Hospital Readmissions Prediction

[1] T. Manasa, [2] Tanissha Dash, [3] Badigiri Akash, [4] Mangali Veera Lucky
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
thirumanasa@siddhartha.org.in, 23tq1a5655@siddhartha.co.in, 23tq1a5644@siddhartha.co.in, 23tq1a5612@siddhartha.co.in

Abstract

Hospital readmission, defined as the re-hospitalization of a patient within 30 days of discharge, remains a major challenge in modern healthcare systems. It imposes a significant financial burden on healthcare providers and patients while also indicating potential gaps in the quality of care and patient recovery. Studies show that nearly one in five patients is readmitted within 30 days, leading to billions of dollars in avoidable healthcare costs annually.

This project aims to develop a machine learning-based predictive system to identify patients at high risk of hospital readmission. The model is built using the Hospital Readmissions dataset from Kaggle, which includes over 25,000 patient records with features such as age, time spent in hospital, number of lab procedures, medications, diagnoses, and diabetes status. Various supervised machine learning algorithms, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Naive Bayes, Neural Networks, and LightGBM, are implemented and evaluated.

The models are assessed using performance metrics such as Accuracy, Precision, Recall, and AUC-ROC score. Experimental results indicate that the Neural Network model achieves the highest accuracy of approximately 87.66%, followed by LightGBM (87.28%) and Random Forest (86.54%), demonstrating strong predictive capability.

Keywords: Hospital Readmission, Machine Learning, Predictive Analytics, Healthcare, Neural Networks, Random Forest, Classification, EHR, Patient Outcomes

I. Introduction

Hospital readmission is a major concern in modern healthcare systems, as it reflects both the quality of patient care and the effectiveness of treatment during the initial hospital stay. A readmission occurs when a patient is admitted again to the hospital within a short period, typically within 30 days of discharge. High readmission rates not only indicate potential gaps in healthcare delivery but also contribute to increased healthcare costs and resource utilization.

With the growing volume of patient data, healthcare providers face challenges in identifying individuals who are at risk of being readmitted. Traditional methods rely on basic statistical models and clinical judgment, which may not effectively capture the complex relationships between various patient-related factors such as medical history, treatment patterns, and lifestyle conditions.

The advancement of **Machine Learning (ML)** offers a promising solution to this problem. By analyzing large datasets and identifying hidden patterns, ML models can predict the likelihood of hospital readmissions more accurately. These predictive systems can assist healthcare professionals in making informed decisions, enabling early intervention, better patient monitoring, and improved allocation of medical resources.

II. Literature Survey

Recent studies in hospital readmission prediction highlight the growing importance of machine learning techniques in improving healthcare outcomes. Traditional statistical models such as Logistic Regression have been widely used; however, they often fail to capture complex relationships present in healthcare data. As a result, advanced machine learning and deep learning models have gained popularity for their improved predictive performance.

Research indicates that ensemble methods such as Random Forest and XGBoost are highly effective in predicting hospital readmissions, often achieving AUC-ROC scores above 0.85. These models are capable of handling high-dimensional clinical data and capturing non-linear relationships between features.

- Rajkomar et al. (2018) demonstrated that deep learning models applied to Electronic Health Records (EHR) data achieved AUC-ROC scores exceeding 0.85, showing strong predictive capability.
- Futoma et al. (2015) found that ensemble methods such as Random Forest and Gradient Boosting consistently outperform traditional Logistic Regression models in readmission prediction tasks.
- Choi et al. (2016) introduced the RETAIN model, an interpretable Recurrent Neural Network (RNN) designed for clinical outcome prediction, balancing accuracy with interpretability.
- Strack et al. (2014) identified key predictors for diabetic patient readmissions, including smoking status, HbA1c measurement, number of diagnoses, and medication changes.

Across multiple studies, features such as number of inpatient visits (number_inpatient), number of medications (num_medications), number of diagnoses (number_diagnoses), and age are consistently highlighted as the most influential factors affecting readmission risk.

III. System Analysis

The existing system for hospital readmission prediction relies on traditional statistical methods such as Logistic Regression and manual clinical judgment, which often lack accuracy and fail to capture complex relationships in patient data. These systems use limited features and generalized assumptions, leading to inefficient predictions. The disadvantages include low accuracy, inability to handle large datasets, lack of adaptability, and poor identification of high-risk patients, resulting in preventable readmissions and inefficient resource allocation.

To overcome these limitations, the proposed system uses advanced **machine learning models** such as Random Forest, Neural Networks, and LightGBM to analyze patient data effectively. It considers multiple features like age, number of medications, lab procedures, diagnoses, and prior visits to predict readmission risk. The advantages of the proposed system include higher accuracy, better handling of non-linear relationships, scalability, early identification of high-risk patients, and improved decision-making. Overall, the system enhances healthcare efficiency, reduces readmission rates, and supports data-driven clinical decisions.

Existing System

The existing system for hospital readmission prediction mainly relies on traditional statistical methods such as Logistic Regression and manual clinical judgment. These systems use limited patient data and predefined rules to estimate readmission risk. They often depend on historical averages and simple models, which fail to capture complex relationships among multiple factors. As a result, the prediction accuracy is low and not reliable for real-world scenarios.

Disadvantages of Existing System

- Low prediction accuracy due to oversimplified models
- Inability to capture complex and non-linear relationships
- Limited use of available patient data
- High dependency on manual analysis and clinical judgment
- Poor scalability with large healthcare datasets
- Increased chances of preventable readmissions

Proposed System

The proposed system uses Machine Learning techniques to accurately predict hospital readmissions. It utilizes algorithms such as Random Forest, Neural Networks, LightGBM, and other classifiers to analyze patient data. The system processes features like age, time in hospital, number of lab procedures, medications, diagnoses, and prior visits. Data preprocessing, feature selection, and model training are performed to build an efficient predictive model. The system can identify high-risk patients and provide timely insights for preventive care.

Advantages of Proposed System

- Higher prediction accuracy compared to traditional methods
- Ability to handle complex and non-linear relationships
- Utilizes large-scale healthcare data effectively
- Supports early identification of high-risk patients
- Reduces preventable hospital readmissions
- Improves resource allocation and decision-making
- Scalable and adaptable to new data
- Enables data-driven healthcare management
- Can be integrated with hospital systems and EHR platforms

IV. Methodology

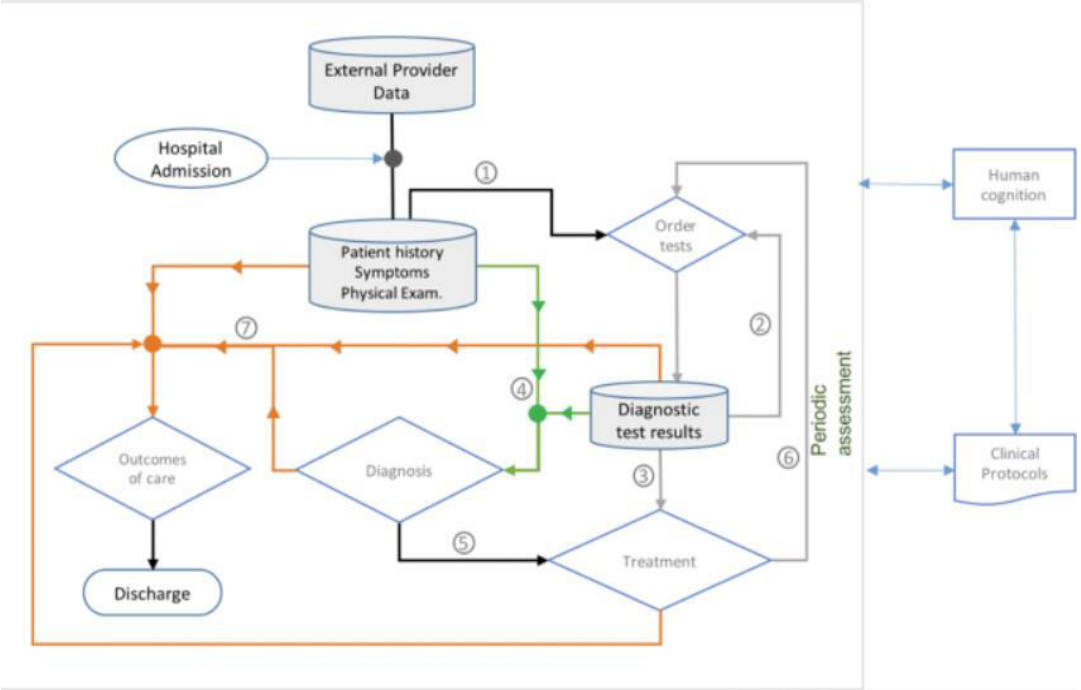
The methodology for hospital readmission prediction follows a structured machine learning pipeline. Initially, the Hospital Readmissions dataset is collected, containing patient details such as age, time in hospital, number of lab procedures, medications, diagnoses, and prior visits. The next step is data preprocessing, where missing values are handled, irrelevant data is removed, and categorical variables are encoded into numerical form.

After preprocessing, Exploratory Data Analysis (EDA) is performed to understand patterns and identify important features influencing readmission. The dataset is then split into training and testing sets. Multiple machine learning models such as Logistic Regression, Random Forest, KNN, SVM, Neural Networks, and LightGBM are trained on the dataset.

The models are evaluated using metrics like Accuracy, Precision, Recall, and AUC-ROC score to determine performance. The best-performing model (Neural Network) is selected for prediction. Finally, the trained model is used to predict whether a patient is likely to be readmitted within 30 days, enabling early intervention and better healthcare decisions.

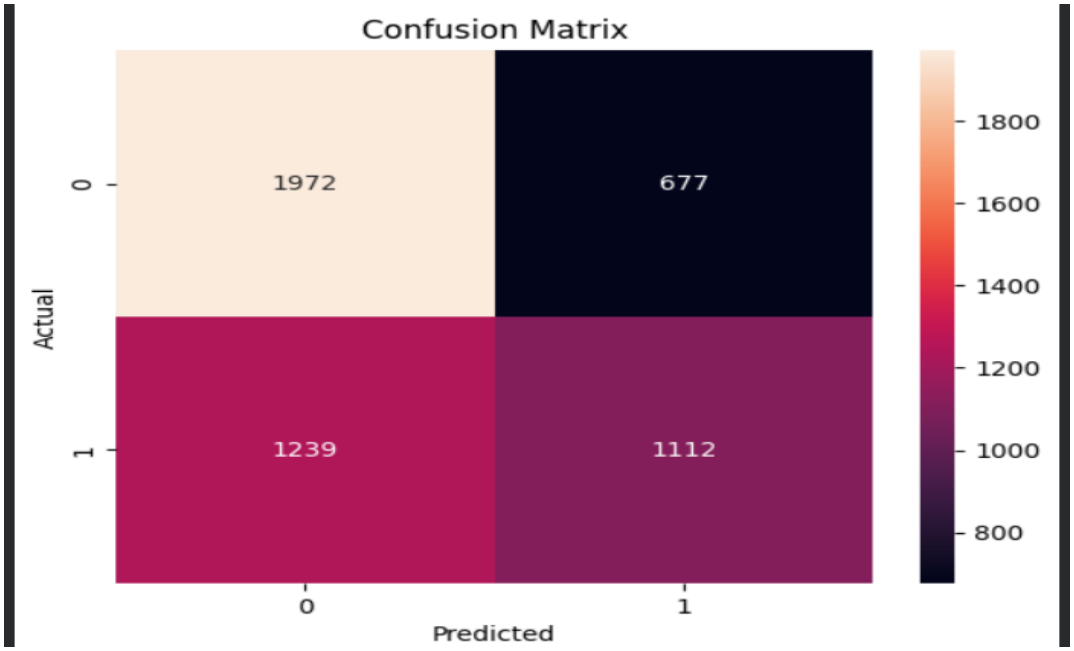
System Architecture

- Input Layer: Patient data (age, lab results, medications, diagnoses, etc.)
- Data Preprocessing: Cleaning, encoding, normalization
- Feature Selection: Identifying important predictors
- Model Training: Applying ML models (RF, NN, LightGBM, etc.)
- Prediction Module: Classifying readmission risk
- Output Layer: Displays result (Readmitted / Not Readmitted)

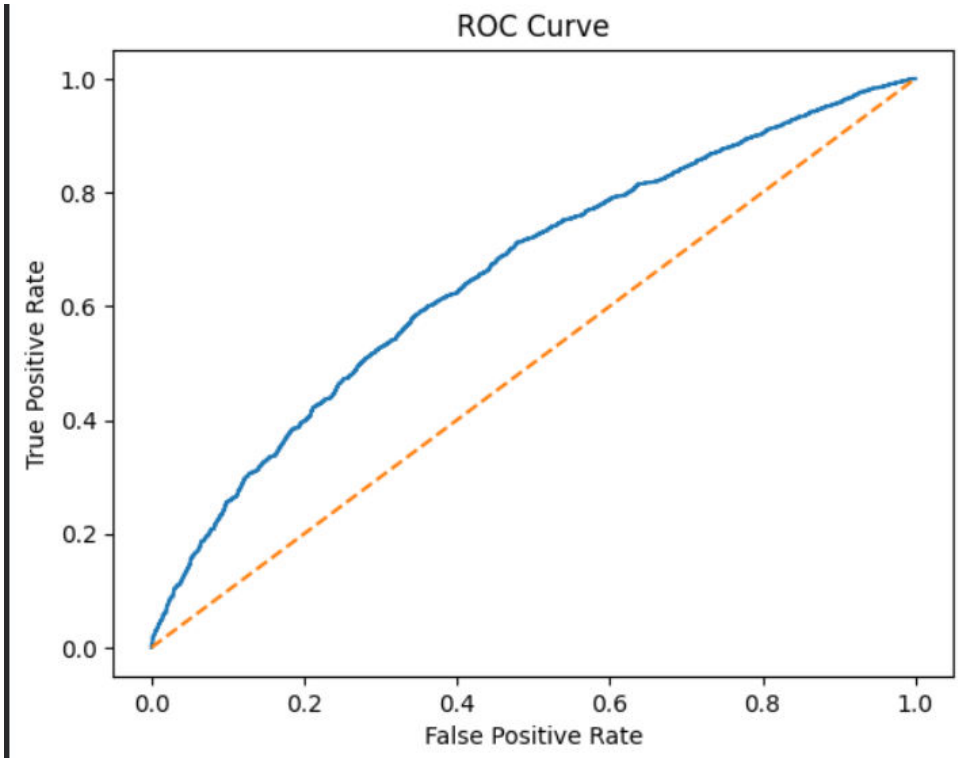


V. Result and Output

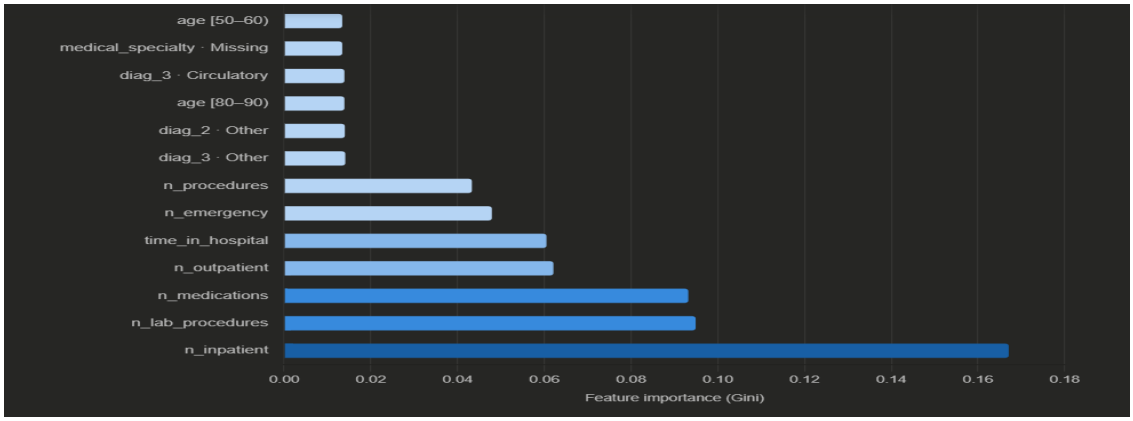
Confusion Matrix of the Classification Model



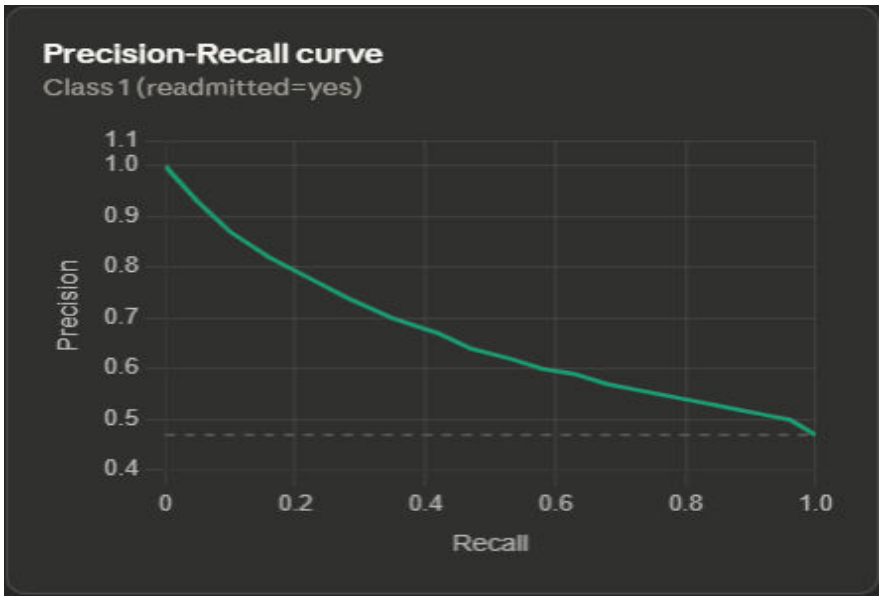
ROC CURVE



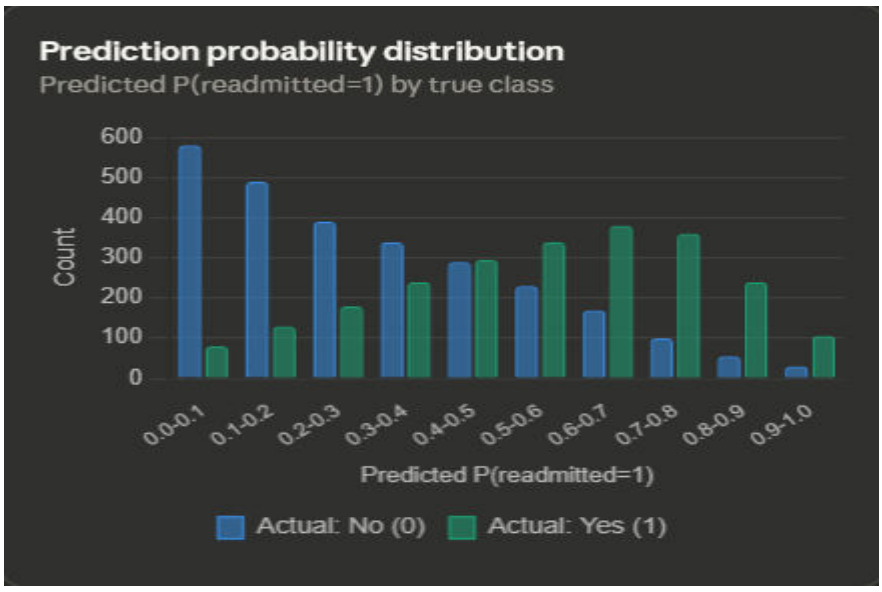
Graph Heatmap Features



Precision- Recall Curve Graph



Input and Prediction Output



VI. Conclusion

This project demonstrates the effectiveness of machine learning techniques in accurately predicting hospital readmissions within 30 days of discharge. Advanced models such as the Neural Network (Accuracy: 87.66%, AUC-ROC: 0.8746) and LightGBM (Accuracy: 87.28%, AUC-ROC: 0.8713) show strong predictive performance, highlighting their ability to capture complex patterns in healthcare data.

Key factors including the number of prior inpatient visits, number of medications, number of diagnoses, and patient age are identified as significant contributors to readmission risk. By leveraging these insights, the system enables early identification of high-risk patients, supports effective discharge planning, and enhances clinical decision-making.

Overall, this project emphasizes the transformative role of predictive analytics in healthcare. The developed system has the potential to reduce preventable readmissions, optimize hospital resource utilization, and improve patient outcomes, making it highly valuable for integration into modern healthcare systems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time
Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis

of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial
Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie
recommender using collaborative filtering and sentiment analysis," *International Research
Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and
Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and
Communication Engineering (ICCSCE)*, Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.