

Hospital Readmission With Machine learning : A Comprehensive Survey and Analysis

Krupa Savant¹, Bhagatsingh Jitkar²

¹MTech student, Department of Computer Science and Engineering, D.Y. Patil College of Engineering & Technology, Kolhapur, India.

²Asst. Prof., Department of Computer Science and Engineering, D.Y. Patil College of Engineering & Technology, Kolhapur, India.

Abstract

Hospital readmissions is a significant concern in healthcare, leading to increased costs and poorer patient outcomes. Predicting which patients are at high risk of readmission allows healthcare providers to take preventative measures, improving patient care and reducing the strain on medical resources. In this project, we design a machine learning model to predict hospital readmissions within a specific time frame (e.g., 30 days post-discharge) using patient data, medical history, and treatment information. The goal of this project is to implement and compare three mostly used machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. These algorithms will be trained on a dataset consisting of demographic data, clinical measurements, treatment details, and previous readmission history. We will evaluate and compare their performance using various metrics, like accuracy, precision, recall, and F1-score, to determine which model provides the most accurate predictions. Through this analysis, we aim to identify key factors contributing to hospital readmissions and provide healthcare providers with a predictive tool that can help optimize patient care, reduce readmission rates, and lower healthcare costs. The project also explores the interpretability of the models, with a particular focus on Random Forest due to its ability to provide feature importance, making it a valuable tool for understanding the factors influencing readmission risks. In conclusion, the use of machine learning techniques in predicting hospital readmissions has the potential to improve the efficiency of healthcare systems and provide data-driven insights for better patient management.

Keywords: Hospital Readmission Prediction, Machine Learning, K-Nearest Neighbor, Random Forest, Decision Tree, Artificial Neural Network. Patient Data Analysis, Healthcare Optimization, Predictive Modeling, Feature Importance.

1. INTRODUCTION

Hospital readmissions, particularly within a short period after discharge, pose a critical challenge to healthcare systems worldwide. They contribute to higher healthcare costs, overburdened medical facilities, and poorer patient outcomes. Identifying patients at risk of readmission has become a key focus in improving the quality of healthcare services and reducing unnecessary hospitalizations. Accurate prediction of hospital readmissions can lead to timely interventions, improved patient care, and optimized healthcare resource utilization.

Traditionally, predicting patient readmissions has been a manual, data-intensive process, relying on clinical expertise and statistical models. However, with the rise of machine learning, it is now possible to leverage vast amounts of patient data to make more accurate and automated predictions. Machine learning models can uncover patterns and relationships within data that may not be evident through conventional analysis. By training on historical data, these models can learn to predict which patients are at the highest risk of readmission.

In this project, we explore the use of three machine learning algorithms—K-Nearest Neighbors (KNN), Decision Tree, and Random Forest—to predict hospital readmissions. Each of these

algorithms has its own strengths:

K-Nearest Neighbors (KNN) is a easy and intuitive algorithm that classifies a new data point depend on the majority class of its nearest neighbors.

Decision Tree is a flowchart-like structure that models decisions and their potential outcomes, providing an interpretable model of how patient factors influence readmissions.

Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions, often leading to improved accuracy and robustness.

The goal of this project is to analyze patient data, including demographic, clinical, and treatment-related variables, to develop models that can accurately predict whether a patient is likely to be readmitted. By comparing the performance of KNN, Decision Tree, and Random Forest, we aim to determine which model provides the most reliable predictions and insights into the factors driving readmissions.

Machine learning techniques, when applied effectively, can transform the healthcare industry by predicting hospital readmissions and offering data-driven recommendations for personalized patient care. This project contributes to the growing field of predictive healthcare analytics, offering potential solutions to enhance the quality of care, reduce costs, and improve result of patient outcomes

2. Literature Review

Hospital Hospital readmission is a significant issue that affects the quality and efficiency of healthcare delivery. Numerous studies have explored methods for predicting hospital readmissions, with many focusing on machine learning approaches due to their ability to process large datasets and uncover complex relationships between variables. This literature review highlights key studies and methodologies relevant to hospital readmission prediction using machine learning techniques, specifically K-Nearest Neighbors (KNN), Decision Trees, and Random Forests.

[1] Aksa Urooj, Md Tabrez Nafis , Mobin Ahmad, "An Analytical and Comparative Study of Hospital Re-admissions in Digital Health Care", Computer Networks, Big Data and IoT, vol.66, pp.717, 2021:

Medical or Hospital re-admissions are costly and reflect poor hospital quality. A high re-admission rate can have significant financial repercussions for both patients and healthcare institutions. With the growing use of electronic health records, a vast amount of data is now available for analysis, helping to identify high-risk patients and potentially lowering mortality rates. Under the 2010 Affordable Care Act, hospitals may face penalties if patients are re-admitted within 30 days of discharge. However, many hospitals argue that re-admissions are often influenced by the populations they serve. This paper reviews various techniques previously employed to predict the likelihood of patient re-admissions after discharge. Additionally, it provides an overview of the socioeconomic and demographic factors that play crucial roles in medical re-admissions and discusses strategies that could help reduce these rates.

[2] Inés Robles Mendo, Gonçalo Marques, Isabel de la Torre Díez, Miguel López- Coronado, Francisco Martín-Rodríguez, "Machine Learning in Medical Emergencies: a Systematic Review and Analysis", Journal of Medical Systems, vol.45, no.10, 2021:

Due to the rising demand for artificial intelligence (AI) research in the field of medicine, the practical applications of these methods in healthcare emergencies remain ambiguous. To address this gap, the authors conducted a systematic review and global overview study to identify, analyze, and evaluate existing research across various platforms, particularly focusing on their implementation in healthcare emergencies. The methodology for selecting and identifying relevant studies and applications involved two approaches. First, the PRISMA methodology was applied to databases including Google Scholar, PubMed, IEEE Xplore, Scopus, and Science Direct. Second, the authors reviewed commercial applications available on well-known platforms like Android and iOS. This review included a total of 20 studies, with the majority focusing on clinical decision-making (n = 4, 20%) or

medical and emergency services ($n = 4$, 20%). Only 10% ($n = 2$) of the studies centered around mobile health (m-health). In addition, 12 apps were fully tested on different devices, addressing either pre-hospital medical care ($n = 3$, 25%) or clinical decision support ($n = 3$, 25%). Notably, half of these applications utilized machine learning techniques, particularly based on natural language processing. Machine learning is becoming more prevalent in healthcare, providing solutions to enhance the quality and efficiency of healthcare services. With the rise of mobile health devices and real-time health assessment applications, machine learning is a growing trend in the healthcare industry.

[3] Lo, Y. T., Liao, J. C., Chen, M. H., Chang, C. M., & Li, C. T. (2021). "Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms". *BMC medical informatics and decision making*, 21(1), 1-11:

Unplanned early hospital readmissions are linked to increased patient harm, higher medical expenses, and a negative impact on the hospital's reputation. Identifying at-risk patients is a critical step toward improving care, as it allows for the implementation of appropriate interventions to prevent readmission. This study aimed to develop machine learning models to predict 14-day unplanned readmissions. A retrospective cohort study was conducted on 37,091 adult patients with 55,933 discharges from an 1193-bed university hospital between September 1, 2018, and August 31, 2019. Patients who were under 20 years of age, admitted for cancer-related treatment, participating in clinical trials, discharged against medical advice, deceased during admission, or residing abroad were excluded from the study. The analysis included predictors across 7 variable categories extracted from the hospital's medical record dataset. Four machine learning algorithms—logistic regression, random forest, extreme gradient boosting, and categorical boosting—were used to build prediction models for 14-day unplanned readmissions. The performance of these models was assessed using precision, recall, F1-score, the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPRC).

[4] Morel, D., Kalvin, C. Y., Liu-Ferrara, A., Caceres-Suriel, A. J., Kurtz, S. G., & Tabak, Y. P. (2020). "Predicting hospital readmission in patients with mental or substance use disorders". a machine learning approach. *International Journal of Medical Informatics*, 139, 104136:

High hospital readmission rates have become a significant challenge for mental health services, as they are directly tied to the quality of patient care. By developing predictive models using machine learning algorithms, it is possible to assess the readmission risk in hospitals. This paper aims to predict the readmission risk for patients with schizophrenia in a specific region of Spain by employing machine learning techniques. The study utilized a dataset comprising 6089 electronic admission records from 3065 patients diagnosed with schizophrenia disorders, with data collected between 2005 and 2015 from acute units in 11 public hospitals across the region. The Random Forest classifier yielded the best results for predicting readmission risk, achieving metrics of accuracy=0.817, recall=0.887, F1-score=0.877, and AUC=0.879. This research identifies the algorithm with the highest accuracy and determines the factors contributing to the readmission risk among patients with schizophrenia in this population. Additionally, it demonstrates that predictive models developed through machine learning can enhance patient care quality and support the creation of preventive treatment plans.

[5] Min, X., Yu, B., & Wang, F. (2019). "Predictive modeling of the hospital readmission risk from patients" claims data using machine learning: a case study on COPD. *Scientific reports*, 9(1), 2362:

Chronic Obstructive Pulmonary Disease (COPD) is a widespread chronic lung condition, impacting hundreds of millions of individuals globally. Many patients with COPD are readmitted to the hospital within 30 days of discharge due to various factors, though such readmissions can often be prevented if appropriate attention is given to those at high risk. This emphasizes the importance of early prediction of hospital readmission risk. The objective of this paper is to systematically develop and assess various machine learning models, both deep learning and traditional methods, for predicting the readmission risk of COPD patients. We test these models on a real-world medical claims database comprising records from 111,992 patients treated at the Geisinger Health System between January

2004 and September 2015. The models are based on both knowledge-driven features—extracted based on clinical insights potentially related to COPD readmission—and data-driven features, derived from the patient data itself.

[6] Molly K. Bailey, Audrey J. Weiss, Marguerite L. Barrett and H. Joanna Jiang, "Characteristics of 30-Day All-Cause Hospital Readmissions 2010–2016", Statistical Brief No. 248 hcup-us.ahrq.gov, February 2019:

This HCUP Statistical Brief presents trends on 30-day all-cause readmissions from 2010 through 2016 using the HCUP Nationwide Readmissions Database (NRD). Trends and changes in readmissions by expected payer are provided. The rate of readmissions and a comparison of costs for the index admission (the initial inpatient stay) and the readmission in 2016 is also provided by the type of principal diagnosis

[7] Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). "Application of machine learning in predicting hospital readmissions" .A scoping review of the literature. BMC medical research methodology, 21(1), 1-14:

This review followed the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Extension for Scoping Reviews (PRISMA-ScR) Statement. The Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) framework was also used to guide item extraction. A systematic search of electronic databases including PUBMED, MEDLINE, and EMBASE was conducted, covering the period from January 1, 2015, to December 10, 2019. Articles were imported into the COVIDENCE online software for title/abstract screening and full-text review for eligibility. Only observational studies that applied machine learning techniques to hospital readmissions among US patients were eligible. Studies without full text available in English were excluded. The qualitative synthesis covered study characteristics, machine learning algorithms employed, and model validation methods, while the quantitative analysis focused on model performance. Performance in terms of Area Under the Curve (AUC) was analyzed using R software, and the Quality in Prognosis Studies (QUIPS) tool was employed to assess the quality of the included studies.

[8]L.-C. Hung, S.-F. Sung and Y.-H. Hu, "A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack", Applied Sciences, vol. 10, no. 18, pp. 6337, 2020:

Readmissions following a stroke are linked not only to increased disability and higher mortality risk but also to rising medical costs. Accurately predicting the risk of readmission and understanding the contributing factors is critical for efficient healthcare resource distribution and planning for quality improvement. This study aimed to develop models to predict the likelihood of readmission or mortality after stroke by applying machine learning techniques to data collected during the initial hospital admission. To ensure balanced class distribution, resampling methods were used, and a two-layer nested cross-validation approach was employed to create and assess the models. The analysis included data from 3,422 patients, with a 90-day readmission or mortality rate of 17.6%. Several key predictive factors were identified, such as age, prior emergency department visits, pre-stroke functional status, stroke severity, body mass index, consciousness level, and the use of a nasogastric tube. The Naïve Bayes model, utilizing class weighting to address class imbalance, demonstrated the highest predictive accuracy with an area under the receiver operating characteristic curve of 0.661. While there is room for further enhancement, these models could serve as tools for early risk assessment in stroke patients, enabling early discharge planning and interventions for transitional care.

Table1:-Summary of different techniques used for heart disease diagnosis

Technology	Ref.	Advantages	Limitations
Machine Learning in Digital Healthcare	[1]	Improved patient monitoring and analysis of hospital readmissions using big data and IoT.	Limited to specific healthcare datasets, may require advanced infrastructure to handle big data.
Machine Learning in Medical Emergencies	[2]	Efficient in medical emergencies, providing quick decision support, and systematic review of data analysis.	Models may not generalize well to all emergency situations, data bias can affect outcomes.
Predictive Modeling for Unplanned Readmissions,	[3]	Accurate 14-day readmission predictions using advanced machine learning algorithms	Requires large datasets for training and may not work well for smaller or less diverse patient populations.
Machine Learning for Mental Health Readmissions	[4]	Effective in predicting hospital readmission in patients with mental or substance use disorders.	Models may struggle with diverse patient mental health conditions and societal factors influencing readmissions.
Predictive Modeling using Claims Data	[5]	Machine learning on claims data for predicting readmission risk in chronic obstructive pulmonary disease (COPD) patients..	Claims data can be incomplete, leading to inaccuracies; models may miss clinical details not covered in claims data.
Statistical Analysis of 30-Day Readmissions	[6]	Statistical analysis provides detailed insights over long periods (2010-2016), highlighting trends in 30-day readmission rates.	Does not leverage advanced machine learning techniques, which can provide more personalized predictions.
Machine Learning in Predicting Readmissions	[7]	Comprehensive review on machine learning models applied to hospital readmissions, focusing on potential applications and best practices.	Lack of focus on specific diseases, general models may not perform well for specialized conditions or populations.
ML for Stroke or TIA Readmission Prediction	[8]	Effective prediction of stroke or transient ischemic attack readmissions using machine learning, showing good accuracy.	May not be easily generalizable to other diseases or conditions.

Medical diagnosis is considered as an important things that needs to be conducted in accurate and efficient manner. Recent advancements in explainable AI (XAI) and interpretable machine learning are beginning to address these issues, making machine learning models more transparent and useful in clinical decision-making. The integration of social determinants of health (e.g., housing, income, access to care) into predictive models is another promising avenue, as these factors have been shown to significantly influence readmission rates.

3. Methodology

Module 1: Importing the Dataset

The dataset module serves as the foundation for training and assessing predictive models. It includes essential patient-related variables such as age, gender, medical history, lab results, vital signs, and treatment details. This dataset plays a crucial role in building machine learning models for predicting hospital readmission.

Module 2: Data Preprocessing

This module prepares the dataset for machine learning by handling missing values, eliminating inconsistencies, normalizing features, and splitting the data into training and testing sets. Proper preprocessing ensures higher accuracy in predictions. Boxplots help detect and eliminate outliers by displaying the distribution of variables. Data points falling outside the whiskers indicate potential anomalies that need to be removed to avoid skewed results. Missing values are replaced with the mean of the corresponding variable to maintain statistical consistency and preserve data integrity.

Module 3: Feature Selection

Feature selection involves identifying the most relevant attributes from the dataset, reducing dimensionality, and enhancing model performance. Determines the significance of each variable in predicting hospital readmission. Methods such as correlation analysis and recursive feature elimination help refine the dataset. The dataset is divided into training and testing subsets to evaluate the model's ability to generalize new cases. Used to train the model. Assesses model performance on unseen data.

Module 4: Classification Algorithms

The classification module employs machine learning models to predict hospital readmissions based on labeled data. The key algorithms include:

A. k-Nearest Neighbors (KNN)

A widely used supervised learning algorithm that predicts readmission likelihood by finding the K-nearest data points based on distance metrics like Euclidean distance.

B. Random Forest

An ensemble learning approach that combines multiple decision trees to enhance predictive accuracy and stability. It identifies key interactions between patient characteristics affecting readmission.

C. Decision Tree

A tree-based supervised learning algorithm that divides data into smaller subsets based on significant predictor variables. It helps determine key factors influencing readmission.

Module 5: Performance Evaluation

To analyze the effectiveness of the model, the following metrics are utilized:

Accuracy: Measures the proportion of correctly predicted instances.

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

- **TP (True Positive):** Correctly predicted readmitted cases.
- **TN (True Negative):** Correctly predicted non-readmitted cases.
- **FP (False Positive):** Incorrectly predicted readmitted cases.
- **FN (False Negative):** Incorrectly predicted non-readmitted cases.

Precision: Evaluates the correctness of positive predictions.

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

Recall: Measures the proportion of actual positive cases correctly identified.

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

F1-Score: A balance between precision and recall, calculated as:

$$F1\text{-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

CONCLUSION

The prediction of hospital readmissions using machine learning techniques, specifically K-Nearest Neighbors (KNN), Decision Trees, and Random Forest, provides valuable insights into improving patient care and reducing unnecessary healthcare costs. These algorithms, by analyzing patient data from electronic health records (EHRs), help identify patients at higher risk of readmission, allowing for timely interventions and personalized treatment plans.

Among the algorithms evaluated, Random Forest typically delivers the most accurate and robust predictions due to its ability to handle large datasets, manage missing data, and reduce overfitting. Decision Trees, while highly interpretable and useful for understanding the most significant factors affecting readmission, are prone to overfitting in complex datasets. KNN, though simple and intuitive, performs best in scenarios with smaller datasets but may struggle with scalability in larger datasets.

This study demonstrates that machine learning models have the potential to significantly enhance predictive accuracy for hospital readmissions when compared to traditional statistical methods. By leveraging these models, healthcare providers can optimize resource allocation, reduce the risk of readmission, and improve patient outcomes. Future research should focus on further refining these models to enhance their generalizability, particularly across diverse patient populations, and integrating social determinants of health to capture a broader range of factors influencing readmission risk.

ACKNOWLEDGEMENTS

We express our heartfelt gratitude to everyone who contributed to the successful completion of this project, Hospital Readmission with Machine Learning: A Comprehensive Survey and Analysis.

First and foremost, we would like to extend our sincere appreciation to our institution and faculty members for their constant support, guidance, and encouragement throughout this research. Their valuable insights and constructive feedback have played a crucial role in shaping our work.

We would also like to acknowledge the contributions of researchers and scholars whose work provided a strong foundation for our study. Their extensive research and publications in the field of hospital readmission and machine learning have been instrumental in enhancing our understanding of the subject.

Our deepest gratitude goes to our mentors, whose expertise and advice have helped refine our approach and methodologies. Their unwavering support and insightful suggestions have significantly contributed to the quality and direction of this project.

We would also like to express our appreciation to our peers and colleagues for their valuable discussions, suggestions, and encouragement, which have helped us refine our analysis and improve the scope of our research.

Lastly, we extend our gratitude to our families and friends for their constant motivation and support throughout this journey. Their encouragement has been a driving force behind our dedication and commitment to this project.

This research is the result of collective efforts, and we sincerely thank everyone who played a role in its development.

REFERENCES

- [1] Aksha Urooj, Md Tabrez Nafis , Mobin Ahmad, "An Analytical and Comparative Study of Hospital Re-admissions in Digital Health Care", *Computer Networks, Big Data and IoT*, vol.66, pp.717, 2021.
- [2] Inés Robles Mendo, Gonçalo Marques, Isabel de la Torre Díez, Miguel López- Coronado, Francisco Martín-Rodríguez, "Machine Learning in Medical Emergencies: a Systematic Review and Analysis", *Journal of Medical Systems*, vol.45, no.10, 2021.
- [3] Lo, Y. T., Liao, J. C., Chen, M. H., Chang, C. M., & Li, C. T. (2021). "Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms". *BMC medical informatics and decision making*, 21(1), 1-11.
- [4] Morel, D., Kalvin, C. Y., Liu-Ferrara, A., Caceres-Suriel, A. J., Kurtz, S. G., & Tabak, Y. P. (2020). "Predicting hospital readmission in patients with mental or substance use disorders". a machine learning approach. *International Journal of Medical Informatics*, 139, 104136.
- [5] Min, X., Yu, B., & Wang, F. (2019). "Predictive modeling of the hospital readmission risk from patients claims data using machine learning", a case study on COPD. *Scientific reports*, 9(1), 2362.
- [6] Molly K. Bailey, Audrey J. Weiss, Marguerite L. Barrett and H. Joanna Jiang, "Characteristics of 30-Day All-Cause Hospital Readmissions 2010–2016", *Statistical Brief No. 248 hcup-us.ahrq.gov*, February 2019.
- [7] Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). "Application of machine learning in predicting hospital readmissions" .A scoping review of the literature. *BMC medical research methodology*, 21(1), 1-14.
- [8] L.-C. Hung, S.-F. Sung and Y.-H. Hu, "A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack", *Applied Sciences*, vol. 10, no. 18, pp. 6337, 2020.
- [9] A.A.Ram, Z.Ali, V.Krishna, N.Nishikaand A.Sharma, "A Guided Neural Network Approach to Predict Early Readmission of Diabetic Patients," in *IEEE Access*, vol. 11, pp. 47527-47538, 2023, doi: 10.1109/ACCESS.2023.3275086.
- [10] M.Alabdulhafithetal., "A Clinical Decision Support System for Edge/Cloud ICU Readmission Model Based on Particle Swarm Optimization, Ensemble Machine Learning, and Explainable Artificial Intelligence," in *IEEE Access*, vol. 11, pp. 100604-100621, 2023, doi:10.1109/ACCESS.2023.3312343.
- [11] K.Alghatani,N. Ammar,A.Rezgui andA.Shaban-Nejad,"PrecisionClinicalMedicineThroughMachineLearning: Using High and Low Quantile Ranges of Vital Signs for Risk Stratification of ICU Patients," in *IEEE Access*, vol. 10, pp. 52418-52430, 2022, doi: 10.1109/ACCESS.2022.3175304.
- [12] K. Alghatani, N.Ammar, A.Rezguiand A.Shaban-Nejad, "Precision Clinical Medicine Through Machine Learning: Using High and Low Quantile Ranges of Vital Signs for Risk Stratification of ICU Patients," in *IEEE Access*, vol. 10, pp. 52418-52430, 2022, doi: 10.1109/ACCESS.2022.3175304.
- [13] G.Confortola, M.Takata, N.Yokoi and M.Egi, "Enhancing Predictive Models to Lower Re hospitalization Risk: Utilizing Historical Medical Records for AI-Driven Interventions," in *IEEE Access*, vol. 12, pp. 78911-78921, 2024, doi: 10.1109/ACCESS.2024.3409152.
- [14] Yhdego HH, Nayebnazar A, Amrollahi F, Boussina A, Shashikumar S, Wardi G, Nemati S. Prediction of Unplanned Hospital Readmission using Clinical and Longitudinal Wear able Sensor Features. *medRxiv[Preprint].2023Apr11:2023.04.10.23288371*.doi:10.1101/2023.04.10.23288371.PMID:37090626; PMCID: PMC10120790
- [15] N Hassan, S Wilson, K Marley, R Slight, S Slight, Artificial intelligence informing clinical decision making on the risk of hospital readmissions in multi-morbid patients a systematic review, *International Journal of Pharmacy Practice*, Volume32, IssueSupplement_1, April2024, Pagesi40–



i41,<https://doi.org/10.1093/ijpp/riac013.05>