Assessing Advanced Machine Learning Techniques for Predicting Hospital Readmission

Samah Alajmani[1], Kamal Jambi[2]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Predicting the probability of hospital readmission is one of the most important healthcare problems for satisfactory, high-quality service in chronic diseases such as diabetes, in order to identify needful resources such as rooms, medical staff, beds, and specialists. Unfortunately, not many studies in the literature address this issue. Most studies involve forecasting the probability of diseases. For prediction, several machine learning methods can be implemented. Nonetheless, comparative studies that identify the most effective approaches for the method prediction are also insufficient. With this aim, our paper introduces a comparative study in the literature across five popular methods to predict the probability of hospital readmission in patients suffering from diabetes. The selected techniques include linear discriminant analysis, instance-based learning (K-nearest neighbors), and ensemble-based learning (random forest, AdaBoost, and gradient boosting) techniques. The study showed that the best performance was in random forest whereas the worst performance was shown by linear discriminant analysis.**

Keywords*—Boosting; random forest; linear discriminant analysis; k-nearest neighbor; machine learning; hospital readmission; predictive analytics

I. Introduction

Healthcare is globally recognized as an important sector. It is one of the most rapidly growing divisions in the employment industries worldwide [1]. According to Russell Reynolds Associates, healthcare costs are predicted to reach a staggering $12 trillion within the next 7 years. Current costs are between $6–$7 trillion [2]. Looking at these figures, it is obvious that healthcare is at a crucial point in the growth and evolution of medicine. A perfect example of this can be seen in the United States, where the total expenditure on healthcare increased by up to 5.3%, and also topped $3 trillion nationally. In 2012, more than 1.5 million people died of diabetes, which is one of the most common and chronic illnesses of our times [3]. This serious disease continues to affect many people around the world. Diabetes affects more than twenty-three million person in the United States alone [4]. Furthermore, the main concern in diabetes care is hospital readmission, reasoned by spending of more than two hundred fifty million dollars on medications for diabetic patients who were readmitted in 2011 [4]. The Agency for Healthcare Research and Quality (AHRQ) revealed more than three million readmissions during a 30-day window in the US. Contribution of these hospital readmissions reached about forty-one billion dollars in hospital costs [5]. Additionally, the totals deaths in 2012 was estimated to be 3.7 million, which included 1.5 million deaths due to diabetes, and an additional 2.2 million deaths due to cardiovascular diseases, chronic kidney disease, and tuberculosis related to higher-than-optimal blood glucose. Research has shown that 43% of these 3.7 million deaths occur before individuals have reached the age of 70. Statistically, diabetes and high blood pressure are more prevalent in lower and middle-class people, as opposed to the higher-income group [3]. Hence, these diabetic patients tend to frequently visit healthcare facilities and hospitals, which guarantees them access to hospital resources and services, for example, the availability of sufficient services by care providers and other hospital staff, medical equipment, early detection and diagnosis of illnesses and diseases, medical treatments, and check-ups and plans developed by the medical staff for the patient. Preventing hospitalization is a distinguished aspect of limiting an affected person's morbidity, improving their results, and constraining healthcare costs [6]. Accordingly, their ultimate purpose is to predict readmission possibility. In reality, readmission for 30 days, that is, within one month of discharge is an excellent indicator of high priority healthcare. The aim of this study is to discuss this issue as well [7].

Machine learning is one of the most popular and leading analytical techniques developed in modern times. It intends to solve highly complicated tasks because it is essential to concentrate on the most appropriate data from seemingly enormous amounts of data [8]. This type of learning includes gathering information from various fields to redefine issues beyond the normal limitations and to arrive at solutions based on a novel understanding of complicated attitudes. It expands over the fields of statistics, algebra and knowledge, data processing, analytics, etc. This type of learning is also influenced by artificial intelligence, control theory, biology, philosophy, information technology, cognitive science, and mathematical calculations. With machine learning, gathering accurate data—ranging from medical records, financial transactions, applications of loans, and supply maintenance—has become possible [9].

Machine learning can be categorized into three groups. Supervised learning is subject to learning as a data scientist attempts to teach a data training algorithm and its possible outcomes. Classification and regression are two different models in machine learning. The classification model aims to forecast unique classes such as blood groups, whereas the regression model predicts numerical values [10]. On the other hand, unsupervised learning tries to look for a pattern in hidden data, associations among variables, or trends in data

[10] [11]. The major goal of this type of learning is the capability to determine either the distribution of data or the hidden structure without the earlier categorization of training data or without being subject to supervision [12]. Lastly, reinforcement learning is learning by an individual where they can study behavior by means of both interactions (the trial and error method) and dynamic environment. Through this learning, a computer program can provide access to the dynamic environment for executing a special goal. It is essential to understand that the system did not have prior knowledge of the behavior of the environment and the only probable method will be through trial and error [10] ,[13]. It is worth noting that in this study, we extend our work in paper [14] by adding five other techniques to compare and detect the best accuracy among the following techniques regarding hospital readmission prediction: 1) Linear discriminant analysis (LDA); 2) Instance-based learning: K-nearest neighbor; and 3) Ensemble-based learning: AdaBoost, gradient boosting, and random forest.

The remainder of the paper is presented as follows: Section 2 includes a background discussion on machine learning algorithms used in this study. Section 3 presents related work on comparative study of machine learning techniques in the healthcare sector. Section 4 discusses the study methodology and presents the outcomes of the experiments. Section 5 summarizes and concludes the study.

II. Background

Developments in machine learning are clearly visible in different fields and industries in the past years. Hence, researchers have discussed the possibility of using machine learning technologies in healthcare, outlining different initiatives in the healthcare domain. Machine learning has many benefits for healthcare, such as predicting readmission in hospitals and diseases, among others. In addition, machine learning is capable of discovering a solution to form a strong relationship between the patient and doctor for reducing the increasing healthcare costs [1]. This section addresses different machine learning techniques used in this study.

A. Linear Discriminant Analysis

LDA is an essential data analysis approach, which has been widely applied in the past for recognition, dimensionality reduction, and supervised classification [15]. it is a mathematical classification technique which can search for a collection of predictors to distinguish between two targets. It is also correlated to regression analysis in that both try to express the relation between an independent variables group and a single dependent variable [16].

B. Instance-Based Learning

This algorithm can be called "Memory-based learning." The most broadly utilized technique of instance-based learning for classification is K-nearest neighbor (KNN) [16]. This method is largely applied to sample classification. The KNN technique can measure the distance from training samples number N [17]. This technique does not attempt to build an internal model, and computations are not executed prior to the classification. In the features space, the training data instances are hardly retained. Next, depending upon the vote's majority from the neighbors, a class of instance is determined. Moreover, an instance is determined for a class that is most common among the neighbors. For variables that are continuous, Murkowski, Euclidian, or Manhattan distance techniques are used, whereas the Hamming technique is used for variables that are categorical [16]. Depending upon the distance, the neighbors are determined using the KNNs. The determined distances are used to recognize and allocate labels to training instances' (k) groups that are nearest to the new point. Despite its simplicity, the KNN has been utilized in a considerable number of applications.

C. Ensemble-Based Learning

Ensemble-based learning techniques lead to predictions that depend on a collection of several classifier outputs. Ensemble learners consist of boosting methods, for example, AdaBoost and gradient boosting, along with bagging methods such as random forest [16]."Boosting" signifies a general and effective strategy to yield highly precise prediction by gathering hard and slightly inexact thumb rules [18]. On the other hand, "bagging" depends on a bootstrapping strategy, in which different classification trees are improved by constantly choosing arbitrary training data subsets [19].

1) Boosting-Based techniques: AdaBoost and Stochastic Gradient Boosting rely on the concept of boosting [16]. AdaBoost (AB) is based on creating a prediction rule, which is extremely accurate, by joining a number of comparatively weak and inexact rules [20]. Furthermore, it is simple, swift, and easy to use with an iterative algorithm which requires only one parameter, iteration number. Moreover, it is not subject to over-fitting and simply determines outliners which are incorrectly classified or are difficult to classify [21]. However, misclassified and/or difficult instances are given significance by gradient boosting (GB), via the remaining errors—also known as pseudo-residuals—of a strong learner. With every iteration, errors are measured, and a weak learner adapts to them. Afterwards, the weak learner contributes to minimizing the total error of the strong learner [16].

2) Baging techniques: Random forests are a combination of tree predictors. It is an ensemble learning method (in addition to the thought that it could be a form of the nearest neighbor predictor). It creates a number of decision trees at the time of training and produces a class, which is the output of classes through individual trees. Furthermore, it attempts to reduce increased bias and variance issues through averaging to detect a balance between the two extremes [22].

III. Related Work

Apart from predicting the probability of hospital readmission, many studies have attempted to use machine learning techniques in healthcare problems. For example, AOA et al. [23] clarified the importance of machine learning techniques in identifying predictive and diagnostic indicators in a set of wide-scale data with extremely elevated geometric relation to genetics. They used SVM, logistic regression (LR), and NBs and proved that SVM had the best accuracy among others. Arun and Sittidech [24] used decision tree (DT), KNN, and NBs to build diabetes classification models. Then,

boosting and bagging were executed using the base classifiers of KNN, NBs, and DT. Based on their tests, they concluded that the greatest accuracy was obtained when bagging was applied with DT. Sisodia and Sisodia [25] presented a model that could predict the probability of diabetes with high accuracy. In their experiment, they applied three techniques: NBs, DT, and SVM, to discover diabetes in its early stages. In accordance with their outcomes, NB achieved better accuracy than other algorithms. Singh [26] conducted an experiment to predict diabetes through the utilization of various machine learning techniques; the accuracy of the proposed technique was 87-95% better than others: DT (C4.5) at 81-85%, Bays classifier at 84-88%, and KNN at 80-82%. However, Shahon et al. had a different intention; they tried using AdaBoost to improve the overall accuracy of models. Consequent to these experiments, it was clear that AdaBoost had a better accuracy than standalone DTs, such as J48, and bagging [27]. Orabi et al.'s approach [28] to fuse regression included randomization. This method achieved an 84% accuracy rate when predicting diabetes according to age. Other investigators suggested a predictive model. They used three techniques of machine learning—SVMs, RFs, and LR—to predict diabetes in Indian women, as well as the factors that could cause the disease. Their comparative study proved that the RFs had the highest performance among other models [29].

In comparison, only a few studies have discussed the prediction of hospital readmission probability. For example, Strack et al. [30] utilized traditional statistical models toward this end. Some investigators concentrated on the comparison of various machine learning techniques to address the issue. For example, Alexander et al. suggested two methods. First, they merged unsupervised and supervised techniques of classification, and subsequently, merged DT and NB. They proved that the former method had better accuracy than the latter method with regard to readmission prediction [31]. Finally, Alajmani and Elazhary [14] used LR, multi-layer perceptron (MLP), NB, SVM, and DT to predict hospital readmission and evaluate accuracy among models. Based on their results, SVM achieved the highest accuracy of 95.22% among other techniques.

In general, very few studies in the healthcare sector are devoted to predicting the possibility of hospital readmission. In addition, there is a lack of research on comparison among different machine learning techniques for prediction. Consequently, this paper attempts to address and discuss both issues regarding the probability prediction of hospital readmission, which depends on real data with different algorithms.

IV. Methodology

It is imperative to clearly understand the data prior to commencing a comparative study, conduct pre-processing when needed, and choose appropriate features for the experiments. It is also important to mention that all experiments in this study were conducted using Python.

A. Dataset Explanation and Features

*1) Data comprehension:*This paper utilizes a sample dataset of diabetic patients from different hospitals across the US [32] , [30]. Such a dataset encompasses 13460 instances from age groups 30–50, with eighteen features. In Table I, the dataset variables and their associated descriptions are presented. Scientific interpretations of these features are beyond this article's scope. In addition, the distribution of features is depicted in Fig. 1.

2) Data pre-processing: This phase, which encompasses both data transformation and data cleaning, is considered to be a significant step. We tried to use an approach that is frequently used and more general in converting categorical variables into variables of real-value; this approach is called one-hot encoding [33]. First, with regard to data transformation, certain categorical variables such as Gender, Change, Age and DiabetesMed are converted into binary forms 0 or 1. Second, with regard to data cleaning, missing values of categorical data need to be accounting for. Toward this aim, the imputation is performed via the categorical data mode. This imputation method helps us with better prediction model performance in cases where missing data has already hidden helpful information [34]. After preprocessing the data became 3090 instances.

3) Feature selection: Here, feature selection is applied to reduce dimensionality, meaning we opt for features that are most relevant. In this research paper, the effect of variables on our target is evaluated. Moreover, this results in the elimination of low-importance variables. The most significant among them are features with high influence on accuracy [16]. The GB technique has been utilized [35] for categorical variables. The variables' average weights are demonstrated in Table II. Subsequently, a threshold of 0.014 is used to attain the variable set. Consequently, the features Age, Admission_source_id, and DiabetesMed are excluded as their weights are less than 0.014. However, the other features demonstrated in Fig. 2 are chosen and selected.

B. Constructing Models of Machine Learning

In this paper, the chosen models have 1 target/output with 2 values, which can be true or false regarding readmission to the hospital within a span of one month. This means the value of the readmission variable is TRUE if the patient is readmitted within a time span of one month. However, if there is no readmission, or if readmission has been carried out after the one-month period, then the value will be FALSE. As mentioned earlier, the driver set for forecasting consist of the selected features. The datasets for training and testing are selected randomly. Moreover, by choosing a 40% testing dataset and a 60% training dataset, a ten-fold cross-validation is applied.

1) Linear discriminant analysis: This model is built using the next parameters n_components, solver, and tol, where n_components is the number of components (< n_classes-1) for reducing dimensionality. Solver "svd" is the decomposition of a singular value. Finally, tol "1e-5" is the threshold to be utilized for estimation of rank in solver of svd. The accuracy of LDA is 0.6388515 and a 10-fold cross-validation is conducted for this model.

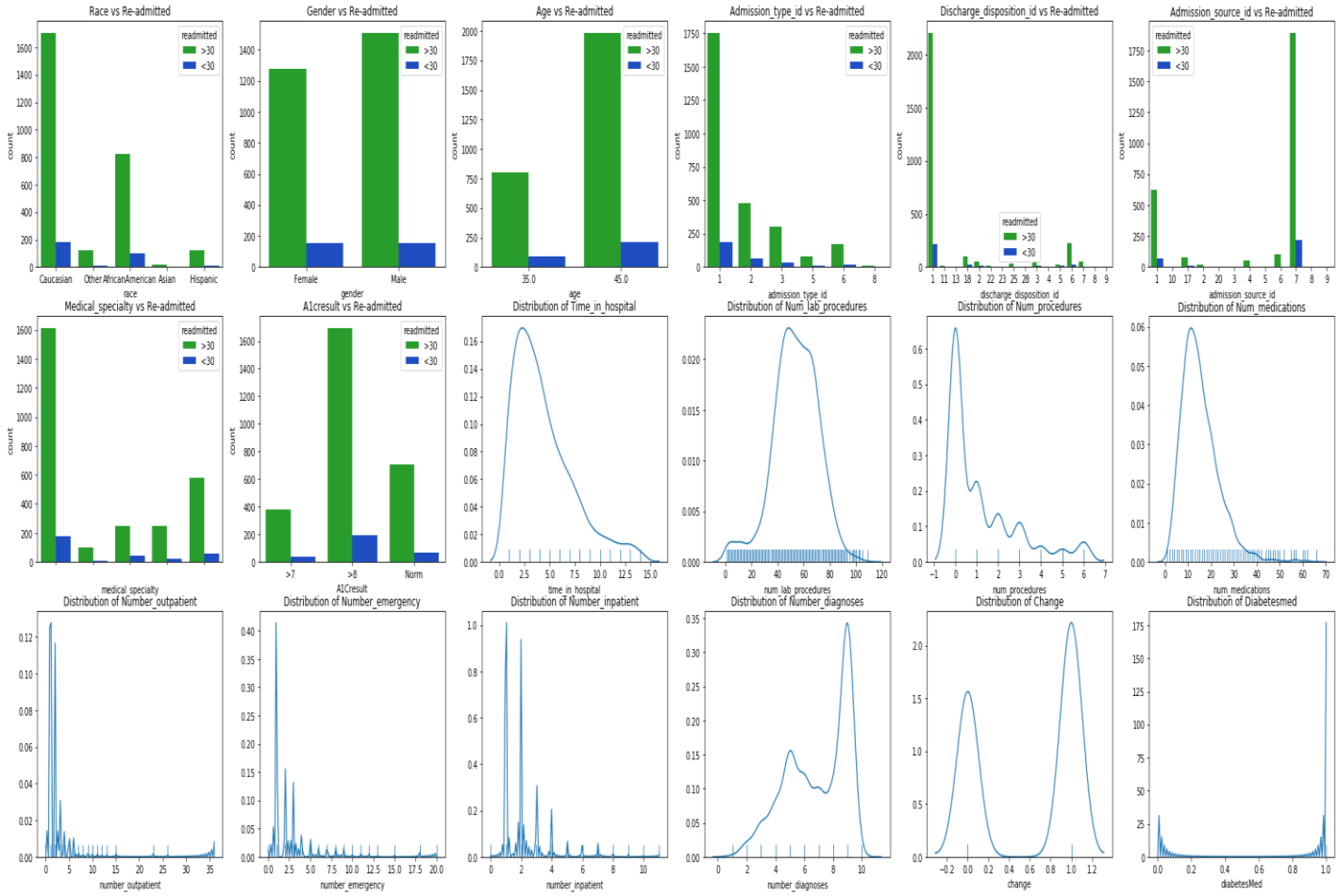

Fig. 1. Features Distribution.

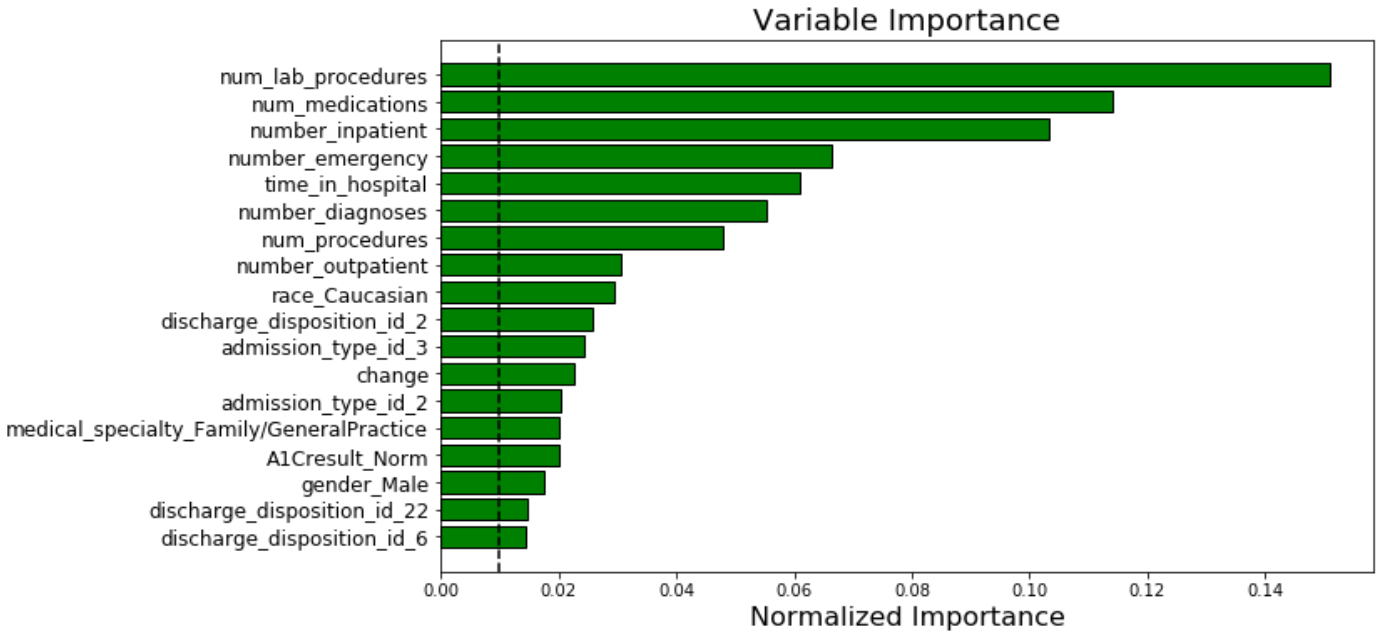

Fig. 2. Variables Importance.

2) K-Nearest neighbor: In this model, the most important parameter is n_neighbors, which represents the number of neighbors for use by default for k neighbors queries. Cross-validation is executed using different values of n_neighbors. Table III illustrates that the highest accuracy = 0.8847016 when n_neighbors = 5.

3) Adaboost: AdaBoost needs three important parameters: (1) n_estimators indicate the number of weak learners for repeat training, (2) learning_rate contributes to weak learners' weights, and (3) algorithm "SAMME" or "SAMME.R." This model uses grid search to evaluate the optimal accuracy and hyperparameters. The best parameters are n_estimators = 5000, algorithm = "SAMME," and learning rate=0.9; thus, the best accuracy of 0.9318079 is clarified in Table IV below.

4) Gradient boosting: This model is built using the following important parameters: (1) n_estimators, which can present number of boosting stages for execution, (2) learning rate indicates the shrinking of learning rate of every tree contribution, (3) creation is the function for measuring the split quality, and (4) max_depth refers to the maximum depth which can limit the number of nodes in the tree. Tuning the max_depth is important to get the best performance. Grid search is used to measure optimum accuracy and hyperparameters. Table V demonstrates that the best accuracy = 0.9362943 when n_ estimators = 200.

5) Random forest: We construct this model using 250 trees in the forest where 26 is the maximum depth of the tree and 10 is the lowest number of samples for dividing an inner node. In addition, grid search is used to find the best accuracy and the best parameters. The following Table VI presents the results.

TABLE. I. DIABETSE DATASET

Variable	Data type
Race	Categorical
Change	Categorical
DiabetesMed	Categorical
Age	Categorical
A1Cresult	Categorical
Gender	Categorical
Num_lab_procedures	Integer
Num_procedures	Integer
Num_inpatient	Integer
Num_oupatient	Integer
Num_medications	Integer
Num_diagnosis	Integer
Num_emergency	Integer
Medical_spacialty	Categorical
time_in_hospital	Integer
Admission_type_id	Integer
Admission_source_id	Integer
Discharge_disposition-id	Integer

TABLE. II. FEATURES IMPORTANCE

Variable	Importance	Decision
Race	0.029016	Acceptable
Change	0.023027	Acceptable
DiabetesMed	0.008867	Unacceptable
Age	0.010165	Unacceptable
A1Cresult	0.020177	Acceptable
Gender	0.020294	Acceptable
Num_lab_procedures	0.149317	Acceptable
Num_procedures	0.046521	Acceptable
Num_inpatient	0.104099	Acceptable
Num_outpatient	0.030696	Acceptable
Num_medications	0.111058	Acceptable
Num_diagnosis	0.055811	Acceptable
Num_emergency	0.066718	Acceptable
Medical_spacialty	0.019117	Acceptable
time_in_hospital	0.062025	Acceptable
Admission_type_id	0.023854	Acceptable
Admission_source_id	0.008961	Unacceptable
Discharge_disposition-id	0.027554	Acceptable

TABLE. III. KNN ACCURACY

n_neighbors	Accuracy
5	0.8847016
10	0.8501570
15	0.8205473

TABLE. IV. ADABOOST ACCURACY

n_neighbors	Accuracy
500	0.9255271
1000	0.9286675
5000	0.9318079

TABLE. V. GRADIENT BOOSTING ACCURACY

n_estimators	Accuracy
100	0.9344997
150	0.9358456
200	0.9362943

TABLE. VI. RANDOM FOREST ACCURACY

n_estimators	Accuracy
150	0.9349484
250	0.9358456
350	0.9344997

V. DISSCUSIOM AND RESULTS

In this study, different performance measures are utilized to compare the studied techniques [36]. Particularly, precision, accuracy, F1 scores, and recalls are relied upon for this reason. As presented in Equations 1, 2, 3, and 4, these parameters are described by true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Furthermore, TPs refer to cases where the prediction is YES, that is, patients will be readmitted in hospital within a duration of 30 days and when there is a match, meaning that the patients are indeed readmitted. Whereas, TNs refer to those cases where the prediction is a NO, and when the patients are NOT readmitted. On a different note, FPs refer to cases where the prediction is a YES, but patients are NOT readmitted, that is, a type I error. Lastly, FNs refer to cases where the prediction is a NO, but the patients are actually readmitted, that is, a type II error.

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

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

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

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

Accuracy refers to the frequency of the classifier being true. The recall is a sensitivity measure, for example, the proportion of TPs to the total number of TPs and FNs. It indicates the rate of cases where the model predicts patient readmission within a time span of 30 days, related to the number of events where the subject is actually readmitted. Alternatively, precision is a calculation of the rate of events when the model accurately predicts the patient's readmission during the 30-day time period, in contrast to sum of events when the model forecasts the patient's readmission. In Table VII, the performance measure values are illustrated.

As previously mentioned, we performed 10-fold cross-validation of the listed techniques. For each model, the training and testing accuracy with respect to 10-fold cross-validation is shown in Table VIII and Fig. 3.

Lastly, for the chosen models, the lowest, highest, and mean accuracies are illustrated in Table IX. It is obvious that ensemble-based learning (RF and AdaBoost) techniques accomplish the maximum accuracy of 0.9579 and 0.9550, respectively. Further, GB's accuracy is 0.9459 while KNN's accuracy is 0.9161. The least value of performance accuracy is 0.6835 in LDA. The complexity of each algorithm followed by each classification is the main reason behind the performance variation.

TABLE. VII. PERFORMANCE MEASURES FOR THE SELECTED MODELS

Models / Measures	Accuracy	Precision	F1_score	Recall
Random Forest	0.932705	0.988024	0.929577	0.877660
AdaBoost	0.931808	0.992929	0.928234	0.871454
Gradient Boosting	0.932705	0.970192	0.930812	0.894504
K-Nearest Neighbor	0.884702	0.857847	0.890405	0.925532
Linear Discriminant Analysis	0.638852	0.646952	0.638527	0.630319

TABLE. VIII. PERFORMANCE MEASURES FOR THE SELECTED MODELS

Random Forest	AdaBoost	Gradient Boosting	K-Nearest Neighbor	Linear Discriminant Analysis
0.937313	0.937313	0.925373	0.889552	0.623881
0.940299	0.952239	0.931343	0.904478	0.683582
0.937313	0.940299	0.943284	0.889552	0.614925
0.934328	0.934328	0.934328	0.889552	0.656716
0.931343	0.931343	0.916418	0.865672	0.600000
0.916168	0.913174	0.898204	0.838323	0.610778
0.931138	0.928144	0.934132	0.892216	0.679641
0.952096	0.955090	0.943114	0.916168	0.634731
0.957958	0.936937	0.945946	0.900901	0.630631
0.942943	0.927928	0.930931	0.909910	0.642643

TABLE. IX. SELECTED MODELS ACCURACY

Model	Min	Max	Mean
Random Forest	0.916168	0.957958	0.938090
AdaBoost	0.913174	0.955090	0.935679
Gradient Boosting	0.898204	0.945946	0.930307
K-Nearest Neighbor	0.838323	0.916168	0.890229
Linear Discriminant Analysis	0.600000	0.683582	0.637753

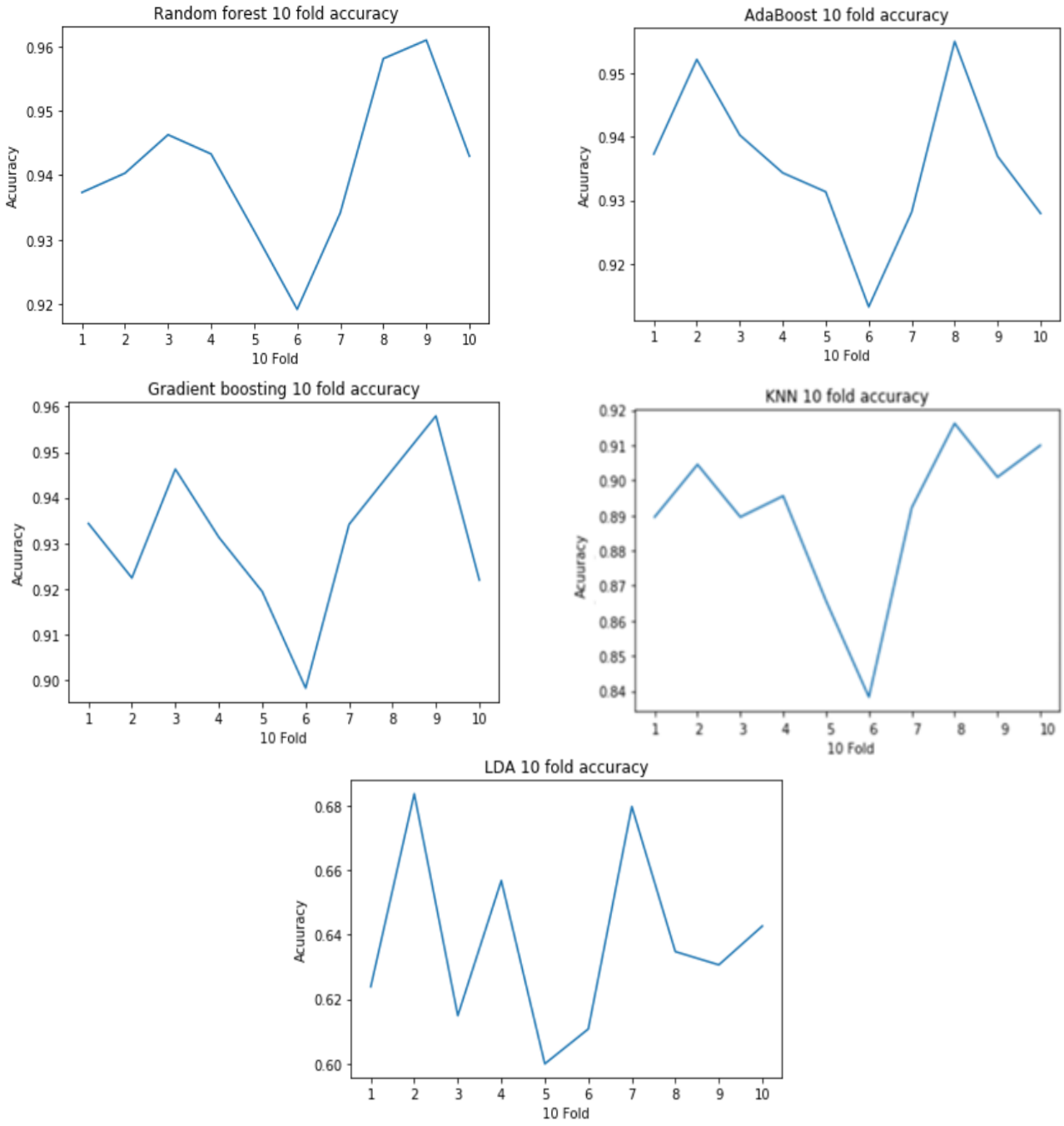

Fig. 3. 10-Folds Cross Validation for Models.

VI. Conclusion and Future Work

For the prediction of readmission, this research seeks to offer a standard for the most commonly applied modern features. The reliability of the chosen models is measured on a real dataset of diabetes from different hospitals in the USA. Based on the outcomes of this study, ensemble-based learning algorithms are suggested, and the highest accuracy is shown by the RF model, followed by the AdaBoost model. Nevertheless, the research will be expanded to a bigger dataset using different machine learning techniques.

References

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

[2] J. Bouwens and D. M. Krueger, "Embracing Change: The Healthcare Industry Focuses on New Growth Drivers and Leadership Requirements," Rusell Reynolds Associates. [Online]. Available: https://www.russellreynolds.com/insights/thought-leadership/embracing-change-the-healthcare-industry-focuses-on-new-growth-drivers-and-leadership-requirements.

[3] G. Roglic, "Global Report on Diabetes.," World Heal. Organ., vol. 58, no. 12, pp. 1–88, 2016.

[4] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying

Diabetic Patients with High Risk of Readmission," arXiv Prepr. arXiv1602.04257., 2016.

[5] "AHRQ: The Conditions that Cause the Most Readmissions," Advisory, 2014. [Online]. Available: https://www.advisory.com/daily-briefing /2014/04/22/ most-common-readmissions. [Accessed: 30-Oct-2019].

[6] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S. C. Chin, and B. Muckian, "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients," in 2013 IEEE International Conference on Big Data.IEEE., 2013, pp. 64–71.

[7] D. J. Rubin, K. Donnell-Jackson, R. Jhingan, S. H. Golden, and A. Paranjape, "Early Readmission among Patients with Diabetes: A Qualitative Assessment of Contributing Factors," J. Diabetes its Complicat. Elsevier., vol. 28, no. 6, pp. 869–873, 2014.

[8] A. L. Bluma and P. Langley, "Artificial Intelligence Selection of relevant features and examples in machine," vol. 97, no. 97, pp. 245–271, 1997.

[9] T. Mitchell, "Machine Learning, McGraw-Hill Higher Education," New York, 1997.

[10] 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.

[11] P. Chowriappa, S. Dua, and Y. Todorov, "Introduction to Machine Learning in Healthcare Informatics," in Machine Learning in Healthcare Informatics.Springer, 2014, pp. 1–23.

[12] 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.

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

[14] S. Alajmani and H. Elazhary, "Hospital Readmission Prediction Using Machine Learning Techniques: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, pp. 212–220, 2019.

[15] P. Markopoulos, "Linear Discriminant Analysis with Few Training Data," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4626–4630.

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

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

[18] Y. Freund and R. Schapire, "A Short Introduction to Boosting," Journal-Japanese Soc. Artif. Intell., vol. 14, no. 771–780, p. 1612, 1999.

[19] G. Chirici et al., "Stochastic Gradient Boosting Classification Trees for Forest Fuel Types Mapping Through Airborne Laser Scanning and IRS LISS-III imagery," Int. J. Appl. Earth Obs. Geoinf., vol. 25, no. 1, pp. 87–97, 2013.

[20] R. E. Schapire, "Explaining Adaboost," in Empirical inference. Springer, 2013, pp. 37–52.

[21] N. Emanet, H. R. Öz, N. Bayram, and D. Delen, "A Comparative Analysis of Machine Learning Methods for Classification Type Decision Problems in Healthcare," Decis. Anal., vol. 1, no. 1, p. 6, 2014.

[22] L. Breiman, "Random Forests," Mach. Learn. Springer, vol. 45, no. 1, pp. 5–32, 2001.

[23] L. Cai, H. Wu, D. Li, K. Zhou, and F. Zou, "Type 2 Diabetes Biomarkers of Human Gut Microbiota Selected via Iterative Sure Independent Screening Method," PLoS One, vol. 10, no. 10, pp. 1–15, 2015.

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

[25] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes Using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018.

[26] A. Singh, "Comparing Data Mining Algorithms for Diabetes Disease Prediction," Int. J. Contemp. Technol. Manag., 2018.

[27] 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.

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

[29] D. Dutta, D. Paul, and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction Using Machine Learning Debadri," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, IEEE, 2018, pp. 924–928.

[30] 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.

[31] J. Kerexeta, A. Artetxe, V. Escolar, A. Lozano, and N. Larburu, "Predicting 30-day Readmission in Heart Failure Using Machine Learning Techniques," in Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies., 2018, vol. 5, no. Biostec, pp. 308–315.

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

[33] K. Potdar, T. S., and C. D., "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," Int. J. Comput. Appl., vol. 175, no. 4, pp. 7–9, 2017.

[34] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases," Appl. Intell., vol. 11, no. 3, pp. 259–275, 1999.

[35] Z. E. Xu, K. Q. Weinberger, and A. X. Zheng, "Gradient Boosted Feature Selection Categories and Subject Descriptors," Kdd, pp. 522–531, 2014.

[36] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.