Feature Selection & Optimization*

Nature inspired meta heuristic firefly technique was applied for feature selection and optimization. It was proposed by Yang *et al.* [19]. This algorithm is inspired and incorporate the flashing light property of the firefly. The aim of the flashing property is to communicate with other flies and to attract the food. Furthermore, it can also be used for the protection purpose. The intensity of the flashing light is inversely proportional with the distance.

$$I \propto \frac{1}{r^2} \qquad (1)$$

In addition, to the distance the air also effects the power of the flashing light. For the optimization problem, the light intensity was represented interms of objective function with the aim to maximize the utilization. The objective function is represented as

$$f(x) = \{x_1, x_2, \dots, x_n\} \qquad (2)$$

The initial number of fireflies are

$$x_j = \{j = 1, 2, \dots, m\}$$

The intensity of light I_j I for the fly is represented by $f(x_j)$. The absorption of light is represented by γ. Firefly algorithm uses the light intensity to select the features. Highly relevant features are represented as the features with high intensity light. In our study we generated the flies equal to the total number of attributes in the data set n=33. Lower bound was set to 10 as the least number of features in the literature was 10 by Choudhury *et al.* [5]. The optimal number of selected features using firefly were 23.

Similarly, for optimization, several combination parameters were created for each classifier to find the optimal combination of classifiers. Table 2 contains the parameters tunning classifier optimization using Firefly.

TABLE II
FIREFLY PARAMETER TUNNING FOR CLASSIFIERS OPTIMIZATION

Parameter Name	Value
Total population of fireflies (N)	33
Light absorption coefficient (γ)	1.0
Alpha (α)	0.5
Beta (β)	0.2
Maximum Iteration	20

D. Classification Methods:

In this section classifiers such as Logistic regression, Decision tree, and Random forest will be discussed.

1) Decision Tree: Decision Tree is one of the widely used classification, regression, and feature selection technique. It is the hierarchical classifier consist of root, leaf, and branches. The execution of the information starts from the root node to the leaves. Attribute selection can be performed using entropy and information gain. Entropy indicates the measure of the randomness in the data while information gain measures the relative change in the entropy. This will result in highest quality decisions and more succinct by putting the biggest information gain attribute at the top of the tree.

The following equation represent how to calculate entropy, where H(S) is the entropy:

$$H(S) = \sum_{x \epsilon X} p(\mathrm{x}) \, \log_2 p(\mathrm{x}) \qquad (3)$$

Where $p(\mathrm{x})$ $is\ the\ probablity\ of\ x.X$ is the attributes in the data sets including the target class. The following equation represent how to calculate information gain, where IG(S,A) is the information gain:

$$IG(S, A) = H(S) - \sum_{i=0}^{n} p(\mathrm{x}) * \mathrm{H(S, X)} \qquad (4)$$

Where *H(S)* is the entropy of the target attribute, and *H(S,X)* is the entropy of the attribute with target class. The hyperparameters for decision tree classifier used in the model is shown in the Table 3.

TABLE III
OPTIMIZED HYPERPARAMETER FOR DECISION TREE USING FIREFLY

Parameter Name	Value
Criterion	Gini
Max_depth	8
Max_features	Log2

2) Random Forest (RF): RF is an ensemble-based machine learning algorithm using regression trees [20]. The algorithm recursively built trees to divide the attributes into regions. The objective of the recursively generating the trees is to reduce the variance. The iterations will be terminated when there is no further reduction in the variance. For testing was performed by analyzing the tree from root and test the outcomes. In addition, RF has mostly the same hyperparameters as the decision trees or bagging classifiers. RF adds more randomness to the model when the trees grow. Therefore, it searches for the appropriate feature between random set of features instead of looking for a most important feature when splitting a node. Thus, when using RF, a random set of features are considered when using the algorithm to split a node. Additionally, you can make the trees to be more random by using thresholds for every feature rather than looking for the best threshold.

RF uses two ensemble techniques such as bagging and random subspace. Bagging is the bootstrapping aggregation used simple random sampling with replacement for the imbalanced data sets. RF contains the below mentioned steps.

- Step 1: Initially the sample is selected using sample random sampling bootstrapping resampling method.
- Step 2: Random subspace will be used to select the attributes from the complete attribute set in the data set.
- Step 3: Construct a tree using ID3 algorithm using bootstrap samples.
- Step 4: Repeat process 1-3 to build the required number of trees to achieve the specified outcome.

Given a data set D consists of $X = \{x_1, x_2, \dots, x_n\}$ and a target class label y. This approach is a typical representation for the random forest. RF is represented by the following equation when training a family of classifiers.

$$D = \{(x_{i,}y_{i,})\}^n i = 1$$

The quality of the split is measured using Gini index. GI of the node is given by

$$GI(N) = \sum_{i \neq j} P(\omega_i) \; P(\omega_j)$$

where $P(\omega_i)$ is the ratio of records in the data set for the class category i. The hyperparameter for Random Forest classifier using Firefly is represented in Table 4.

TABLE IV
OPTIMIZED HYPERPARAMETER FOR RANDOM FOREST USING FIREFLY

Parameter Name	Value
N_estimators	80
Max_depth	18
Min_samples_split	18
Max_features	auto

3) Logistic Regression: Logistic Regression is used for classification and regression, a statistical model that uses maximum-likelihood ratio concept. Logistic regression also known as logit model, that use sigmoid function. A model that enumerates a proper function of the fitted probability of the event is a linear function of the observed values of the target variables. The benefit of logistic regression is that it uses a simple probabilistic formula of classification. The drawback of linear regression model is that it cannot deal with non-linear problem.

The training process depends on choosing the parameters, the parameters should define the function that maximizes the posteriori likelihood function. For example, let C is the number of classes identified as $C \in \{1,2,\dots,C\}$, and let X is the feature vector of length n. Thus, the given equation 5 below represents the probability that X belongs to one of the C classes. The $\beta_1, \beta_2, \dots, \beta_k$ represents the parameter vectors that define regression coefficients, and $\langle \beta_k, X \rangle$ is the vectors inner product.

$$\mathrm{P}(Y = k|x) = \frac{e^{\langle \beta_k, X \rangle}}{\sum_{i=1}^{K} e^{\langle \beta_i, X \rangle}}\ for\ k = 1,2,\dots,k, \qquad (5)$$

Where p is the probability of success. K represents the class label. k | x represents that x belongs to the k class label. From the training process the coefficients βk can be obtained. Equation 6 and 7 will be used to predict the outcome of feature vector X.

$$k^* \in argmaxPr(Y = k|X), k \in \{1,2,\dots,K\} \qquad (6)$$

$$k^* \in argmax\langle \beta_x, X \rangle, k \in \{1,2,\dots,K\} \qquad (7)$$

The hyperparameter for LR using Firefly optimization is shown in Table 5.

TABLE V
OPTIMIZED HYPERPARAMETER FOR LOGISTIC REGRESSION USING FIREFLY

Parameter Name	Value
C	2.0
Dual	False
Max_iterations	100

E. Evaluation Metrics

To investigate the performance of the proposed model, several standard evaluation parameters will be used such as accuracy, precision, recall and Area Under the Curve (AUC). Below are the equations

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

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

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

$$AUC = \sum_{i \in (TP+TN+FP+FN)} \frac{(TP_i + TP_{i-1}).FP_i + TFP_{i-1}}{2}$$

Where the acronym TP represents true positive, TN true negative, FP false positive, and FN false negative, respectively.

III. RESULTS AND DISCUSSION

Models were implemented in python 3.8.5 version using Jupyter notebook 6.0.3. The sklearn_nature_inspired_ algorithms library was used for the optimization of the classifiers using Firefly. The NatureInspiredSearchCV technique was applied for each classifier optimization. Firefly was also used for the feature selection. The EvoPreprocess library was employed for the feature selection, which is using NiaPy library as a backend for the Natural inspired algorithms.

Experiments were conducted using all 50 features and subset of features. Subsequently, feature importance technique was used to ranks the features based on their level of importance. Three set of experiments were conducted such as.

- The first set contains all 50 features.
- The second set contains initial 33 features after the preprocessing, which is the same as the first set, excluding 17 features.
- And the third set is the features selected using Firefly feature selection. The number of features selected by Firefly feature selection was 23.

K-fold cross validation data sampling technique was used, with K=10. Data imbalance may lead to model overfitting, SMOTE oversampling technique was applied to alleviate the data imbalance. SMOTE technique was applied on training data. It is a data augmentation technique for generating the minority class records [21]. The generation of the minority class instances can be represented by equation.

$$Y^{'} = Y^i + (Y^j - Y^j).* \gamma$$

Where Y^i represents the minority, class records. While Y^j is the randomly selected minority class. γ represents vector combination of randomly generated number between 0 and 1 [22].

The random number for generator was set to 2, which means it seeded a new RandomState object. Finally, each training data from three set of features was input in all the three classifiers such as LR, DT, and RF, respectively. Experiments were conducted with and without using SMOTE. The number of records before and after 30 days readmission is increased to 54861 records after the SMOTE. Figure 7 represents the number of records per class label with and without SMOTE.

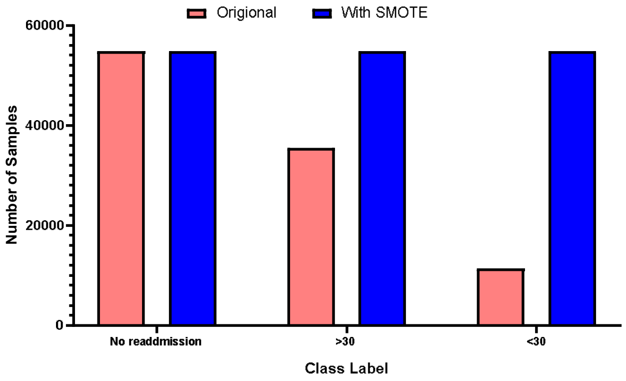

Fig. 7 Number of records per class label with and without SMOTE

As discussed previously, three set of experiments were conducted using feature sets i.e. all features, selected after preprocessing (33) and selected features using firefly (23), respectively. The models were trained and tested using all the three-feature set. However, we obtained the highest results from the third set. Table 6 presents the results of the classifiers using various feature sets with original and SMOTE data.

TABLE VI
EXPERIMENTAL RESULTS OF CLASSIFIERS USING VARIOUS FEATURE SET AND WITH ORIGINAL AND SMOTE DATA

Feature set	No. of features	Classifier	SMOTE	Acc	Prec	Rec	AUC
All Features	50	RF	With	0.94	0.98	0.9	0.94
		DT		0.90	0.90	0.90	0.90
		LR		0.66	0.72	0.65	0.68
		RF	With out	0.92	0.99	0.92	0.96
		DT		0.87	0.93	0.92	0.93
		LR		0.91	0.97	0.91	0.96
Selected Features after pre-processin g	33	RF	With	0.92	0.94	0.90	0.92
		DT		0.90	0.93	0.89	0.91
		LR		0.66	0.66	0.66	0.66
		RF	With out	0.92	0.98	0.91	0.93
		DT		0.87	0.95	0.92	0.93
		LR		0.91	0.98	0.91	0.95
Selected features using Firefly	23	RF	With	**0.99**	**0.99**	**0.94**	**0.98**
		DT		0.90	0.93	0.88	0.90
		LR		0.85	0.80	0.87	0.83
		RF	With out	0.91	0.98	0.91	0.95
		DT		0.88	0.95	0.92	0.93
		LR		0.91	0.98	0.91	0.95

Furthermore, Table 7 contains the comparison of the proposed model with the benchmark studies. The selection criteria for the benchmark depends on the data set used in the study. We attempt to improve the outcomes achieved by the previous studies using machine learning techniques for diabetic readmission. Furthermore, the study attempts to identify the significant features for readmission of diabetic patients.

TABLE VII
COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH THE BENCHMARK STUDIES

Reference	**Year**	**Technique**	**Data Augmentation**	**No. of Feature**	**Findings**	
					Acc	AUC
[5]	2018	GBM	No	10	0.96	-
[10]	2018	SVM	Yes	32	0.81	-
[11]	2019	Hybrid Ensemble	No	12	0.79	0.81
[12]	2019	NB	No	18	0.56	0.64
[13]	2018	CNN	Yes	43	0.92	0.95
[14]	2019	RF	No	45	0.99	0.99
[15]	2020	DNN	Yes	35	0.95	0.97
[16]	2020	RF	No	-	0.83	0.84
Our study	2020	RF	Yes	33	0.99	0.98

As shown in the Table 7, the proposed study outperformed most of the studies in the benchmark. The current study achieved the highest accuracy of 0.99 and AUC of 0.98. Furthermore, our study achieved the highest precision of 0.99 and the precision of 0.99 similar to the precision achieved by Ramirez *et al. study* [14]. However, [14] study have shown higher recall and AUC when compared to our study. In the [14], three classes were converted into binary class. Furthermore, the number of features were 45, however, in the current study, it is 33 features.

Ghazo [11] study has achieved an accuracy of 0.73 and AUC of 0.94. However, the experiments were performed in their study for the binary class. Sarthak *et al.* [15] achieved similar AUC as compared to the proposed study. However, the current study was performed for multiclass while Sarthak study convert the data set into binary class. In conclusion, the proposed model outperformed most of the benchmark studies except Ramirez *et al.* interms of recall and AUC.

IV. CONCLUSION

In many countries, huge costs associated with hospital readmissions can affect hospitals reputation and burden on the country. Therefore, early prediction of diabetes patient's hospital readmission is of key importance. In current study, Random Forest outperformed the other models using feature set containing medication, demographic and service utilization attributes. Moreover, the logistic regression results and the decision tree models performed in this work have shown better results on several performance metrices compared to other studies that used the same models. Despite of significant outcomes achieved by the study there are some limitation of the proposed study. The data set contains huge number of missing values for some attributes. Furthermore, the data set suffers from huge imbalance. Since diabetic hospital readmission is considered as a critical healthcare measure and early prediction can reduce the cost. Furthermore, it will help the management to identify areas that need further attention. Hence, it is recommended to investigate other machine models and multiple data sets that may achieve better results and identify significant feature for early prediction of diabetic patients' readmissions.

REFERENCES

[1] "Diabetes-WHO." https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed Dec. 11, 2020).

[2] S. Bolourani *et al.*, "Using machine learning to predict early readmission following esophagectomy," *J. Thorac. Cardiovasc. Surg.*, 2020, doi: 10.1016/j.jtcvs.2020.04.172.

[3] P. Wolff, M. Graña, A. R. Sebastián, and M. B. Yarza, "Machine Learning Readmission Risk Modeling : A Pediatric Case Study," vol. 2019, 2019.

[4] Y. Tseng, H. Wang, T. Lin, J. Lu, C. Hsieh, and C. Liao, "Development of a Machine Learning Model for Survival Risk Stratification of Patients With Advanced Oral Cancer," vol. 3, no. 8, 2020, doi: 10.1001/jamanetworkopen.2020.11768.

[5] A. Choudhury and D. C. M. Greene, "Evaluating Patient Readmission Risk: A Predictive Analytics Approach," *Am. J. Eng. Appl. Sci.*, vol. 11, no. 4, pp. 1320–1331, 2018, doi: 10.3844/ajeassp.2018.1320.1331.

[6] F. Alshakhs, H. Alharthi, N. Aslam, I. U. Khan, and M. Elasheri, "Predicting postoperative length of stay for isolated coronary artery bypass graft patients using machine learning," *Int. J. Gen. Med.*, vol. 13, 2020, doi: 10.2147/IJGM.S250334.

[7] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying Diabetic Patients with High Risk of Readmission." arXiv preprint arXiv:1602.04257.

[8] R. Duggal, S. Shukla, S. Chandra, and B. Shukla, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. December, pp. 469–476, 2016, doi: 10.1007/s13410-016-0495-4.

[9] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in India," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 4, pp. 519–528, 2016, doi: 10.1007/s13410-016-0511-8.

[10] S. Cui, D. Wang, Y. Wang, P. W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Comput. Methods Programs Biomed.*, vol. 166, pp. 123–135, 2018, doi: 10.1016/j.cmpb.2018.10.012.

[11] Ghazo, Esraa. Prediction of Diabetic Patient Readmission Using Hybrid Ensemble Learning Diss. State University of New York at Binghamton, 2019.

[12] M. Alloghani *et al.*, “Implementation of machine learning algorithms to create diabetic patient re-admission profiles,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. Suppl 9, pp. 1–16, 2019, doi: 10.1186/s12911-019-0990-x.

[13] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, “Predicting hospital readmission among diabetics using deep learning,” *Procedia Comput. Sci.*, vol. 141, pp. 484–489, 2018, doi: 10.1016/j.procs.2018.10.138.

[14] J. C. Ramírez and D. Herrera, “Prediction of Diabetic Patient Readmission Using Machine Learning,” *Commun. Comput. Inf. Sci.*, vol. 1096 CCIS, pp. 78–88, 2019, doi: 10.1007/978-3-030-36211-9_7.

[15] Sarthak, S. Shukla, and S. Prakash Tripathi, “Embpred30: Assessing 30-days readmission for diabetic patients using categorical embeddings,” *Adv. Intell. Syst. Comput.*, vol. 1168, pp. 81–90, 2021, doi: 10.1007/978-981-15-5345-5_7.

[16] C. I. Ossai and N. Wickramasinghe, “Intelligent therapeutic decision support for 30 days readmission of diabetic patients with different comorbidities,” *J. Biomed. Inform.*, vol. 107, no. June, p. 103486, 2020, doi: 10.1016/j.jbi.2020.103486.

[17] B. Strack *et al.*, “Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records,” *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/781670.

[18] C. Feng *et al.*, “Log-transformation and its implications for data analysis,” *Shanghai Arch. Psychiatry*, vol. 26, no. 2, pp. 105–109, 2014, doi: 10.3969/j.issn.1002-0829.2014.02.

[19] X. Yang, Nature-Inspired Metaheuristic Algorithms Second Edition, vol. 4, no. C. 2010.

[20] Y. L. Pavlov, “Random forests,” *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780367816377-11.

[21] W. P. K. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, “Smote: synthetic minor- ity over-sampling technique,” *J. Mach. Learn. Res.*, vol. 16, pp. 321–357, 2002.

[22] T. Zhu, Y. Lin, and Y. Liu, “Synthetic minority oversampling technique for multiclass imbalance problems,” *Pattern Recognit.*, vol. 72, pp. 327–340, 2017, doi: 10.1016/j.patcog.2017.07.024.