

# Hospital Readmission Prediction using Machine Learning Techniques

## A Comparative Study

Samah Alajmani<sup>1</sup>

Computer Science Department  
Faculty of Computing & Information Technology  
King Abdulaziz University  
Jeddah, Saudi Arabia

Hanan Elazhary<sup>2</sup>

College of Computer Science & Engineering  
University of Jeddah, Jeddah, Saudi Arabia  
Computers & Systems Department  
Electronics Research Institute, Cairo, Egypt

**Abstract**—One of the most critical problems in healthcare is predicting the likelihood of hospital readmission in case of chronic diseases such as diabetes to be able to allocate necessary resources such as beds, rooms, specialists, and medical staff, for an acceptable quality of service. Unfortunately relatively few research studies in the literature attempted to tackle this problem; the majority of the research studies are concerned with predicting the likelihood of the diseases themselves. Numerous machine learning techniques are suitable for prediction. Nevertheless, there is also shortage in adequate comparative studies that specify the most suitable techniques for the prediction process. Towards this goal, this paper presents a comparative study among five common techniques in the literature for predicting the likelihood of hospital readmission in case of diabetic patients. Those techniques are logistic regression (LR) analysis, multi-layer perceptron (MLP), Naïve Bayesian (NB) classifier, decision tree, and support vector machine (SVM). The comparative study is based on realistic data gathered from a number of hospitals in the United States. The comparative study revealed that SVM showed best performance, while the NB classifier and LR analysis were the worst.

**Keywords**—Decision tree; hospital readmission; logistic regression; machine learning; multi-layer perceptron; Naïve Bayesian classifier; support vector machines

### I. INTRODUCTION

Nowadays, numerous chronic diseases, such as diabetes, are widespread in the world; and the number of patients is increasing continuously. The estimated number of diabetic adults in 2014 was 422 million versus 108 million in 1980 [1]. Such patients visit hospitals frequently, requiring continuous preparation for ensuring the availability of required resources including hospital beds, rooms, and enough medical staff for an acceptable quality of service. Accordingly, predicting the likelihood of readmission of a given patient is of ultimate importance. In fact readmission during a one month period (30 days) of discharge indicates "a high-priority healthcare quality measure" and the goal is to address this problem [2].

Machine learning, which is one of the most important branches of artificial intelligence, provides methods and techniques for learning from experience [3]. Researchers often use it for complex statistical analysis tasks [4]. It is a wide multidisciplinary domain which is based on numerous

disciplines including, but not limited to, data processing, statistics, algebra, knowledge analytics, information theory, control theory, biology, statistics, cognitive science, philosophy, and complexity of computations. This field plays an important role in term of discovering valuable knowledge from databases which could contain records of supply maintenance, medical records, financial transactions, applications of loans, etc. [5].

As indicated in Fig. 1, machine learning techniques can be broadly classified into three main categories [3]. Supervised learning techniques involve learning from training data, guided by the data scientist. There are two basic types of learning missions: classification and regression. Models of classification attempt to predict distinguished classes, such as blood groups, while models of regression prognosticate numerical values [3]. In unsupervised learning, on the other hand, the system could attempt to find hidden data patterns, associations among features or variables, or data trends [3], [4]. The main objective of unsupervised learning is the ability to specify hidden structures or data distributions without being subject to supervision or the prior categorization of the training data [6]. Finally, in reinforcement learning the system attempts to learn through interactions (trial and error) with a dynamic environment. During this learning mode, the computer program provides access to a dynamic environment in order to perform a specific objective. It is worth noting that in this case, the system does not have prior knowledge regarding the environment's behavior, and the only way to figure it out is through trial and error [3], [7], [8].

According to Kaelbling et al., the term healthcare informatics refers to the combination between machine learning and healthcare with the purpose of specifying interest patterns [9]. In addition to this, it has the potential for establishing a good relationship between patients and doctors, and minimizing the increasing cost of healthcare [10]. The goal of this paper is to apply machine learning techniques, and specifically prediction techniques, for predicting the likelihood of readmission of patients to hospitals. This problem hasn't been adequately addressed in the literature. In fact most research efforts are oriented towards prediction of diseases. Machine learning includes numerous analytic techniques for prediction and the literature lacks adequate comparative studies

that assist in selecting a suitable technique for this purpose. Our research is based on a large data set collected by numerous United States hospitals [11], [12]. In short, this paper has two main contributions as follows:

- Analyzing five most common machine learning techniques for prediction and providing a comparative study among them.
- Addressing the problem of patient readmission to hospitals, since it has been rarely addressed by researchers.

Organization of the rest of the paper is as follows: First, we present background about the machine learning techniques considered in this research. This is followed by related work to highlight the contributions of the paper. We then present our methodology and discuss the results of the experiments. Finally, we sum up this work via a conclusion and discussion of possible future work.

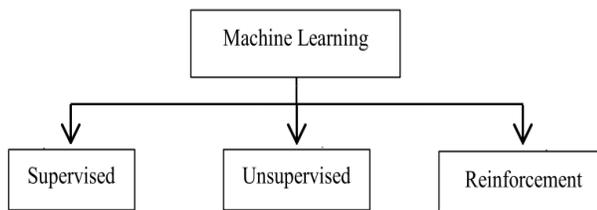


Fig. 1. Classification of Machine Learning Techniques.

## II. BACKGROUND

This section discusses the five basic machine learning techniques employed in this research study.

### A. Logistic Regression Analysis

Regression is a statistical notion that can be used to identify the relationship weight between one variable called the dependent variable and a group of other changeable variables denoted as the independent variables. Logistic regression (LR) is a non-linear regression model, used to estimate the likelihood that an event will occur as a function of others [13].

### B. Artificial Neural Network

An Artificial Neural Network (ANN) is a computational model which attempts to emulate the human brain parallel processing nature. An ANN is a network of strongly interconnected processing elements (neurons), which operate in parallel [14] inspired by the biological nervous systems [15]. ANNs are broadly used in many researches because they are capable of modeling non-linear systems, where relationships among variables are either unknown or quite complicated [14]. An example of an ANN is the Multi-Layer Perceptron (MLP), which is typically formed of three layers of neurons (input layer, output layer, and hidden layer) and its neurons use non-linear functions for data processing [16].

### C. Naïve Bayesian Classifier

Naïve Bayesian (NB) classifier relies on applying Bayes' theorem to estimate the most probable membership of a given event in one of a set of possible classes. It is described as being

naïve, since it assumes independence among variables used in the classification process [15], [17], [18].

### D. Support Vector Machine

Support vector machines (SVMs) are supervised learning models, which can be applied for classification analysis and regression analysis. They have been proposed by Vapnik in 1995. They can perform both linear and non-linear classification tasks [5], [12], [17], [19].

### E. Decision Tree

Decision trees are one of the most famous techniques in machine learning. A decision tree relies on classification by using attribute values for making decisions. In general, a decision tree is a group of nodes, leaves, a root and branches [20]. Many algorithms have been proposed in the literature for implementing decision trees. One important algorithm is CART (Classification and Regression Tree). It is used for dealing with continuous and categorical variables [8], [21].

## III. RELATED WORK

Many researchers attempted to use machine learning techniques in healthcare problems other than hospital readmission likelihood prediction. For example, Arun and Sittidech used K-Nearest Neighbor (KNN), NB, and decision trees with boosting, bagging, and ensemble learning in diabetes classification. Their experiments confirmed that the highest accuracy is obtained by applying bagging with decision trees [22]. On the other hand, Perveen et al. attempted to improve the performance of such algorithms using AdaBoost. The evaluation of experimental outcomes showed that AdaBoost had better performance in comparison to bagging [23]. Orabi et al. [24] suggested integrating regression with randomization for predicting diabetes cases according to age, with an accuracy of 84%. Other researchers proposed building a predictive model using three machine learning techniques, which are random forests (RFs), LR, and SVMs; for predicting diabetes in Indians females, in addition to the factors causing diabetes. Their comparative study concluded that RFs had the best performance among the others [25].

Relatively few research studies addressed the problem of hospital readmission likelihood prediction. For example, Strack et al. used statistical models for this purpose [12]. Other researchers focused on comparing different machine learning techniques for addressing this problem. For example, Kerexeta [26] proposed two approaches. In the first, they combined supervised and unsupervised classification techniques, while in the latter, they combined NB and decision trees. They showed that the former approach had a better performance in comparison to the latter in terms of readmission prediction.

To sum up, relatively few research efforts in healthcare are concerned with the problem of prediction of hospital readmission likelihood. Additionally, there is a shortage of adequate comparative studies for comparing machine learning techniques used for prediction. Hence, this paper attempts to tackle those two problems by comparing five common machine learning techniques for tackling the problem of hospital readmission likelihood prediction based on real data.

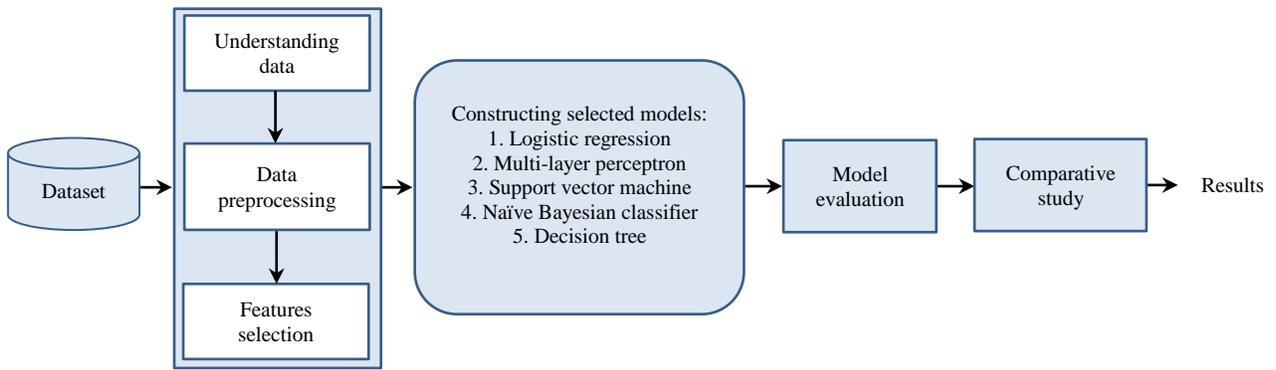


Fig. 2. Proposed Methodology.

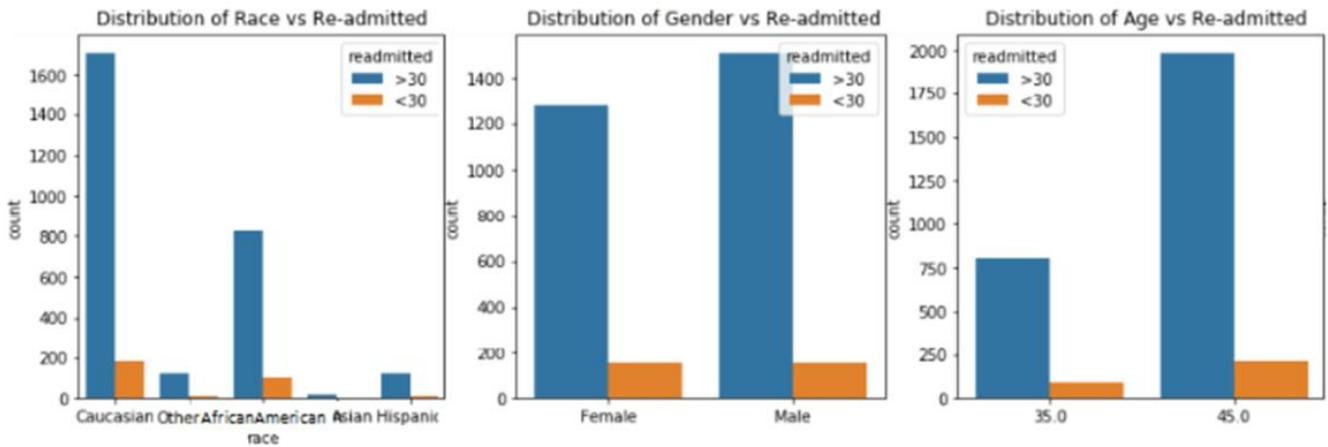


Fig. 3. Analysis of Variables.

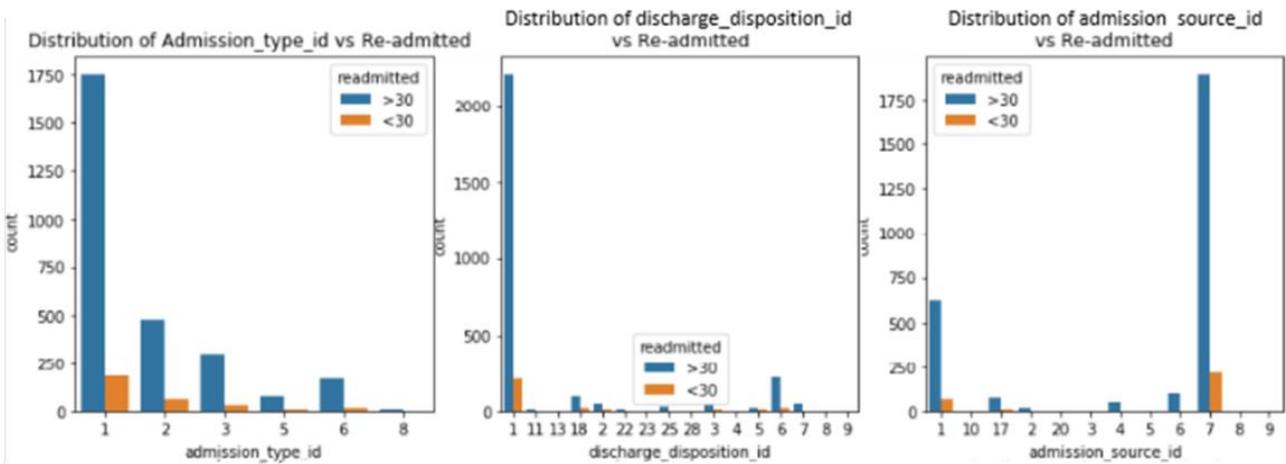


Fig. 4. Analysis of Variables (cont.).

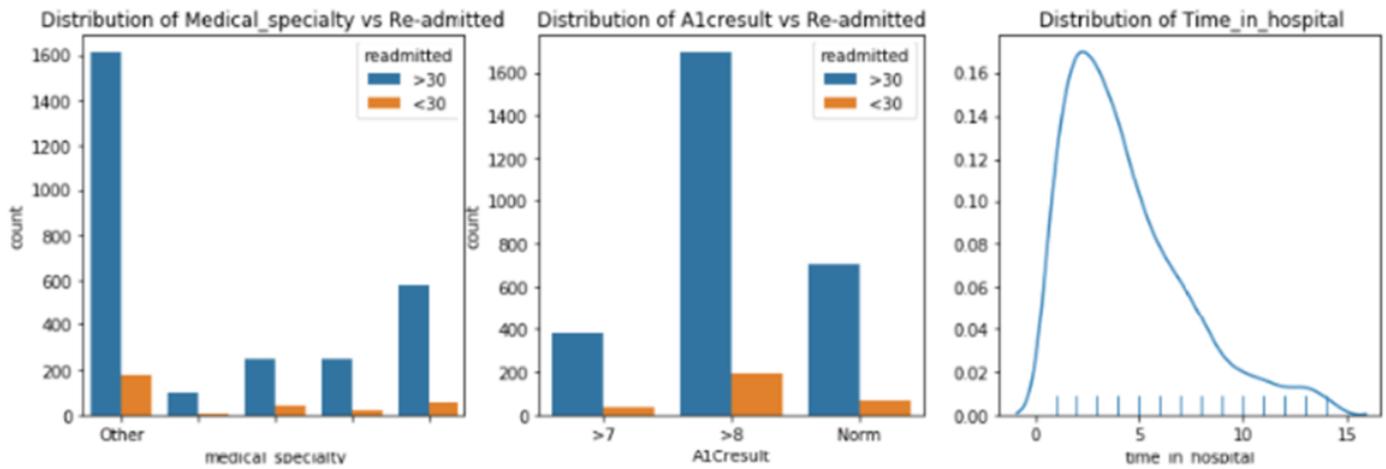


Fig. 5. Analysis of Variables (Cont.).

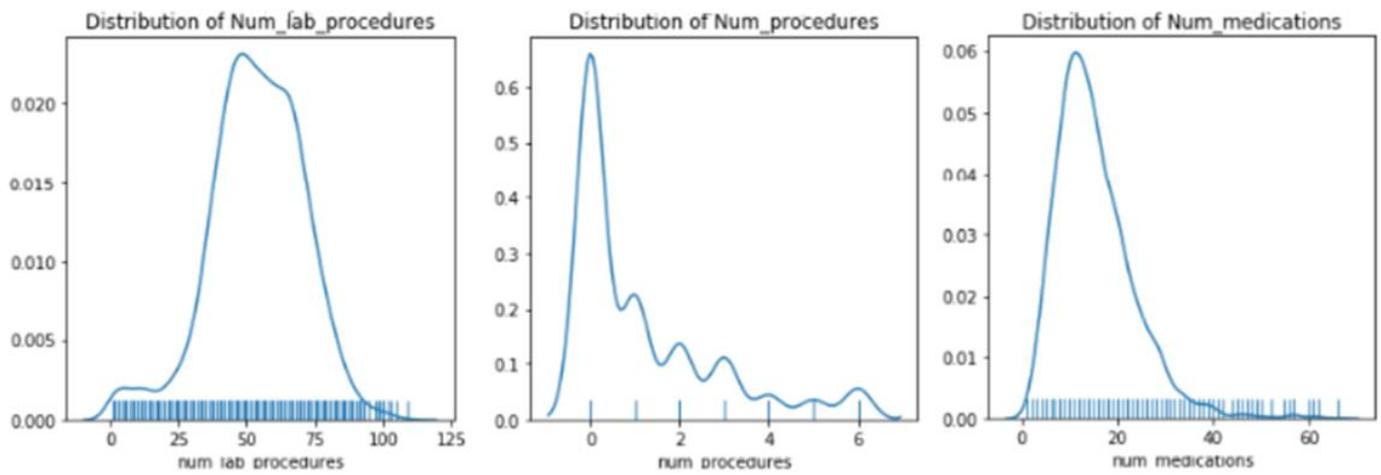


Fig. 6. Analysis of Variables (Cont.).

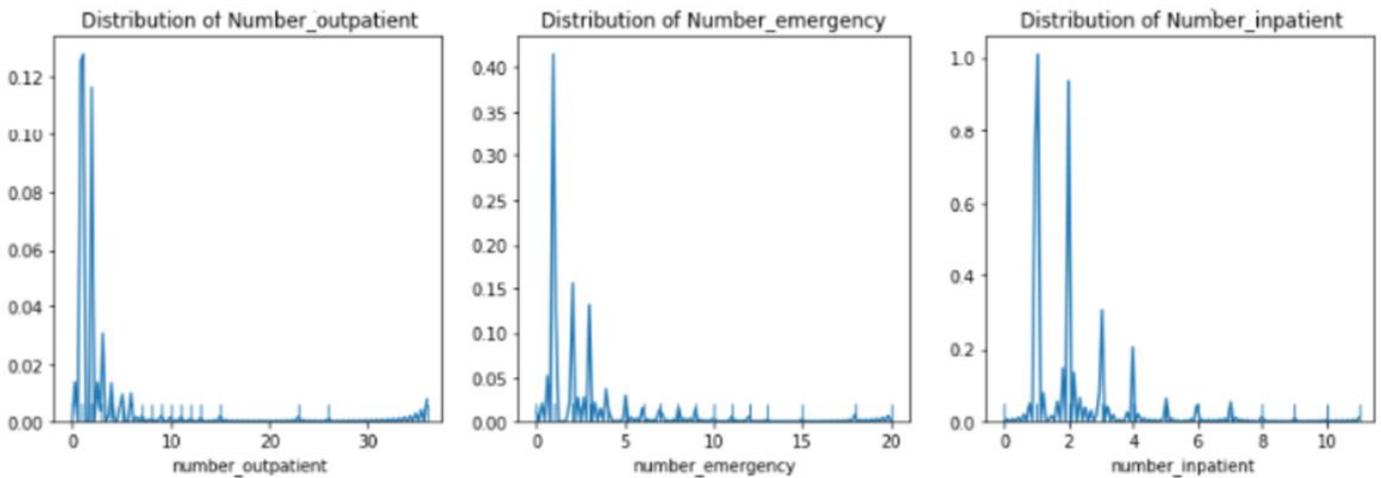


Fig. 7. Analysis of Variables (Cont.).

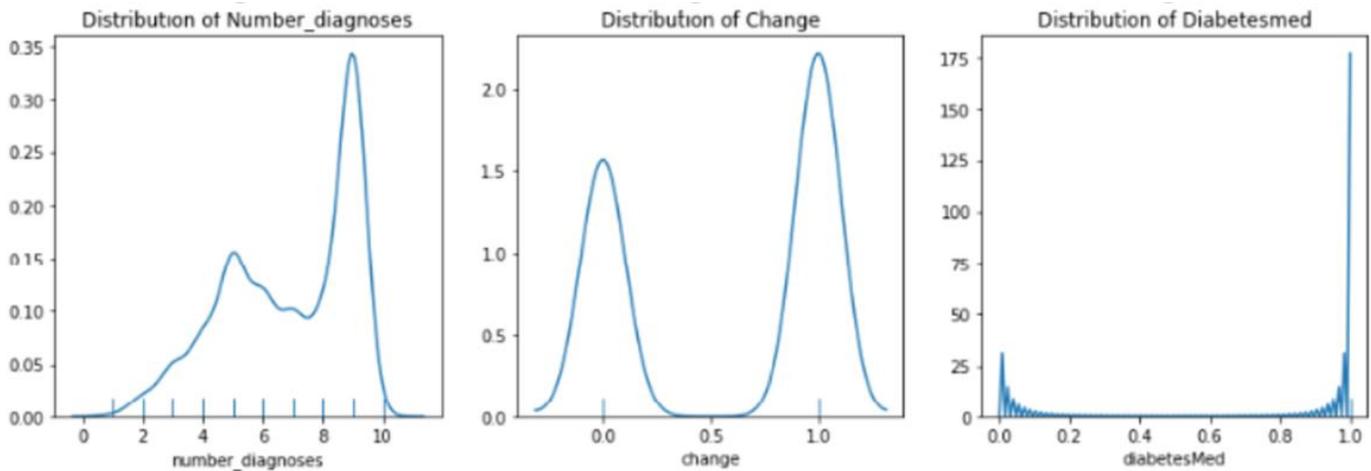


Fig. 8. Analysis of Variables (Cont).

#### IV. METHODOLOGY

Before starting the comparative study, it is important to understand the data, perform preprocessing if necessary, and select features appropriate for the experiments as depicted in Fig. 2. Those tasks are explained below. It is worth noting that all the experiments were conducted using Python.

##### A. Data Preparation

1) *Understanding data*: In this study, we exploited a sample of a diabetes patients’ dataset, which has been extracted from many hospitals in the United States [11], [12]. This dataset includes 3090 instances in the age range of 30-50 and with 18 attributes. Table I depicts the variables of the dataset together with their descriptions. The scientific meanings of those variables are beyond the scope of this paper. Fig. 3 through 8 depict the distribution of those features.

2) *Data pre-processing*: This is a very important stage which includes data transformation and cleaning. In data transformation, some variables were transformed from categorical to binary (0/1) such as (Change, DM, G, and A). Some other variables were transformed from integer to string such as AS, DI, and AS. In data cleaning, some values of categorical data were missing and had to be accounted for. For this purpose, we employed imputation (substitution) via the mode of the categorical data.

3) *Feature Selection*: In this step, we perform feature selection for dimensionality reduction. In other words, we select the most relevant features. In this study, towards this goal, we assessed the impact of variables on our target. This helped us eliminate variables with low importance. Features which have high influence on accuracy are the most important [27]. We used the Gradient Boosting technique [28] for categorical features. Table II demonstrates the average

weights of the variables. We then utilized a threshold of 0.014 to obtain our feature set [29], [30]. Accordingly, the features A, AS, and DM were rejected since their weights were lower than 0.014. All the other features depicted in Fig. 9 were selected.

TABLE I. DIABETES PATIENTS’ DATA

Variable	Variables Abbreviation	Data type
Race	R	Categorical
Gender	G	Categorical
Age	A	Categorical
Admission type Id	AT	Integer
Discharge disposition Id	DI	Integer
Admission source Id	AS	Integer
Medical specialty	MS	Categorical
A1Cresult	A1Cresult	Categorical
Time in hospital	TH	Integer
Number of lab procedures	NL	Integer
Number of procedures	NP	Integer
Number of medications	NM	Integer
Number of outpatient	NO	Integer
Number of emergency	NE	Integer
Number of inpatient	NI	Integer
Number of diagnosis	ND	Integer
Change	Change	Categorical
DiabetesMed	DM	Categorical
DM	0.008867	Rejected

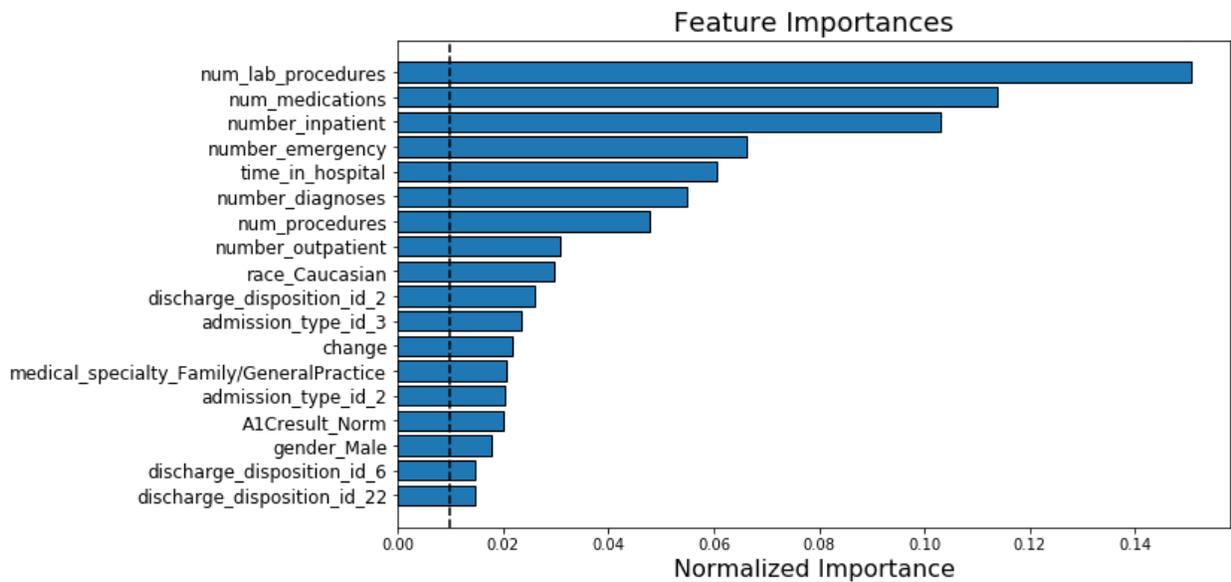


Fig. 9. Importance of Variables.

TABLE II. FEATURES IMPORTANCE

Variable	Importance (average)	Decision
R	0.029016	Confirmed
G	0.020294	Confirmed
A	0.010165	Rejected
AT	0.023854	Confirmed
DI	0.027554	Confirmed
AS	0.008961	Rejected
MS	0.019117	Confirmed
AlCresult	0.020177	Confirmed
TH	0.062025	Confirmed
NL	0.149317	Confirmed
NP	0.046521	Confirmed
NM	0.111058	Confirmed
NO	0.030696	Confirmed
NE	0.066718	Confirmed
NI	0.104099	Confirmed
ND	0.055811	Confirmed
Change	0.023027	Confirmed

### B. Constructing Machine Learning Techniques Models

In this comparative study, the selected models included one output/target with two values (True or False) regarding hospital readmission during a period of 30 days. In other words, the value of the readmission parameter is true if readmission is done during a period of 30 days. Otherwise, in case of no readmission or in case readmission is done after 30 days, its value is false. The set of drivers for the prediction was comprised of the selected features as discussed above. The training dataset and the testing dataset were selected randomly. Additionally, 10-fold cross validation was applied by selecting 40% of the data for testing and the rest for training. The settings of the various models are discussed below.

1) *Logistic regression*: This model was built by importing the Logistic regression module and using it to generate the classifier. Grid search was employed to detect the optimal accuracy and the best hyper-parameters.

2) *Support vector machine*: Kernel SVM was trained using the training set. A support vector classification (SVC) task was used. In this technique, there are many parameters such as C, kernel, and gamma; where C represents the error term penalty parameter, kernel determines the kernel type which can be utilized in the algorithm (in our case it is 'rbf'), and gamma indicates the coefficient of the kernel, such that a high value of gamma attempts to completely fit the set of training data. Grid search was employed to determine the optimal parameters and accuracy. Table III illustrates that the optimal accuracy for C is 10. On the other hand, the optimal accuracy for gamma was 0.3.

TABLE III. SVM ACCURACY

C	Accuracy
0.1	0.8169582
1.00	0.9102736
10.00	0.9246298

3) *Decision tree*: This model was generated using the 'gini' function to evaluate the split quality of the tree. In our study, the min\_samples\_split = 30 is the minimal number of samples needed for splitting an internal node, and max\_depth is the maximal tree depth. Grid search was conducted and the best accuracy for max\_depth was 15 as depicted in Table IV.

TABLE IV. DECISION TREE ACCURACY

max_depth	Accuracy
5	0.8326603
10	0.8627187
15	0.8788694

TABLE V. MLP WEIGHT MATRIX RESULTS

	Hidden2.1	Hidden2.2	Hidden2.3	Hidden2.4	Hidden2.5	Readmitted<30	Readmitted>30
num_lab_procedures	-0.0021	0.0263	-0.2052	0.2031	-0.0536	0.0000	0.0000
num_medications	0.0055	-0.1040	2.1573	1.3116	0.0572	0.0000	0.0000
number_inpatient	-0.0141	-0.0330	-3.7337	3.9118	0.0097	0.0000	0.0000
num_emergency	-0.0105	-0.0008	0.1937	-0.2044	0.0118	0.0000	0.0000
time_in_hospital	0.0113	0.0307	-1.0520	1.3083	-0.0158	0.0000	0.0000
number_diagnosis	-0.0064	0.0004	-0.8371	-0.0016	0.0135	0.0000	0.0000
num_procedures	0.0005	-0.0362	1.3875	0.0360	0.0034	0.0000	0.0000
number_oupatient	0.0096	-0.0106	0.0029	0.7529	0.0561	0.0000	0.0000
race_caucasian	-0.0238	-0.1076	-0.0529	-1.2512	4.5722	0.0000	0.0000
discharge_disposition_id_2	-0.7015	-3.2359	3.0491	1.3420	-0.8454	0.0000	0.0000
admission_type_id_3	-0.0043	0.0155	2.0813	1.5240	-2.4905	0.0000	0.0000
change	0.0019	0.0294	-0.5155	-0.9567	-0.0422	0.0000	0.0000
medical_speciality_Family/GeneralPractice	-0.0643	-2.8393	5.2166	0.8710	-0.0294	0.0000	0.0000
admission_type_id_2	0.0163	0.0751	0.1518	0.1636	-2.2760	0.0000	0.0000
A1Cresult_Norm	4.1982	-5.7995	-1.5765	-1.0361	-0.0567	0.0000	0.0000
gender_Male	-0.0045	0.0313	-0.7745	-1.3289	-0.0169	0.0000	0.0000
discharge_disposition_id_6	0.0011	-0.0197	0.4592	0.6655	0.0107	0.0000	0.0000
discharge_disposition_id_22	-0.0233	-2.7176	-0.0397	2.7966	2.1024	0.0000	0.0000
Hidden2.1	0.0000	0.0000	0.0000	0.0000	0.0000	-2.4213	0.3857
Hidden2.2	0.0000	0.0000	0.0000	0.0000	0.0000	-1.0671	1.4849
Hidden2.3	0.0000	0.0000	0.0000	0.0000	0.0000	1.0464	1.0589
Hidden2.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0924	0.6212
Hidden2.5	0.0000	0.0000	0.0000	0.0000	0.0000	-4.5618	-0.0239

4) *Naïve bayesian classifier*: A NB model was created using Gaussian Naive Bayes, which assumes that the attributes follow a natural distribution.

5) *Multi-Layer perceptron*: We built a MLP network using 18 inputs. The number of neurons in a hidden layer was 5. The function of the neurons was stochastic gradient descent. The maximum number of iterations was 300, and the two outputs were (readmitted < 30 and readmitted > 30). Table V illustrates that result of MLP weight matrix after training.

## V. RESULTS AND DISCUSSION

This work utilized various performance measures to compare the studied techniques [31]. Specifically, we relied on accuracy, recall, precision, and F1 scores for this purpose. Those parameters are defined in terms of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as indicated in equations (1) through (4). TPs are cases in which we predicted yes (they will be readmitted in a month period), and they were really readmitted. TNs are cases in which we predicted no, and they were not readmitted. On the other hand, FPs are cases in which we predicted yes, but they were not actually readmitted; Type I error. Finally, FNs are cases in which we predicted no, but they were actually readmitted; Type II error.

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{FN}+\text{TN}) \quad (1)$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \quad (2)$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (3)$$

$$\text{F1\_score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (4)$$

Accuracy indicates how often the classifier is correct. The recall is a sensitivity measure (ratio of TPs to the sum of TPs and FNs). It indicates the rate of cases the model predicted the patient will be readmitted in a month period (relative to the number of cases the patient was actually readmitted). The precision measures the rate of cases that the model predicts the patient will be readmitted in a month period correctly compared to total number of cases in which the model predicts the patients will be readmitted. Table VI depicts the values of the performance measures.

As previously noted, we used 10-fold cross validation for the models. Table VII and Fig. 10 depict the training and testing accuracy for each model under 10-fold cross validation.

Finally, Table VIII illustrates the minimum, maximum and mean accuracy for every model. It is clear that SVM achieved the highest accuracy of 0.9522. It was followed by DT, with accuracy of 0.9251, and then MLP with accuracy of 0.8358. The lowest performance was detected for NB classifier and LR analysis, which achieved respectively 0.69069 and 0.6865.

TABLE VI. PERFORMANCE MEASURES FOR THE EMPLOYED TECHNIQUES

Model/ Measures	Accuracy	Recall	Precision	F1_score
Decision Tree	0.878869	0.854745	0.908571	0.876033
Logistic Regression	0.641095	0.626773	0.651013	0.638663
Naïve Bayes	0.638852	0.433511	0.746565	0.548514
Support Vector Machine	0.294630	0.942376	0.911664	0.926765
Multi-Layer Perceptron	0.799910	0.781028	0.815741	0.798007

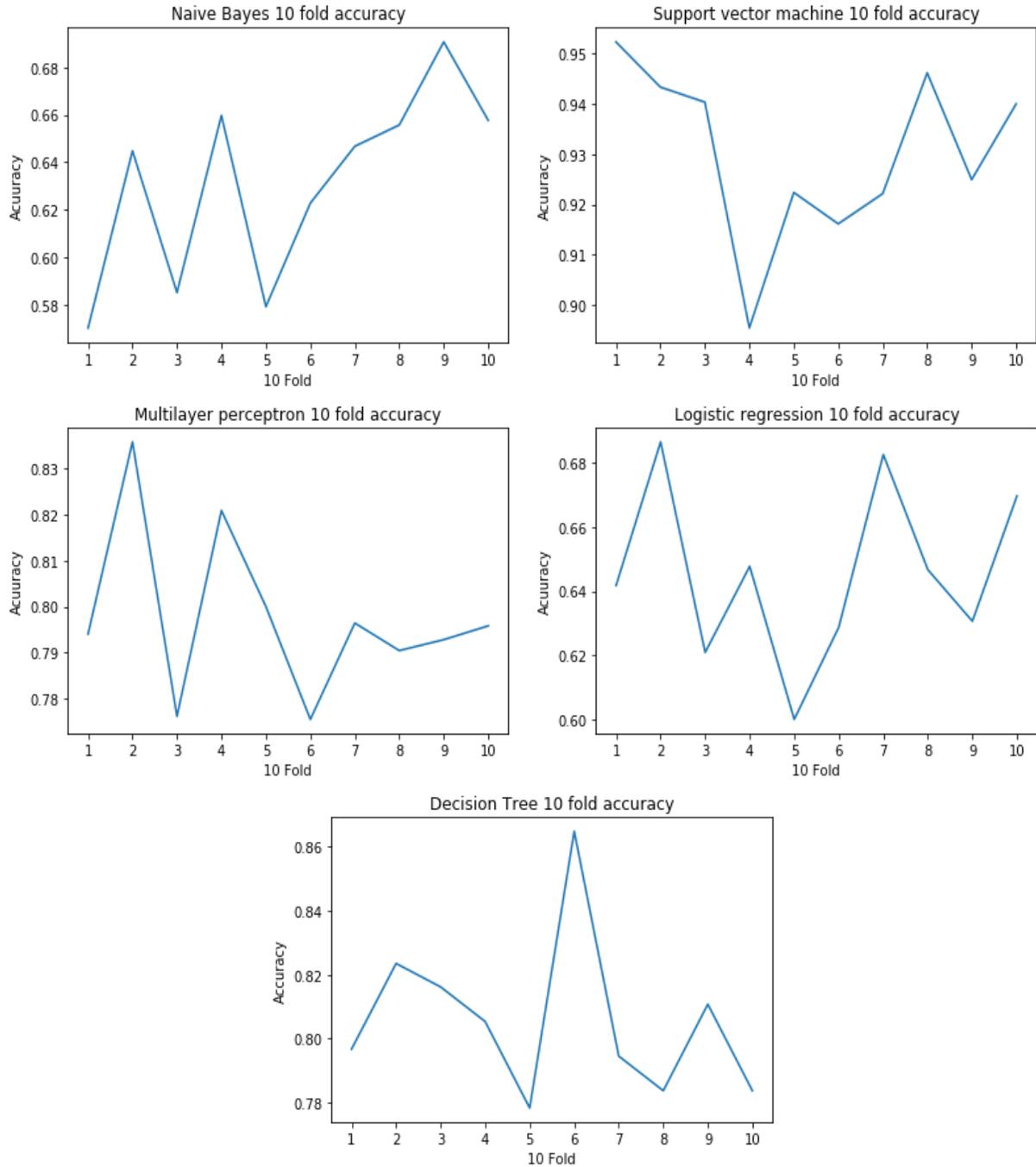


Fig. 10. 10-Fold Cross Validation for Models.

TABLE VII. 10-FOLD CROSS VALIDATION FOR EMPLOYED TECHNIQUES

SVM	CART	NB	LR	MLP
0.952239	0.898507	0.570149	0.641791	0.794030
0.943284	0.901493	0.644776	0.686567	0.835821
0.940299	0.886567	0.585075	0.620896	0.776119
0.895522	0.907463	0.659701	0.647761	0.820896
0.922388	0.853731	0.579104	0.600000	0.800000
0.916168	0.883234	0.622754	0.628743	0.775449
0.922156	0.865269	0.646707	0.682635	0.796407
0.946108	0.925150	0.655689	0.646707	0.790419
0.924925	0.906907	0.690691	0.630631	0.792793
0.939940	0.906907	0.657658	0.669670	0.795796

TABLE VIII. ACCURACY OF EMPLOYED TECHNIQUES

No	Model	Min	Max	Mean
1	Support Vector Machine	0.895522	0.952239	0.930303
2	Decision Tree	0.853731	0.925150	0.893523
3	Multi-Layer Perceptron	0.775449	0.835821	0.797773
4	Naïve Bayes	0.570149	0.690691	0.631230
5	Logistic Regression	0.600000	0.686567	0.645540

## VI. CONCLUSION AND FUTURE WORK

This paper presented a comparative study among five machine learning techniques; namely LR, MLP, NB classifier, decision trees, and SVMs; for predicting the likelihood of hospital readmission of diabetes patients. The study relied on real data collected from hospitals in the United States. Based on the study, the SVM provided the best performance. Nevertheless, the study will be extended to compare additional techniques and larger datasets will be considered as well.

### REFERENCES

[1] G. Roglic, "Global report on diabetes.," World Heal. Organ., vol. 58, no. 12, pp. 1-88, 2016.

[2] D. Rubin, K. Donnell-Jackson, R. Jhingan, S. Golden, and A. Paranjape, "Early readmission among patients with diabetes: A qualitative assessment of contributing factors," J. Diabetes Complications, vol. 28, no. 6, pp. 869-873, 2014.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104-116, 2017.

[4] P. Chowriappa, S. Dua, and Y. Todorov, "Machine Learning in Healthcare Informatics," vol. 56, pp. 1-23, 2014.

[5] T. Mitchell, "Machine learning (mcgraw-hill international editions computer science series)," 1997.

[6] E. Bose and K. Radhakrishnan, "Using Unsupervised Machine Learning to Identify Subgroups among Home Health Patients with Heart Failure Using Telehealth," CIN - Comput. Informatics Nurs., vol. 36, no. 5, pp. 242-248, 2018.

[7] L. Kaelbling, A. Littman, and A. Moore, "Reinforcement learning: A survey," J. Artif. Intell. Res., vol. 4, pp. 237-285, 1996.

[8] K. Shailaja, B. Seetharamulu, and M. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second Int. Conf. Electron. Commun. Aersp. Technol., no. Iceca, pp. 910-914, 2018.

[9] J. Davies and J. Gibbons, "Machine Learning and Software Engineering in Health Informatics," in Proceedings of the First International

Workshop on Realizing AI Synergies in Software Engineering, 2012, pp. 37-41.

[10] R. Bhardwaj, A. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 236-241, 2017.

[11] A. Asuncion and D. Newman, "UCI Machine Learning Repository," 2007. [Online]. Available: <https://archive.ics.uci.edu/ml/index.php>.

[12] S. B. et al., "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," Biomed Res. Int., vol. 2014, 2014.

[13] A. H. Karp, "Using logistic regression to predict customer retention," Proc. Elev. Northeast SAS Users Gr. Conf. <http://www.lexjansen.com/nesug/nesug98/solu/p095.pdf>, 1998.

[14] F. Amato, A. López, E. M. Peña-Méndez, P. Va?hara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," J. Appl. Biomed., vol. 11, no. 2, pp. 47-58, 2013.

[15] S. F., "Machine-Learning Techniques for Customer Retention: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273-281, 2018.

[16] N. Jothi, N. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," Procedia Comput. Sci., vol. 72, pp. 306-313, 2015.

[17] D. Sisodia and D. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578-1585, 2018.

[18] A. Hazra, S. Kumar, and A. Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms," Int. J. Comput. Appl., vol. 145, no. 2, pp. 39-45, 2016.

[19] E. Holzschuh, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting\*," Reports Prog. Phys., vol. 55, no. 7, pp. 1035-1091, 1992.

[20] R. Sharma, V. Sugumaran, H. Kumar, and M. Amarnath, "A comparative study of naive Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal," Int. J. Decis. Support Syst., vol. 1, no. 1, p. 115, 2015.

[21] S. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques?: A Review," vol. 10, no. 7, pp. 2137-2159, 2017.

[22] N. Nai-Arun and P. Sittidech, "Ensemble Learning Model for Diabetes Classification," Adv. Mater. Res., vol. 931-932, pp. 1427-1431, 2014.

[23] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," Procedia Comput. Sci., vol. 82, no. March, pp. 115-121, 2016.

[24] K. Orabi, Y. Kamal, and T. Rabah, "Early Predictive System for Diabetes Mellitus Disease," vol. 1, pp. 420-427, 2016.

[25] D. Paul, "Analysing Feature Importances for Diabetes Prediction using Machine Learning Debadri," 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf., pp. 924-928, 2018.

[26] J. Kerexeta, A. Artetxe, V. Escolar, A. Lozano, and N. Larburu, "Predicting 30-day Readmission in Heart Failure using Machine Learning Techniques," Proc. 11th Int. Jt. Conf. Biomed. Eng. Syst. Technol., vol. 5, no. Biostec, pp. 308-315, 2018.

[27] S. F., "Machine-Learning Techniques for Customer Retention: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273-281, 2018.

[28] K. Kira and L. A. Rendell, A Practical Approach to Feature Selection. Morgan Kaufmann Publishers, Inc., 1992.

[29] D. W. Opitz, "Feature selection for ensembles," Proc. 16th Natl. Conf. Artif. Intell. AAAI, vol. 16, no. 3, pp. 379-384, 1999.

[30] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection 1 Introduction," An Introd. to Var. Featur. Sel., vol. 3, pp. 1157-1182, 2003.

[31] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Managemet, vol. 45, 2009, pp. 427-437.