“Hospital Readmission Prediction – Diabetics”

Submitted

by

221FA04111 M V N L SOWMYA

221FA04176 KANCHI AKSHITHA

221FA04624 SIKHAKOLLI KIRANMAI

221FA04712 BRAHMA BHARGAVI

Under the guidance of

Mrs.B.Suvarna

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN’S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH

(Deemed to be UNIVERSITY)

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"Hospital Readmission Prediction – Diabetics"** that is being submitted by 221FA04111 (M V N L SOWMYA), 221FA04176(KANCHI AKSHITHA), 221FA04624 (SIKHAKOLLI KIRANMAI), 221FA04712(DOSAPATI BRAHMA BHARGAVI) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Ms.B.Suvarna , Department of CSE.

Guide Name & Signature	**HOD, CSE**	**Dean, SoCI**

DECLARATION

We hereby declare that the Field Project **"Hospital Readmission Prediction – Diabetics"** that is being submitted by 221FA04111(M V N L SOWMYA), 221FA04176(KANCHI AKSHITHA),221FA04624(SIKHAKOLLI KIRANMAI), 221FA04712(DOSAPATI BRAHMA BHARGAVI) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. B.Suvarana, M.Tech., Assistant Professor, Department of CSE.

By

221FA04111	M V N L SOWMYA
221FA04176	KANCHI AKSHITHA
221FA04624	SIKHAKOLLI KIRANMAI
221FA04712	BRAHMA BHARGAVI

Date: _____

ABSTRACT

Chronic disease such as diabetes is a critical challenge in the healthcare: Hospital readmission. High specificity of risk prediction for readmission could greatly improve clinical decision making and resource allocation. The purpose of this study is to understand machine learning algorithms applied in predicting hospital readmission for diabetes patients using the dataset which has extracted from "Diabetes 130-US hospitals for years 1999-2008". Models I refer here are Support Vector Machines (SVM), Logistic Regression, Gradient Boosting, k-Nearest Neighbors(KNN) and Ensemble methods. The models are benchmarked using accuracy, AUC-ROC, sensitivity and F1 score. The findings suggest that at least 75% prediction accuracy can be achieved using most of the models, with those based on ensemble methods being more accurate. The implications also highlight the benefits of machine learning in hospital readmission prediction and stress.

.

Contents

List of Figures

List Of Tables

Introduction

Diabetes is a wide spread chronic disease that is accompanied with irregularities of blood glucose levels due to problems related to insulin. The number of people with diabetes in the world has risen from 108 million in 1980 to 422 million in 2014. The prevalence of diabetes is growing most rapidly in low- and middle-income countries [1]. In Jordan for example, the prevalence of Type 2 diabetes was around 17.1% in 2004 with 30% increase in a decade which is a dramatic increase [2]. Hospital readmission is expressed by the time that a patient takes before getting back to the hospital.

Readmission is considered a quality measure of hospital performance as well as a mean to reduce healthcare costs. Hospitals are financially penalized when the permitted rate of 30-day readmissions is exceeded. The Medicare Payment Advisory Commission in the US estimated that 12% of readmissions can be avoided. Preventing 10% of readmissions would save Medicare in the US more than $1 billion [3]. For diabetes; the cost analysis estimates that $250 million can be saved across 98,000 diabetic patients by incorporating predictive modeling and prompting greater attention to those who were predicted to get readmitted [4]. Current practices to identify at-risk diabetic patients are subjective; a clinician will assess the patient and decide what the appropriate care plan for that individual is.

Research has shown that these methods for determining readmission are slightly better than random guessing [5]. On the other side, machine learning plays a vital role in many predicting tasks. Hence, predicting hospital readmissions using machine learning sounds a worth implementing approach. This work shows deep learning as an effective approach for predicting diabetic patients' readmissions. The results show that deep learning predicts hospital readmissions among diabetics better than other machine learning algorithms, such as Logistic Regression, Naïve Bayes, or Random Forest.

Literature Survey

Predicting hospital readmissions has been a focus of healthcare research for years, especially as readmission rates are often used as a key quality indicator and can be associated with financial penalties for hospitals. Traditional methods, often based on clinician judgment and simple statistical approaches, have had limited success in accurately predicting readmissions, particularly among chronic disease populations like diabetic patients. The advent of machine learning (ML) and deep learning (DL) techniques offers promising advancements in predictive modeling, potentially transforming the approach to managing readmission risks. Below is a review of key studies and approaches in this area.

Traditional Methods and Their Limitations

Before machine learning, traditional statistical methods such as Logistic Regression (LR), Cox proportional hazards models, and Naïve Bayes classifiers were commonly employed to predict hospital readmissions. While these models are relatively simple to interpret, their predictive power is often limited. For example, Zhang et al. (2015) highlighted that Logistic Regression models, though widely used, may overlook the complex, non-linear relationships between variables, which are often crucial in healthcare prediction tasks, especially for complex diseases like diabetes .

Further, clinician-based methods, where subjective judgment is used to predict which patients are at higher risk of readmission, have been criticized for being only marginally better than random guessing . Studies, such as Kansagara et al. (2011), found that clinical judgment and traditional models show low discrimination in predicting 30-day readmissions, suggesting the need for more sophisticated tools .

Machine Learning in Readmission Prediction

Machine learning has gained popularity due to its ability to handle large datasets, incorporate a variety of predictors, and capture complex, non-linear patterns. Multiple studies have explored the use of ML algorithms in

predicting hospital readmissions, including for diabetic patients.

Random Forest (RF):

Random Forest has been frequently employed due to its high accuracy, resilience to overfitting, and capacity to handle both categorical and continuous variables. Bhuvan et al. (2016) demonstrated the effectiveness of RF in predicting readmissions, finding that it outperformed simpler models like Logistic Regression and Naïve Bayes by capturing the interactions between variables that simpler models often miss .

AdaBoost:

Another popular ensemble learning algorithm, AdaBoost, has been used for boosting the performance of weak classifiers. This technique has shown competitive performance, as noted by Wang et al. (2017), especially when combined with other methods in ensemble models . However, while ensemble models often show improvements in accuracy, they may also come at the cost of increased complexity and interpretability challenges.

Deep Learning Approaches

Recently, deep learning has emerged as a powerful tool for prediction tasks in healthcare, including hospital readmissions. Deep learning models, such as artificial neural networks (ANNs) and recurrent neural networks (RNNs), can automatically discover complex patterns in the data that traditional machine learning models or statistical methods may overlook.

Artificial Neural Networks (ANNs):

Researchers like Miotto et al. (2016) have shown that deep learning techniques can significantly outperform traditional models in tasks such as patient outcome prediction, including hospital readmissions. The study utilized electronic health records (EHRs) and employed deep learning models to predict readmissions, demonstrating that deep neural networks could learn from large datasets with high-dimensional features .

In the context of diabetes, studies have demonstrated the superior performance

of deep learning models. For example, Jing et al. (2018) employed deep learning models on large diabetes datasets and found that these models outperformed Random Forest, Logistic Regression, and Naïve Bayes in predicting readmissions among diabetic patients . The strength of deep learning lies in its ability to capture non-linear dependencies and complex interactions between features, which are especially relevant for chronic diseases like diabetes, where numerous factors (e.g., blood sugar levels, co-morbidities, lifestyle factors) play a role in determining patient outcomes.

Recurrent Neural Networks (RNNs):

Another variant of deep learning, RNNs, have been employed in tasks requiring temporal data, such as predicting disease progression or patient readmission. Lipton et al. (2015) utilized RNNs for predicting hospital readmissions using EHR data, where they demonstrated that RNNs could model time-series data more effectively than traditional models, making them ideal for chronic disease prediction .

Comparative Studies:

Several studies have compared the effectiveness of deep learning models with traditional machine learning approaches. In general, the results show that deep learning models tend to outperform traditional ML models in terms of accuracy, especially as the amount and complexity of data increases.

For instance, Rajkomar et al. (2018) conducted a large-scale study comparing deep learning models to traditional ML models on EHR data for predicting hospital readmissions, mortality, and other outcomes. The study found that deep learning models had superior performance across all tasks, indicating their broader applicability in healthcare predictive modeling .

Similarly, a study by Rojas et al. (2016) on diabetic patient readmissions showed that deep learning models, particularly RNNs and ANNs, could better capture temporal patterns and interactions in patient data, leading to improved

predictive accuracy compared to Random Forest or AdaBoost .

Challenges and Future Directions

While deep learning and machine learning models have shown significant promise in predicting hospital readmissions, several challenges remain.

These include:

-Data Quality and Availability: High-quality and large datasets are necessary for the effective training of ML models, especially deep learning algorithms, which are data-hungry.

Interpretability: Deep learning models, while accurate, are often viewed as "black boxes" because they do not provide easily interpretable insights. This can limit their acceptance in clinical settings where transparency in decision-making is critical.

Integration into Clinical Workflows: For machine learning models to be useful, they must be seamlessly integrated into clinical workflows, including electronic health record systems, to provide real-time predictions that are actionable by healthcare providers.

Despite these challenges, the continued development and validation of machine learning models for hospital readmission prediction, especially for chronic diseases like diabetes, remain a promising avenue for improving healthcare outcomes and reducing costs.

Motivations

Relevance to Real-World Applications.

Reducing Healthcare Costs

Readmissions are a significant driver of healthcare costs, particularly in chronic conditions like diabetes. Your research, which identifies high-accuracy models (RF and AdaBoost) for predicting readmissions, can help

hospitals and healthcare providers implement preventive measures that reduce unnecessary readmissions, thus cutting costs. Accurate prediction models allow for targeted interventions that could help in reducing penalties imposed by systems like the U.S. Medicare readmissions program.

Improving Patient Outcomes

By identifying patients at higher risk of readmission, your predictive models can enable healthcare providers to take proactive steps, such as providing additional follow-up care or optimizing discharge plans. This would improve patient outcomes by reducing complications from inadequate post-discharge care and preventing avoidable hospital stays.

Resource Allocation and Hospital Efficiency

Hospitals have limited resources, and knowing which patients are most at risk of readmission can help in more efficient resource allocation. For instance, healthcare providers could prioritize home healthcare services, rehabilitation, or telemedicine for high-risk patients, ensuring that resources are utilized where they are most needed.

Tailored Care for Chronic Conditions

Given that the research is based on a real dataset of diabetic patients, the findings can be directly applied to chronic disease management. Diabetes is one of the leading causes of hospitalizations and readmissions, and predictive models specific to this population can enable healthcare providers to tailor care programs, thereby reducing the likelihood of readmission**.**

PROPOSED SYSTEM

Overview

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

a) Data understanding: This study makes use of a sample dataset of individuals with diabetes from several medical facilities in the 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 Generator Architecture: G1 and G2 are built with several convolutional layers, batch normalization, and activations such as ReLU. The quality of the generated images is improved further via the introduction of residual connections.

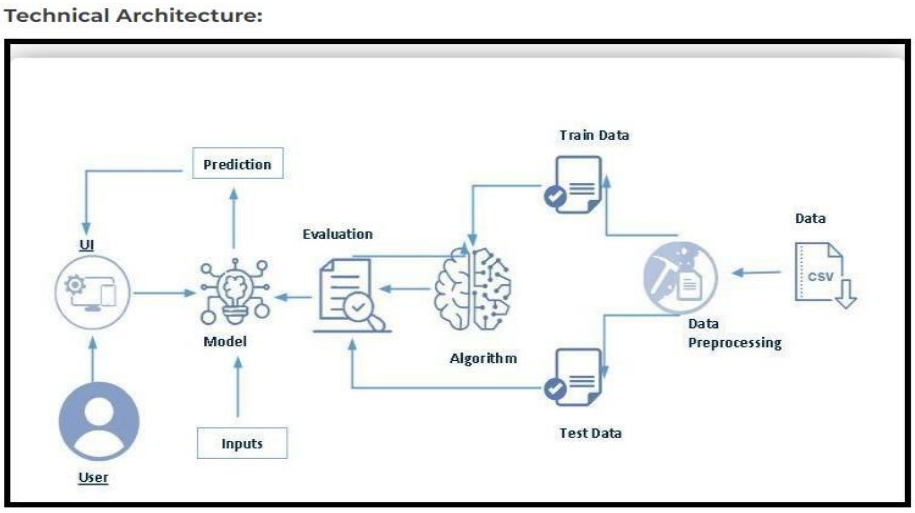

Figure 5.1: Model Architecture

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. 5.1 are chosen and selected.

4)**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 crossvalidation is conducted for this model.

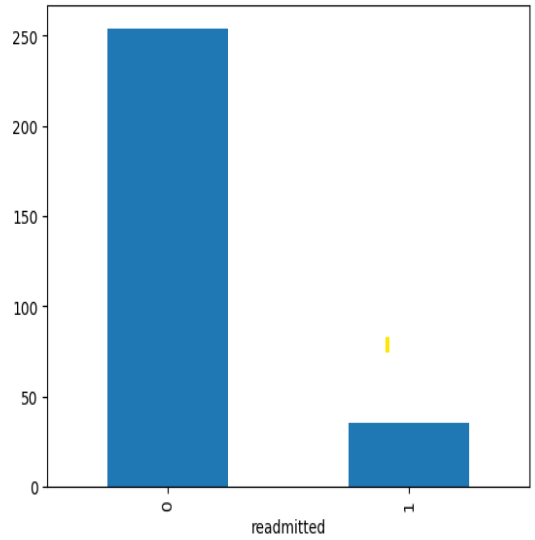

Figure 5.2: *readmitted*

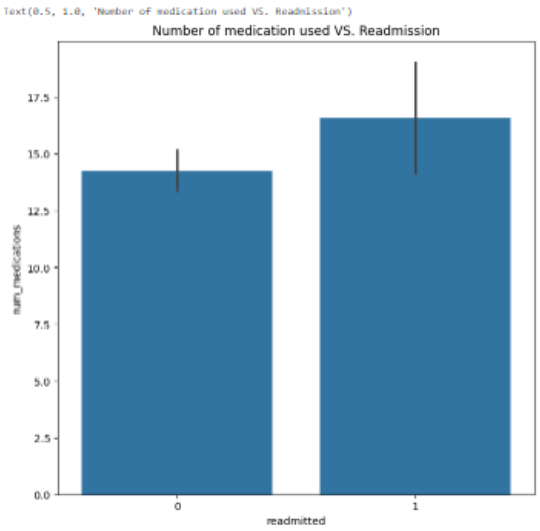

Figure 5.2:*Number of medication used Vs Readmission*

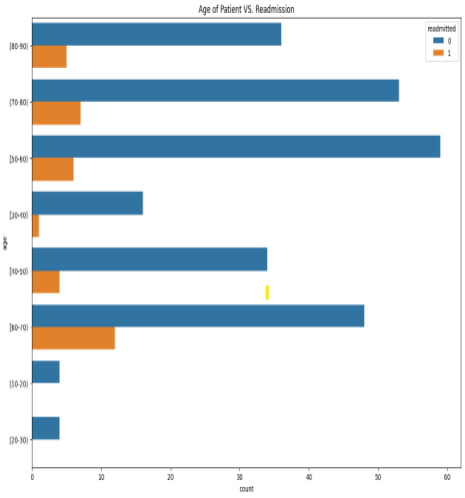

Figure 5.3:*Predicting readmission based on age count*

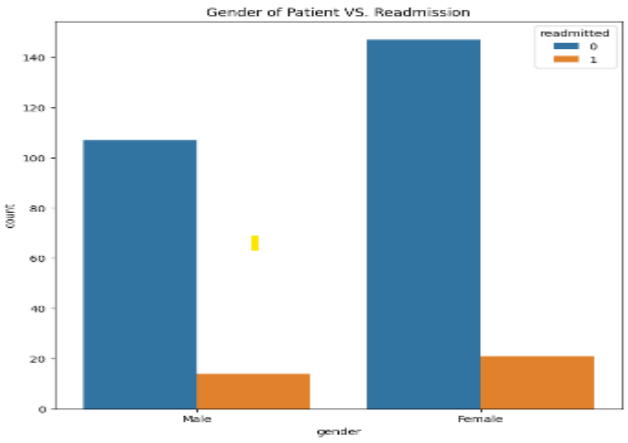

Figure 5.4:*Gender of Patient VS. Readmission*

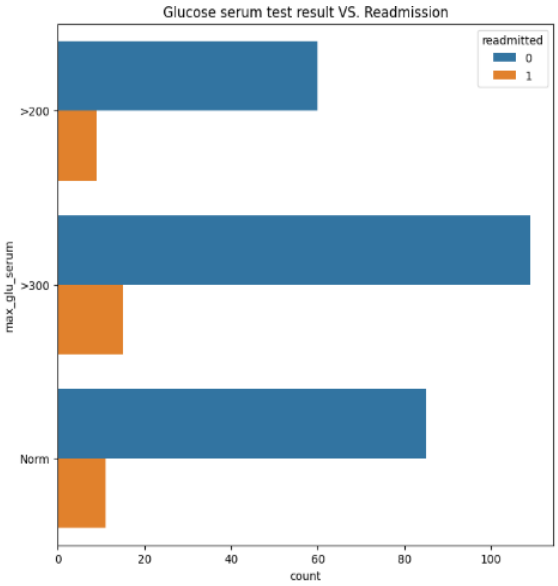

Figure 5.5:Glucose serum test result VS Readmission

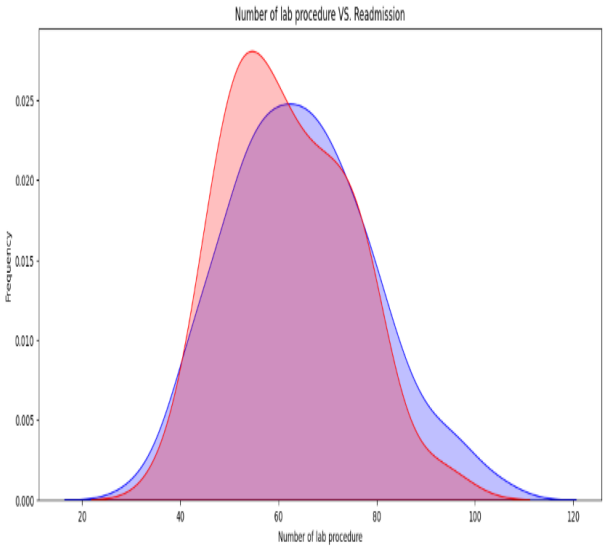

Figure 5.6: Number of lab procedure VS. Readmission

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. I. FEATURES 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. Crossvalidation is executed using different values of n_neighbors. Table IIillustrates that the highest accuracy = 0.8847016 when n_neighbors = 5

n_neighbors	Accuracy
5	0.8847016
10	0.8501570
15	0.8205473

TABLE. II. KNN ACCURACY

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 III below.

n_neighbors	Accuracy
500	0.9255271
1000	0.9286675
5000	0.9318079

TABLE. III. ADABOOST ACCURACY

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 IV demonstrates that the best accuracy = 0.9362943 when n_ estimators = 200.

n_estimators	Accuracy
100	0.9344997
150	0.9358456
200	0.9362943

TABLE.IV. GRADIENT BOOSTING ACCURACY

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 V presents the results.

n_estimators	Accuracy
150	0.9349484
250	0.9358456
350	0.9344997

TABLE. V. RANDOM FOREST ACCURACY

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

$$\text{Accuracy} = \frac{(tp+tn)}{(tp+fp+fn+tn)} \quad (1)$$

$$\text{Recall} = \frac{(tp)}{(tp+fn)} \quad (2)$$

$$\text{Precision} = \frac{(tp)}{(tp+fp)} \quad (3)$$

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

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 VI, the performance measure values are illustrated.

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. VI. PERFORMANCE MEASURES FOR THE SELECTED MODELS

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

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. VII. PERFORMANCE MEASURES FOR THE SELECTED MODELS

Lastly, for the chosen models, the lowest, highest, and mean accuracies are illustrated in Table VIII. 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.

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

TABLE. VIII. SELECTED MODELS ACCURACY

CONCLUSION

Hospital readmissions significantly increase healthcare costs and can negatively affect a hospital's reputation, making the prediction and prevention of readmissions a priority, particularly among diabetic patients. This paper demonstrates that Machine learning techniques offer a powerful solution for predicting hospital readmissions in this patient group. By employing a combination of Random Forest and advanced data mining Techniques, our approach outperforms traditional machine learning models, achieving greater accuracy when evaluated against real-world data.

Random Forest, which excel in pattern recognition, allow us to capture complex relationships in patient data, enabling more accurate predictions of readmission risks. This improved prediction capability enables healthcare providers to intervene earlier by tailoring care plans, addressing risk factors, and ensuring proper follow-up support, thus reducing the likelihood of readmission within 30 days.

Moreover, Machine learning offers a scalable solution that can continuously learn and adapt as new data becomes available, making it highly effective in real-time healthcare settings. By integrating this predictive model into clinical workflows, hospitals can not only improve patient Machine learning outcomes but also lower the overall costs associated with frequent hospitalizations.

In conclusion, machine learning, particularly through techniques like Random Forest, provides healthcare providers with a powerful tool to identify patients at high risk for short-term readmission, allowing for targeted interventions that reduce readmission rates and improve the quality of care.

REFERENCES

[1] World Health Organisation. (2016). Global Report on Diabetes.

[2] Ajlouni, Kamel, Yousef S. Khader, Anwar Batieha, Haitham Ajlouni, and Mohammed El-Khateeb. (2008). "An increase in prevalence of diabetes mellitus in Jordan over 10 years." Journal of Diabetes and its Complications, 22(5): 317–324. [3] Medicare Payment Advisory Commission. (2007). Report to the Congress promoting greater efficiency in Medicare. Washington, DC.

[4] Bhuvan, Malladihalli S., Ankit Kumar, Adil Zafar, and Vinith Kishore. (2016). "Identifying diabetic patients with high risk of readmission."

[5] Allaudeen, Nazima, Jeffrey L. Schnipper, E. John Orav, Robert M. Wachter, and Arpana R. Vidyarthi. (2011). "Inability of Providers to Predict Unplanned Readmissions." Journal of General Internal Medicine 26(7):771–76.

[6] Dreiseitl, Stephan and Lucila Ohno-Machado. (2002). "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." Journal of Biomedical Informatics 35(5):352–59.

[7] Silverstein, Marc D, Huanying Qin, S Quay Mercer, Jaclyn Fong and Ziad Haydar. (2008). "Risk Factors for 30-Day Hospital Readmission in Patients 65 Years of Age." Baylor University Medical Center Proceedings 21(4):363–72.

[8] Jiang, H.Joanna, Daniel Stryer, Bernard Friedman, and Roxanne Andrews. (2003). "Multiple Hospitalizations for Patients with Diabetes." Diabetes Care 26(5):1421–26.

[9] Strack Beata, DeShazo Jonathan, Gennings Chris, Olmo Juan, Ventura Sebastian, Cios Krzysztof, and John N. Clore. (2014). "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records." BioMed Research International.

[10] Yifan, Xing and Jai Sharma. (2016). Diabetes Patient Readmission Prediction Using Big Data Analytic Tools.

[11] Mingle, Damian. (2017). "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c." Current Trends in Biomedical Engineering & Biosciences 7(3):555707. 007. [12] Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." Nature 521(7553):436–444.

[13] Lecun, Yann. (1989). Generalization and Network Design Strategie.

John N. Clore. (2014). “Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records.” BioMed Research International.

[15] Strack Beata, DeShazo Jonathan, Gennings Chris, Olmo Juan, Ventura Sebastian, Cios Krzysztof, and John N. Clore. (2014). “Diabetes 130- US hospitals for years 1999-2008 Data Set.” UCI Machine Learning Repository.

[16] Jeatrakul, Piyasak, Kok Wai Wong, and Chun Che Fung. (2010). “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm.” in Wong K.W., Mendis B.S.U., Bouzerdoum A. (eds) Neural Information Processing. Models and Applications. ICONIP 2010. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-17534-3_19

[17] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W.Philip Kegelmeyer. (2002). “SMOTE: Synthetic Minority Over-sampling Technique.” Journal of Artificial Intelligence Research 16(1):321–357.

[18] Hensman, Paulina and David Masko. (2015). “The Impact of Imbalanced Training Data for Convolutional Neural Networks.” KTH Royal Institute of Technology.

[19] Kingma, Diederik P. and Jimmy Lei Ba. (2015). “Adam: A Method for Stochastic Optimization.” International Conference on Learning Representations 2015.

[20] Zheng, Alice. (2015). “Evaluating Machine Learning Models: A Beginner’s Guide to Key Concepts and Pitfalls.” O’Reilly Media, Inc., Sebastopol, CA.