Frontiers in Health Informatics
2021; 10: 74

www.ijmi.ir
Open Access

Predicting hospital readmission of diabetic patients using machine learning

Boshra Farajollahi[1]*, Maysam Mehmannavaz[2], Hafez Mehrjoo[2], Fateme Moghbeli[3], Mohammad Javad Sayadi[1]

[1]Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran
[2]Doornama Company, Data Science lab, Ilam, Iran
[3]PhD of Medical Informatics, Assistant Professor, Department of HIT, Varastegan Institute for Medical Sciences, Mashhad, Iran

Article Info

Article type:
Research

Article History:
Received: 2020-12-12
Accepted: 2021-03-23
Published: 2021-05-01

**** Corresponding author:***
Boshra Farajollahi

Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

Email:
boshrafarajollahi1373@gmail.com

Keywords:
Prediction
Diabetic Patients
Machine Learning

A B S T R A C T

Introduction: Diabetes is a chronic disease associated with abnormal high levels of glucose in the blood. Diabetes make many kinds of complications, which also leads to a high rate of repeated admission of patients with diabetes. The goal of this study is to Predict hospital readmission of Diabetic patients with machine learning techniques.

Material and Methods: The data used in the study are data obtained from the UCI machine learning repository about diabetic patients. The dataset used contains 100,000 instances and it include 55 features from 130 hospitals in the United States for 10 years.

Results: This article gets results from the final stages of evaluation. In this evaluation process, compared the performance of decision tree, random forest, Xgboost, k-neighbors, Adaboost and deep neural network with accuracy.

Conclusion: The number of selected features by PCA-based feature selection method improve the predictive performance based on accuracy of deep learning and most machine learning models for predicting readmission. The improvement of machine learning models depended on the specific choice of the prediction model, number of selected features, and "k" for k-fold validation.

Cite this paper as:
Farajollahi B, Mehmannavaz M, Mehrjoo H, Moghbeli F, Sayadi M. Predicting hospital readmission of diabetic patients using machine learning. Front Health Inform. 2021; 10: 74. DOI: 10.30699/fhi.v10i1.266

INTRODUCTION

Diabetes is a chronic condition associated with abnormally high levels of sugar (glucose) in the blood [1]. It can be suffered by everyone and until now there is no cure for it [2]. The number of people with diabetes continues to rise [3], this is predicted to increase to 629 million by the end of 2045 [1]. Diabetes could cause many kinds of complications, which also leads to a high rate of repeated admission of patients with diabetes [3]. Hospital readmission is an episode when a patient who had been discharged from a hospital is admitted again within a specified usually short time interval [4]. Within healthcare recent focus has been on 30 days readmission which is when a patient is being readmitted to the hospital within 30 days after being discharged from the hospital [5]. Some of readmissions are avoidable although requires evidence-based treatments [6]. However, poor quality of care and/or ineffective transitions of care play a significant role [7]. As the quality of inpatient care is associated with early readmission, prediction of hospital readmission for diabetes patient would prove invaluable [1]. Hospital readmission is an indicator of the quality of care and is a driver for increasing cost of healthcare [8]. The United States (US) health system endures significant economic burden for diabetes care. This cost reached about $327 billion in 2017 [7]. The ability to predict patient readmissions will ultimately help the hospital to calculate and manage the quality of patient care [9]. Studying data over time can play a role in

predicting the magnitude and frequency of future events [10]. Machine learning (ML) plays a vital role in many predicting tasks. Hence, predicting hospital readmissions using ML sounds a worth implementing approach [11]. In this paper, we applied ML and a deep learning to the diabetes dataset with the aim of predicting readmission of diabetic patients within 30-days based on features in dataset. The classifier used include decision tree, random forest, Xgboost, k-Neighbors, Adaboost and deep neural network.

MATERIAL AND METHODS

The data used in the study are data obtained from the UCI machine learning repository about diabetic patients. The dataset used for this study contains 100,000 instances and it include 55 features from 130 hospitals in the United States for 10 years (1999-2008), and it has missing value. The significant attributes contributing to the analysis are the number of lab procedures, number of medications, age, insulin, number of outpatients, time in hospital, number of diagnoses, number of emergency and number of procedures.

Information was extracted from the database for encounters that satisfied the following criteria:

(1) It is an inpatient encounter (a hospital admission).

(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

(3) The length of stay was at least 1 day and at most 14 days.

(4) Laboratory tests were performed during the encounter.

(5) Medications were administered during the encounter.

Machine learning models

We consider five traditional ML and a deep learning model. The models under consideration are decision tree, random forest (a tree-based ensemble model), K-nearest neighbors, Adaboost (boosting ensemble methods), Xterme gradient boosting (Xgboost), and deep neural networks (a deep learning method). We implemented all these models using the Scikit-learn library in Python programming language.

Medical data in the real world is noisy, inconsistent, and incomplete. So before building the prediction model, it is essential to preprocess the data efficiently and make it appropriate for predictive modelling [12]. Feature selection technique was implemented in this study to make the prediction more accurate.

Feature selection

Optimizing the performance of the classification algorithm model by feature selecting is an important part [9]. Feature selection, as a data preprocessing strategy, has been proven to be effective and efficient in preparing data (especially high-dimensional data) for various data mining and ML problems. The objectives of feature selection include building simpler and more comprehensible models, improving data mining performance, and preparing clean, understandable data [13]. We use the principal components analysis (PCA) for data reduction, which can not only greatly reduce the time of model learning while preserving the data implied information, but also eliminate data noise and data redundancy [14]. In this study PCA-based feature selection is applied to all 55 features. The different choices for reduced dimension are checked and three selected feature sets are 10 features, 15 features and 20 features. A classifier is then built using the selected feature sets as input predictor variables.

Statistical analysis

In order to avoid over-fitting, we evaluated the prediction accuracy of all models under consideration via k-fold cross-validation. In this study 10-fold and 5-fold cross validation are used to separate the data sets into two categories train and test. The benefit of using k-fold cross validation is testing the entire dataset. After making a training set in each k-fold a model is generated, which testing set will be tested based on the model. To test dataset decision tree, random forest, k-neighbors, Xgboost and Adaboost are used. The reason of using several methods is to compare the final result and increase the accuracy and find out the impact of k in k-fold cross validation. Fig 1 illustrates the architecture of methodology.

Deep learning

Deep learning, a subfield of ML, has seen a dramatic resurgence in the past 6 years, largely driven by increases in computational power and the availability of massive new datasets [15]. Deep learning combines advances in computing power and neural networks with many layers to learn complicated patterns in large amounts of data. It is an extension of classical neural network and uses more hidden layers so that the algorithms can handle complex data with various structures [16]. The main difference between traditional ML and deep learning algorithms is in the feature engineering which requires domain expertise and a time-consuming process. Deep learning algorithms involve automatic feature engineering, whereas we need to handcraft the features in traditional ML algorithms [16]. This study present a deep neural network (DNN) model to predict 30-day readmission in patients with diabetes at the time of discharge. In order to make sure that the model generalizes and does not over-fit, the data were split

two parts: 80% for training set and 20% for testing set then the validation set was taken as a random 30% of the training set.

RESULTS

This research gets results from the final stages of evaluation. In this evaluation process, compared the performance of decision tree, random forest, Xgboost, k-neighbors, Adaboost and deep neural network with accuracy. The next section represents the related work.

Comparison of the Performance of Algorithms

Fig 2 shows the accuracy of models based on the comparison by using 10 selected features in the performance. The accuracy of traditional ML algorithms, by 10-fold is more than 5-fold; which among them K-neighbors has the least accuracy in each k-fold and ensemble models (random forest and Adaboost) have the most accuracy. In comparison of the accuracy between deep learning and traditional ML models it is considered that, deep learning has the highest accuracy than other models.

Fig 3 presents the accuracy of models based on the comparison by using 15 selected features in the performance. With the exception of random forest, the accuracy of all traditional ML algorithms by the 10-fold are more than the 5-fold. The accuracy of decision tree in both k-fold (k=5, k=10) are very similar. In this model, in the comparison of the accuracy of deep learning with other models of traditional ML, it is observed that the accuracy of deep learning is higher than those models.

Fig 4 indicates the comparison of the accuracy of traditional ML models using by 20 selected features. With the exception of random forest, the accuracy of all traditional ML algorithms by the 10-fold are more than the 5-fold. In comparison of the accuracy between deep learning and traditional ML models it is considered that, deep learning has the highest accuracy than other models.

In the selected feature sets, the accuracy of traditional ML models by 10-fold is further than 5-fold; except for random forest amongst which k-neighbor has the lowest accuracy results in each k-fold. In order to better assess, the deep learning model has been compared with the traditional ML models in each selected feature sets. Therefore, it can be seen that the best model of prediction was deep learning.

Fig 5 demonstrates the comparison of the accuracy of traditional ML models among three selected feature sets by 5-fold. The accuracy of random forest, k-neighbors and Adaboost clearly increases when the number of selected features expands; whereas the accuracy patterns of decision tree and Xgboost do not grow continuously.

Fig 6 displays the comparison of the accuracy of traditional ML models among three selected feature sets by 10-fold. The accuracy of decision tree, k-neighbors and Adaboost clearly increase when the number of selected features enlarges; whereas the accuracy patterns of random forest and Xgboost do not grow constantly.

Fig 7 pointed out the comparison of deep learning accuracy among three selected feature sets. By increasing the number of features, the accuracy of deep learning will grow. The highest accuracy 86.88% has been achieved by using deep learning with 20 selected features among all other conditions.

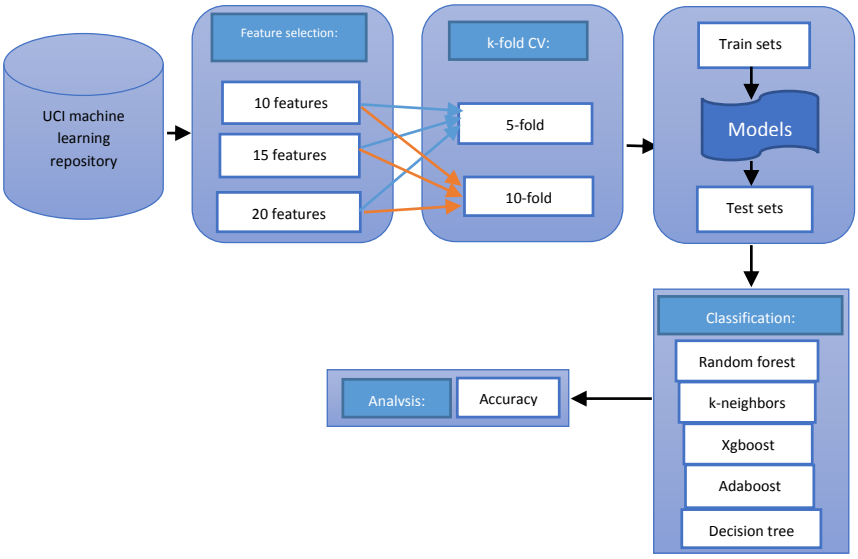

Fig 1: The architecture of methodology

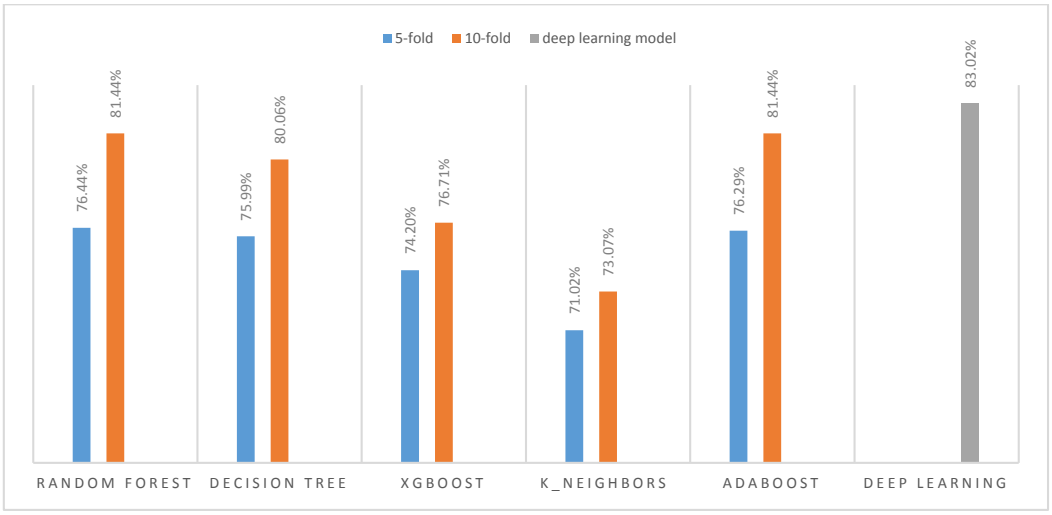

Fig 2: Comparison of the accuracy of models using 10 selected features

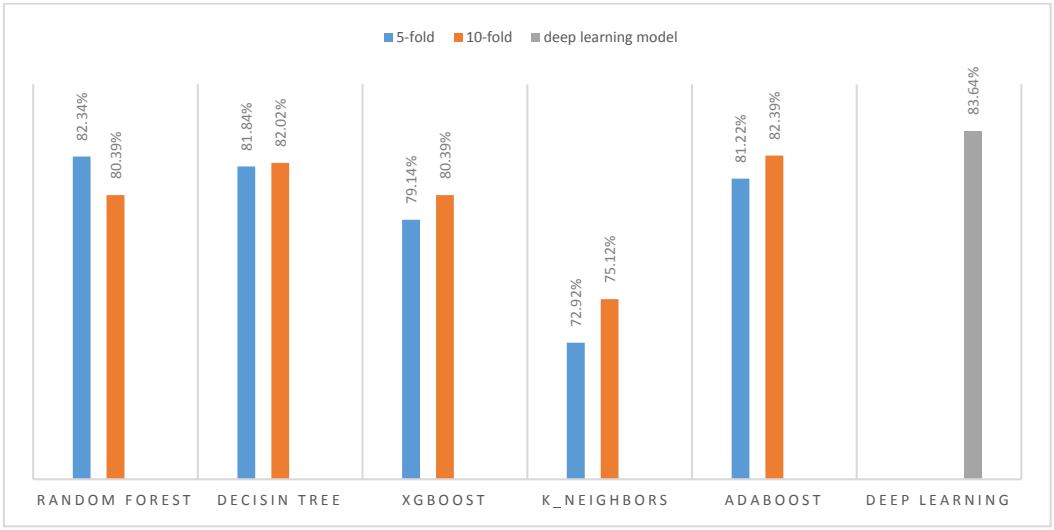

Fig 3: Comparison of the accuracy of models using 15 selected features

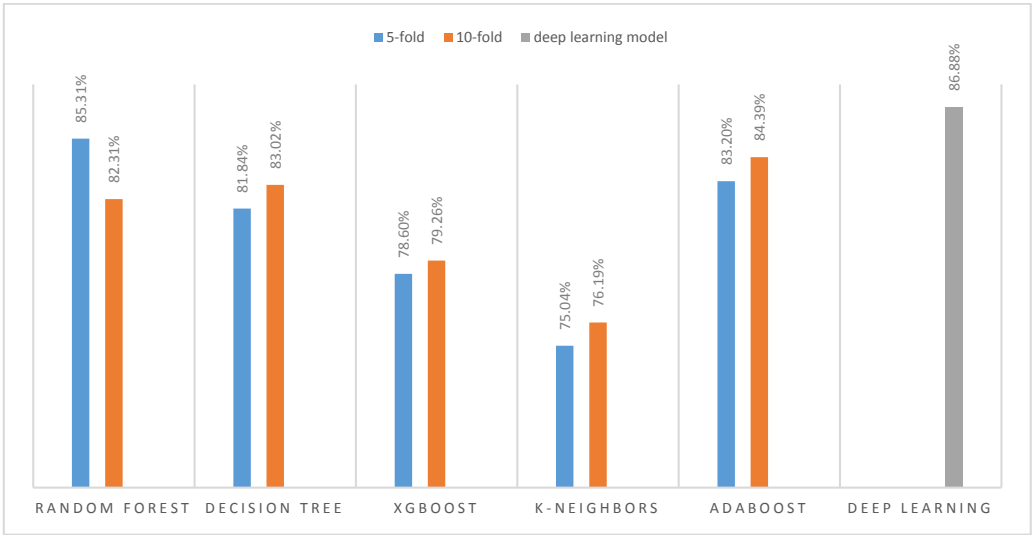

Fig 4: Comparison of the accuracy of models using 20 selected features

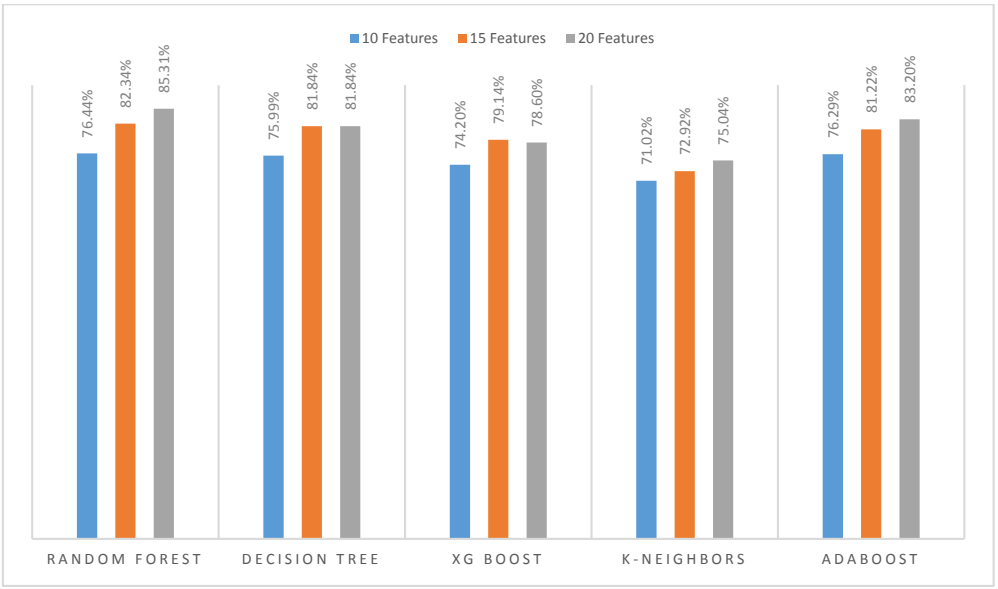

Fig 5: Comparison of the accuracy of models by 5-fold

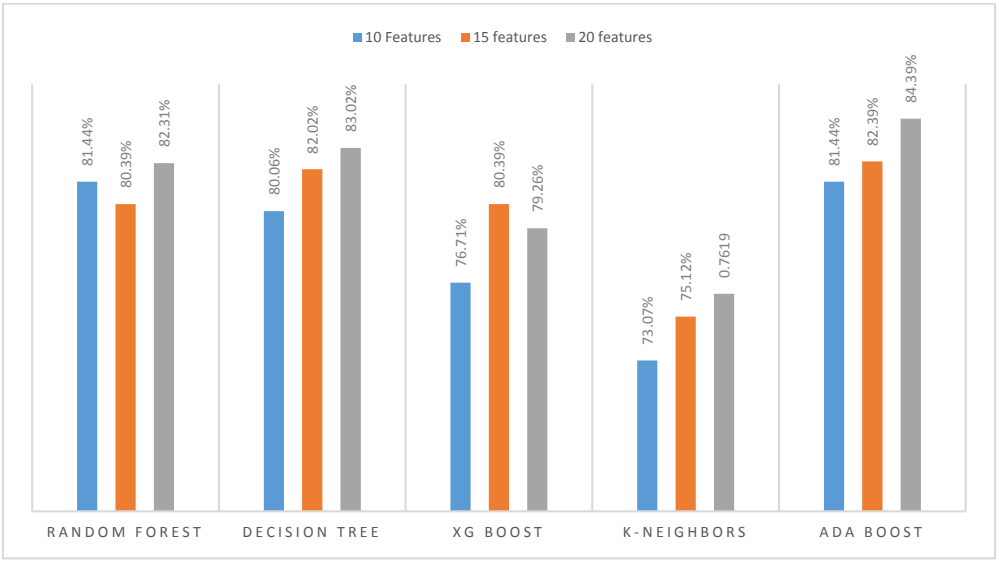

Fig 6: Comparison of the accuracy of models by 10-fold

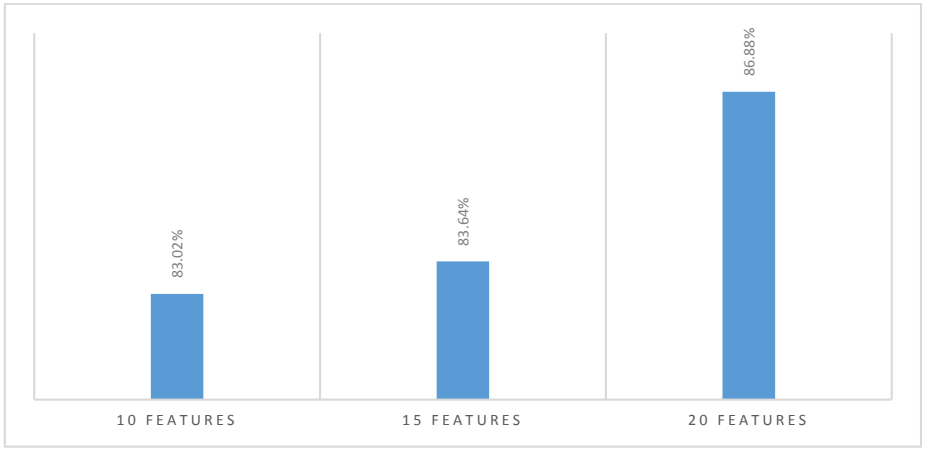

Fig 7: Comparison of the accuracy of deep learning model

DISCUSSION

In the study which has been proved by Alloghani et al. the probable predictors of diabetes hospital readmission, among those hospitals which used ML techniques along with other exploratory methods, has been explored. The classifiers which they used were included Random Forest, Linear Discriminant Analysis, Naïve Bayes, k–Nearest Neighbor, J48 and Support vector machine. The dataset contains 55 attributes which only 18 of those features were used as per the scope of the study. The conventional confusion matrix and ROC efficiency analysis were used to evaluate the performance of the models. The final readmission model is based on the best performing model as per the true positive rates, specificity and sensitivity. In order to validate a model and to improve the total accuracy, the 10-fold cross validation method which were applied for estimating, were lied. Eventually, the most responsive and effective model for classifying, learning, and predicting readmission rates which uses mHealth data, is Naïve Bayes; which has the most area of ROC and which is the most efficient model [6].

Sharma et al have published a project named "Prediction on Diabetes Patient's Hospital Readmission Rates". The purpose of this study is to develop a model which can predict 30-day hospital readmission. The researchers used ML algorithms including random forest, decision tree, logistic regression, Adaboost and Xgboost for prediction. Among all other algorithms, they have achieved the highest accuracy 94% by using Random forest model [17].

CONCLUSION

We applied ML models include decision tree, random forest, Xgboost, k-neighbors, Adaboost and deep neural network for prediction and achieved highest accuracy using deep neural network. Larger number of selected features by PCA-based feature selection method can improve the predictive performance based on accuracy of deep learning and most ML models for predicting readmission. However, the improvement of ML models depended on the specific choice of the prediction model, number of selected features, and "k" for k-fold validation.

AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

1. Pham HN, Chatterjee A, Narasimhan B, Lee CW, Jha DK, Wong EYF, et al. Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis. International Conference on System Science and Engineering. IEEE; 2019.
2. Tamin F, Iswari NMS. Implementation of C4.5 algorithm to determine hospital readmission rate of diabetes patient. International Conference on New Media Studies. IEEE; 2017.
3. Hu P, Li S, Huang Y-a, Hu L. Predicting hospital readmission of diabetics using deep forest. International Conference on Healthcare Informatics. IEEE; 2019.
4. Artetxe A, Beristain A, Grana M. Predictive models for hospital readmission risk: A systematic review of methods. Comput Methods Programs Biomed. 2018; 164: 49-64. PMID: 30195431 DOI: 10.1016/j.cmpb.2018.06.006 [PubMed]
5. Forsman R, Jönsson J. Artificial intelligence and machine learning: A diabetic readmission study [BSc Project]. Kristianstad University, Sweden; 2019.
6. Alloghani M, Aljaaf A, Hussain A, Baker T, Mustafina J, Al-Jumeily D, et al. Implementation of machine learning algorithms to create diabetic patient re-admission profiles. BMC Med Inform Decis Mak. 2019; 19(Suppl 9): 253. PMID: 31830980 DOI: 10.1186/s12911-019-0990-x [PubMed]
7. Alamer AA, Patanwala AE, Aldayyen AM, Fazel MT. Validation and comparison of two 30-day re-admission prediction models in patients with diabetes. Endocr Prac. 2019; 25(11): 1151-7. PMID: 31414904 DOI: 10.4158/EP-2019-0125 [PubMed]
8. Bojja R, El-Gayar O. Predicting hospital readmissions of diabetic patients: A machine learning approach. Annual Research Symposium. Dakota State University; 2019.
9. Pujianto U, Setiawan AL, Rosyid HA, Salah AMM. Comparison of naïve bayes algorithm and decision tree C4.5 for hospital readmission diabetes patients using hba1c measurement. Knowledge Engineering and Data Science. 2019; 2(2): 58-71.
10. Sayadi M, Moghbeli F, Mehrjoo H, Mahaki M. A linear study of the spread of covid19 in China and Iran. Frontiers in Health Informatics. 2020; 9(1): 32.
11. Hammoudeh A, Al-Naymat G, Ghannam I, Obied N. Predicting hospital readmission among diabetics using deep learning. Procedia Computer Science. 2018; 141: 484-9.

12. Duggal R, Shukla S, Chandra S, Shukla B, Khatri SK. Predictive risk modelling for early hospital readmission of patients with diabetes in India. International Journal of Diabetes in Developing Countries. 2016; 36(4): 519-28.

13. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. ACM Computing Surveys. 2017; 50(6): 1-45.

14. Jia M, Tian F. Readmission prediction of diabetic based on convolutional neural networks. International Conference on Computer and Communications. IEEE; 2019.

15. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019; 25(1): 24-9. PMID: 30617335 DOI: 10.1038/s41591-018-0316-z [PubMed]

16. Muniasamy A, Tabassam S, Hussain MA, Sultana H, Muniasamy V, Bhatnagar R. Deep learning for predictive analytics in healthcare. International Conference on Advanced Machine Learning Technologies and Applications. Springer; 2019.

17. Sharma A, Agrawal P, Madaan V, Goyal S. Prediction on diabetes patient's hospital readmission rates. Proceedings of the Third International Conference on Advanced Informatics for Computing Research. ACM. 2019; 1: 1-5.