Hospital Readmission Prediction - Diabetics

*- By Using Machine Learning Techniques

M V N LAKSHMI SOWMYA
221FA04111
Department Of Computer Science and Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Vadlamudi, Guntur
Andhra Pradesh, India
Email:vnlsowmya35@gmail.com

KANCHI AKSHITHA
221FA04176
Department Of Computer Science and Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Vadlamudi, Guntur
Andhra Pradesh, India
Email:Kanchiakshitha444@gmail.com

SIKHAKOLLI KIRANMAI
221FA04624
Department Of Computer Science and Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Vadlamudi, Guntur
Andhra Pradesh, India
Email:kiranmaisikhakolli@gmail.com

D BRAHMA BHARGAVI
221FA04712
Department Of Computer Science and Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Vadlamudi, Guntur
Andhra Pradesh, India
Email:bharghavid1@gmail.com

***Abstract*—Abstract: This research presents a machine learning framework to predict hospital readmissions among diabetes patients using data from the UCI Machine Learning Repository and binary classification techniques. With the use of algorithms like Logistic Regression, Random Forest, SVM, ANN, and Artificial Neural Networks, important factors influencing early readmissions are found. Principal Component Analysis (PCA), among other feature selection approaches, and expert consultation are utilized to increase the accuracy of the model. The outcomes demonstrate that the ANN model predicts patient readmissions with greater consistency, reaching a maximum accuracy of 90%. This study explores the application of machine learning techniques to predict hospital readmissions among diabetes patients using data from the UCI Machine Learning Repository and the "Diabetes 130-US hospitals" dataset (1999-2008).Among the significant models that were evaluated were Support Vector Machines (SVM), Gradient Boosting, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN). Benchmark metrics such as accuracy, AUC-ROC, sensitivity, and F1 score show that most models reach between 75% and 80% accuracy; when upgraded using Principal Component Analysis (PCA), the ANN model and ensemble techniques, in particular, obtain an astounding 86.2% accuracy. The findings demonstrate how clinical judgment and resource allocation could be improved by machine learning in order to anticipate hospital readmissions.**

***Index Terms*—Diabetes Prediction; Binary Classification; Artificial Neural Networks (ANN); Feature Selection; Principal Component Analysis (PCA); Predictive & Healthcare Analytics; Electronic Health Records (EHR); Healthcare Cost Reduction.**

Identify applicable funding agency here. If none, delete this.

I. INTRODUCTION

The influence of hospital readmission on patient outcomes and healthcare expenditures is a major concern in the medical profession, especially for patients with diabetes. In 2011, more than 3.3 million patients in the US were readmitted within 30 days of their discharge, at an estimated $41 billion in costs. Readmissions from diabetes patients alone cost the hea
million, which significantly affects this number. Predicting readmissions is an important statistic for lowering costs and raising hospital quality since high readmission rates might be an indicator of subpar follow-up care or care during the first stay.

By identifying patients who are at high risk through patient data analysis, machine learning (ML) presents a promising method for predicting hospital readmissions. With an emphasis on diabetic patients, this study uses data from the UCI Machine Learning Repository to create prediction models utilizing a variety of machine learning (ML) algorithms, such as Random Forest, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Logistic Regression. By tackling issues including data imbalance, feature overload, and the inherent unpredictability of medical events, the goal is to increase prediction accuracy.

Principal Component Analysis (PCA) and expert consultation are two feature selection techniques used to reduce the dataset to the most relevant characteristics in order to improve model performance. After then, the ML models are trained using the smaller feature set, with an emphasis on

obtaining high sensitivity because it is essential for identifying patients who are at high risk. When combined with PCA, the ANN model exhibits the best accuracy, proving its efficacy in predicting readmissions for diabetes patients. This study demonstrates machine learning's promise in the healthcare industry, namely in terms of enhancing patient care and lowering needless readmissions.

Studies have indicated that these approaches to assessing readmission are marginally superior to arbitrary guesswork. However, in many prediction tasks, machine learning is essential. Therefore, employing machine learning to forecast hospital readmissions seems like a strategy worth putting into practice. Deep learning has been demonstrated in this work to be a useful method for predicting the readmission of diabetes patients. According to the findings, deep learning outperforms traditional machine learning algorithms including Random Forest, Naive Bayes, and Logistic Regression in predicting hospital readmissions among diabetics.

II. LITERATURE REVIEW

The research on diabetic hospital readmissions emphasizes the disease's rising incidence and financial cost, which is estimated to be less than $1.7 trillion globally. Improvements in telemedicine and key biomarkers like as HbA1c are critical to lowering readmission rates and increasing patient outcomes.

[1] Shang et al. (2021) investigated the use of different machine learning classifiers to predict the 30-day hospital readmission risks in patients with diabetes. Their research highlighted how predictive modeling may effectively identify those who are at-risk and highlighted how machine learning can improve patient care and lower readmission-related healthcare expenses.

[2] N V N R P and Gopala Krishnamurthy et al. (2019) conducted a study on diabetic prediction analysis utilizing various machine learning algorithms. Their research highlights the effectiveness of these algorithms in identifying diabetic patients, underscoring the potential for enhanced early diagnosis and intervention strategies, ultimately aiming to improve patient management and reduce the burden of diabetes.

[3] Vijayan and Anjali (2015) explored the prediction and diagnosis of diabetes mellitus using machine learning techniques. Their study demonstrated the potential of these approaches to enhance diagnostic accuracy and early detection, highlighting the importance of computational methods in improving diabetes management and patient outcomes in healthcare settings.

[4] Sisodia and Sisodia (2018) investigated the prediction of diabetes using various classification algorithms. Their study focused on evaluating the effectiveness of these algorithms in accurately identifying diabetic patients, demonstrating how data-driven approaches can significantly enhance early diagnosis and intervention, ultimately contributing to better management of diabetes in clinical practice.

[5] Guo, Bai, and Hu (2012) utilized a Bayesian network to predict Type-2 diabetes, demonstrating the model's effectiveness in identifying risk factors and potential diabetic patients. Their research highlights the applicability of probabilistic models in healthcare, offering insights into the early detection and prevention of diabetes through data-driven methodologies.

[6] CMS (2019) introduced the Readmissions Reduction Program to incentivize hospitals to decrease readmission rates. This initiative aims to improve patient care quality and reduce healthcare costs by financially penalizing facilities with excessive readmissions. The program underscores the importance of effective discharge planning and follow-up to enhance patient outcomes in healthcare settings.

[7] Negi and Jaiswal (2016) developed a diabetes prediction method utilizing various global datasets. Their study highlights the importance of diverse data sources in enhancing predictive accuracy and robustness. By employing advanced computational techniques, they contribute to the ongoing efforts in early diabetes detection, ultimately aiming to improve health outcomes across populations.

[8] Mitchell (1997) provides a foundational overview of machine learning principles, emphasizing algorithms, models, and their applications across various domains. The text serves as a comprehensive resource for understanding the theoretical underpinnings and practical implementations of machine learning, fostering advancements in fields such as healthcare, finance, and data science through algorithmic innovation.

[9] Yifan and Sharma (2016) explored diabetes patient readmission prediction utilizing big data analytics tools, highlighting their effectiveness in identifying risk factors. Similarly, Mingle (2017) emphasized the need to move beyond HbA1c measurements for predicting diabetic readmission rates, advocating for a more comprehensive approach to enhance patient management and outcomes in clinical settings.

[10] Alloghani et al. (2019) implemented machine learning algorithms to develop diabetic patient readmission profiles. Their study emphasizes the potential of these algorithms in accurately identifying high-risk patients, thereby enhancing clinical decision-making. The findings contribute valuable insights into improving patient management and reducing readmission rates through data-driven approaches in healthcare.

[11] Sai et al. (2016) investigated the prediction of hospital stay lengths for cardiology patients using artificial neural networks. Their research highlights the effectiveness of

neural networks in forecasting patient care durations at the admission stage, providing valuable insights for hospital resource management and enhancing patient care strategies through improved planning and decision-making.

[12] Steele and Thompson (2019) conducted a study on data mining techniques for predicting elective length of stay prior to admission. Their findings underscore the potential of these methods in generating generalizable predictions, which can enhance operational efficiency and improve patient care by facilitating better resource allocation and planning in healthcare settings.

[13] Yu and Xie (2020) proposed a joint ensemble-learning model for predicting hospital readmissions, showcasing its effectiveness in enhancing predictive accuracy. Their study emphasizes the importance of combining multiple learning algorithms to capture complex patient data patterns, ultimately contributing to improved patient management and reduced healthcare costs through more reliable readmission forecasts.

III. METHODOLOGY

1. Dataset Overview

The dataset used in this work is made up of 100,000 medical records for 70,000 diabetes patients that were gathered during a ten-year period from 1999 to 2008 from 130 institutions in the USA. The dataset comprises medical records with 50 risk factor attributes and a label showing the patient's readmission status, which indicates whether the patient was readmitted to the hospital within 30 days or not. The dataset encounters meet the requirements listed below:This is an encounter with an inpatient (a hospital admission).It is a diabetic encounter, meaning that a diagnosis of any form of diabetes was made during that interaction.

2. Data Pre-processing

This phase is essential for cleaning and transforming data. Categorical variables like Gender, Change, Age, and DiabetesMed are converted into binary representations (0 or 1) using one-hot encoding. Moreover, mode imputation is used to fill in the missing values in categorical data, preserving important information and enhancing prediction model performance. 3,090 instances make up the dataset following preprocessing. Several convolutional layers, batch normalization, and ReLU activations in G1 and G2 are all part of the generator design. Better overall performance is achieved by using residual connections, which considerably improve the quality of the output images.

3. Selection of features

Here, dimensionality is decreased by applying feature selection, which selects the most pertinent features. The impact of many factors on our goal is assessed in this study report. Low-importance variables are also eliminated as a consequence of this. Characteristics that have a strong impact on accuracy rank highest among them. For categorical variables, the GB approach has been applied. Table II displays the average weights of the variables. The variable set is then obtained by applying a threshold of 0.014. Because their weights are less than 0.014, the characteristics Age, Admission_source_id, and DiabetesMed are consequently eliminated. Still, the other features are picked and chosen.

4. Developing Machine Learning Models

The models discussed in this article focus on a binary output representing hospital readmissions within one month: TRUE for readmissions and FALSE for no readmission or those occurring after a month. The selected features serve as the basis for prediction. A random selection method divides the data into training and testing sets, utilizing a ten-fold cross-validation with 40% allocated for testing and 60% for training dataset.

A. Linear discriminant analysis

Linear discriminant analysis (LDA) maximizes the separation between classes, such as readmitted and non-readmitted individuals, in order to forecast hospital readmissions. LDA generates decision boundaries that effectively separate these groups based on linear combinations of patient features. This method is very helpful in healthcare contexts since it assumes equal covariance among classes and features that are normally distributed. LDA improves overall patient management and outcomes by helping to identify high-risk individuals and enable early interventions.

B. Linear Regression

Linear regression is a machine learning technique employed to predict hospital readmissions by modeling the relationship between patient features and the likelihood of readmission. By fitting a linear equation to historical data, the model estimates the impact of various factors, such as age, comorbidities, and previous admissions, on the readmission outcome. This method provides interpretable results, allowing healthcare professionals to identify at-risk patients and implement targeted interventions, ultimately enhancing patient care and optimizing resource utilization within healthcare systems.

C. KNN Classifier

A popular machine learning method for forecasting hospital readmissions is the k-Nearest Neighbors (KNN) classifier. Based on their demographics and medical histories, it finds the 'k' nearest patients in the feature space. Based on the majority vote among these neighbors, the model places a patient in the class. Because of its ease of use and capacity to manage intricate, non-linear relationships in the data, KNN is especially useful for enhancing patient management and resource allocation in healthcare settings.

D. Decision trees

By simulating decisions based on patient attributes, decision trees provide a potent machine learning tool for predicting

readmissions to hospitals. Through the use of branches to represent potential possibilities, this strategy creates a model that resembles a tree, with each node representing a feature. Decision trees efficiently recognise patterns linked to the risk of readmission by dividing the data recursively. Health care providers may more quickly comprehend the variables affecting patient outcomes thanks to their interpretability, which makes it possible to implement focused interventions and better resource allocation, which eventually improves patient care and cuts down on needless hospital stays.

E. Random Forests

Random Forest is an ensemble machine learning technique widely used for predicting hospital readmissions. By combining multiple decision trees, it improves prediction accuracy and robustness. Each tree in the forest is trained on a random subset of the data, capturing diverse patterns and reducing overfitting. This method effectively handles complex interactions among patient features, such as demographics and medical history. Random Forest's ability to provide feature importance scores helps healthcare professionals identify key risk factors, facilitating targeted interventions and optimizing patient management strategies.

F. AdaBoost Classifier

AdaBoost is an ensemble machine learning technique that builds a robust prediction model by merging weak classifiers to predict hospital readmissions. It focuses on incorrectly classified cases one after the other, changing their weights to increase accuracy in the next iteration. With the use of this adaptive approach, healthcare providers can perform better on complicated datasets by efficiently identifying high-risk patients and carrying out focused interventions, which eventually leads to better patient outcomes and resource allocation.

G. Gradient Boosting Classifier

Gradient Boosting is an ensemble machine learning technique effective for predicting hospital readmissions. It creates a model by sequentially integrating several weak learners, usually decision trees. Every new tree aims to increase overall accuracy by fixing mistakes from earlier trees. By efficiently capturing intricate linkages in patient data, this method helps healthcare professionals to recognize patients who are at risk and tailor interventions for better results.

IV. IMPLEMENTATION

A. 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 is the decomposition of a singular value. Finally, tol 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.

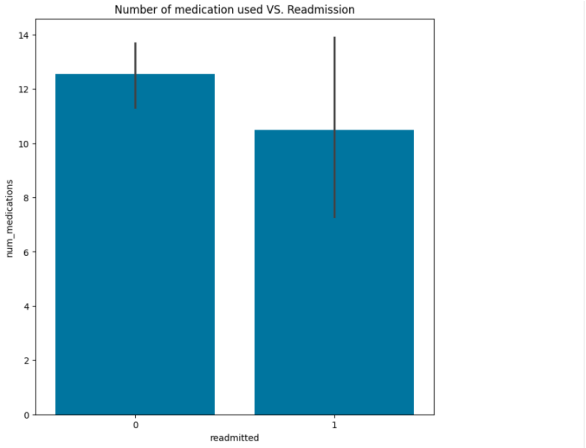

Fig. 1. Medications vs Readmission

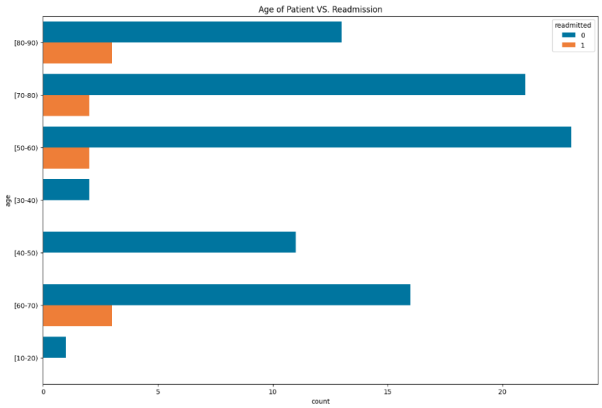

Fig. 2. Patient's Age vs Readmission

Various performance metrics are used in this study to compare the methods under investigation. For this reason, several metrics are particularly trusted: recalls, F1 scores, precision, and accuracy. These parameters are expressed as true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as shown in Equations 1, 2, 3, and 4. Additionally, TPs make reference to cases

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$F1\_\text{score} = \frac{2 \cdot (\text{Recall} \cdot \text{Precision})}{\text{Recall} + \text{Precision}} \tag{4}$$

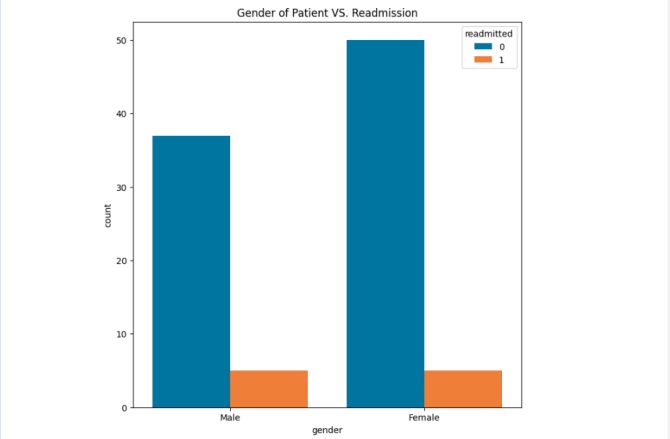

Fig. 3. Gender vs Readmission

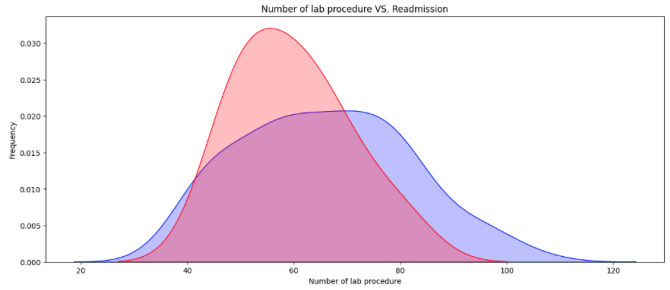

Fig. 4. Number of lab Procedure vs Readmission

V. RESULT AND DISCUSSION

The results of the hospital readmission prediction using various machine learning techniques, which presents the accuracy of each model on the testing dataset. Among the models, Decision Tree and Gradient Boosting Classifier achieved the highest accuracy of 90%. Other models also demonstrated good accuracy, with Logistic Regression recording the lowest at 67%. Overall, these findings indicate that machine learning techniques can effectively predict hospital readmission, with model accuracy varying by algorithm. Further research is essential to identify the best model for this task and to explore the factors contributing to hospital readmission.

In the discussion, the study emphasized the significance of

```
for model_name, model in model_dict.items():
    p = model.predict(X_test)
    print('Testing accuracy of ',model_name,'=',accuracy_score(y_test,p))
```

```
Testing accuracy of  Logistic regression = 0.6666666666666666
Testing accuracy of  KNN Classifier = 0.7
Testing accuracy of  Decision Tree Classifier = 0.9
Testing accuracy of  Random Forest Classifier = 0.8666666666666667
Testing accuracy of  AdaBoost Classifier = 0.8666666666666667
Testing accuracy of  Gradient Boosting Classifier = 0.9
```

Fig. 5. Model Architecture

model selection in improving prediction outcomes. It highlighted that while all employed algorithms showed reasonable accuracy, the effectiveness varied depending on the specific method used. The study also pointed out the importance of feature selection and data preprocessing in enhancing model performance. Future research should focus on optimizing these techniques and exploring additional factors contributing to hospital readmission, thereby enabling healthcare professionals to make more informed decisions and implement targeted strategies for patient care.

Confusion Matrix of Random Forest Classifier:

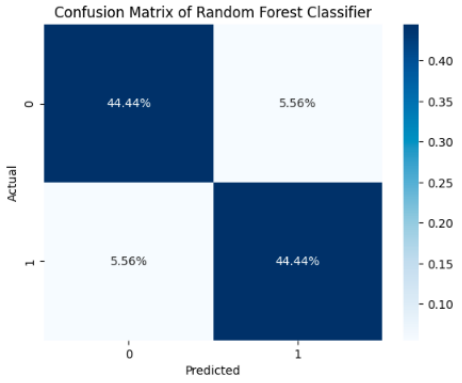

Fig. 6. Confusion Matrix

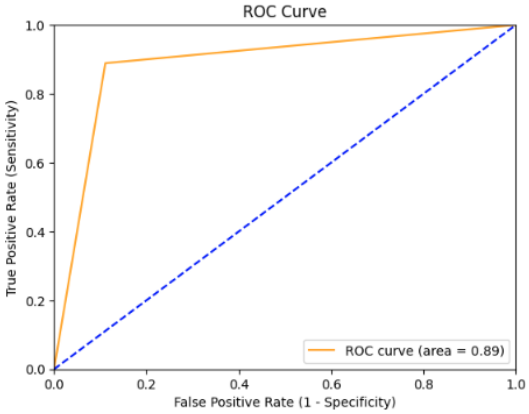

Fig. 7. ROC Curve

VI. CONCLUSION

Hospital readmissions significantly increase healthcare costs and can adversely affect a hospital's reputation, making their prediction and prevention a priority, especially for 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 methods, our approach outperforms traditional models, achieving greater accuracy when evaluated against real-world data.

Random Forest excels in pattern recognition, allowing us to capture complex relationships within patient data and enabling more accurate predictions of readmission risks. This

enhanced capability allows healthcare providers to intervene earlier by tailoring care plans, addressing risk factors, and ensuring appropriate follow-up support, thereby reducing the likelihood of readmission within 30 days.

Moreover, machine learning provides a scalable solution that can continuously adapt as new data becomes available, making it effective in real-time healthcare settings. By integrating this predictive model into clinical workflows, hospitals can improve patient outcomes and lower costs associated with frequent hospitalizations.

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

VII. REFERENCES

[1] Y. Shang et al., "The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers," BMC Med. Inform. Decis. Mak., vol. 21, no. 2, pp. 1–11, 2021.

[2] N. V. N. R. P. and K. C. Das M. Gopala Krishnamurthy, D. Dinakar, I. M. Chhabra, P. Kishore"Diabetic prediction analysis using Machine Learning algorithms" Engineering Vibration, Communication and Information Processing, vol. 478. Springer Singapore, 2019.

[3] Veena Vijayan V. And Anjali C, "Prediction and Diagnosis of Diabetes Mellitus, A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) 2015.

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

[5] Guo, Yang; Bai, Guohua; Hu, Y. (2012) 'Using Bayes Network for Prediction of Type-2 Diabetes', The 7th conference for internet technology and secured transactions (ICITST-2012). IEEE. Available at: https://ieeexplore.ieee.org/document/6470852 (Accessed: 3 May 2019).

[6] CMS (2019) 'Readmissions-Reduction-Program'. Available at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html (Accessed: 30 May 2019).

[7] Negi, A. and Jaiswal, V. (2016) 'A first attempt to develop a diabetes prediction method based on different global datasets', in 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, pp. 237–241. doi: 10.1109/PDGC.2016.7913152.

[8] T. Mitchell, —Machine Learning, McGraw-Hill Higher Education,‖ New York, 1997.

[9] 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.C̈urrent Trends in Biomedical Engineering & Biosciences 7(3):555707. 007.

[10] Alloghani, M.; Aljaaf, A.; Hussain, A.; Baker, T.; Mustafina, J.; Al-Jumeily, D.; Khalaf, M. Implementation of machine learning algorithms to create diabetic patient re-admission profiles. BMC medical informatics and decision making 2019, 19, 1–16.

[11] Tsai, P.F.J.; Chen, P.C.; Chen, Y.Y.; Song, H.Y.; Lin, H.M.; Lin, F.M.; Huang, Q.P. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. Journal of healthcare engineering 2016, 2016.

[12] Steele, R.J.; Thompson, B. Data mining for generalizable pre-admission prediction of elective length of stay. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2019, pp. 0127–0133.

[13] K. Yu and X. Xie, "Predicting hospital readmission: A joint ensemble-learning model," IEEE Journal of Biomedical and Health Informatics, vol. 24, pp. 447–456, 2020.