As per UGC guidelines an electronic bar code is provided to secure your paper

International Journal for Modern Trends in Science and Technology
Volume 12, Issue 06, pages 325-337.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue06.html
DOI: https://doi.org/10.5281/zenodo.20739565

Hospital Re-Admission Prediction System Based on HER Data

V. Sivani, Gottapu Priyanka, Kavuru Durga Prasad, Rayi Pavan Kalyan

Department of Information Technology Swarnandhra College of Engineering & Technology, Seetharamapuram, Narsapur, Andhra Pradesh, INDIA.

To Cite this Article
V. Sivani, Gottapu Priyanka, Kavuru Durga Prasad & Rayi Pavan Kalyan (2026). Hospital Re-Admission Prediction System Based on HER Data. International Journal for Modern Trends in Science and Technology, 12(06), 325-337. https://doi.org/10.5281/zenodo.20739565

Article Info
Received: 16 May 2026; Revised: 07 May 2026; Accepted: 12 June 2026.

KEYWORDS

Personalized medical recommendation system, machine learning in healthcare, disease prediction, treatment recommendation, artificial intelligence in medicine, TF-IDF vectorization, XGBoost algorithm, ensemble learning

ABSTRACT

The rapid growth of healthcare data and advancements in artificial intelligence have created new opportunities for improving medical diagnosis and treatment processes. Traditional healthcare systems often rely on generalized treatment approaches and manual diagnosis, which may lead to inaccuracies and inefficient decision-making. To address these limitations, this project presents a Personalized Medical Recommendation System that leverages machine learning techniques to predict diseases and recommend suitable treatments based on individual patient data. These factors play a crucial role in determining the health condition of an individual and are often overlooked in conventional systems. By incorporating these parameters, the system aims to provide more accurate and personalized healthcare solutions. The system utilizes Term Frequency–Inverse Document Frequency (TF-IDF) vectorization for transforming textual symptom data into numerical features, enabling effective processing by machine learning models. For disease prediction, the project employs the XGBoost algorithm, a powerful ensemble learning technique known for its high accuracy, efficiency, and ability to handle structured data. This model is trained on a dataset containing information about symptoms, diseases, and corresponding treatments. A user-friendly interface is developed using Streamlit, allowing users to easily input their symptoms and receive instant results. The system is designed to be efficient, scalable, and easy to use, making it suitable for real-world healthcare applications.

In conclusion, this project highlights the potential of machine learning in transforming healthcare systems by providing intelligent, data-driven, and personalized medical solutions. Future enhancements may include the integration of real-time patient data, expansion of the disease database, and incorporation of deep learning techniques to further improve system performance.

1. INTRODUCTION

The healthcare industry is undergoing a major transformation with the integration of advanced technologies such as machine learning, artificial intelligence, and data analytics. Traditional healthcare systems primarily rely on generalized treatment approaches, where medical decisions are based on common symptoms and predefined protocols. However, every individual has unique biological characteristics, lifestyle habits, and genetic factors, which influence their health conditions and responses to treatments.

In recent years, there has been a growing demand for personalized healthcare systems that can provide tailored medical recommendations based on individual patient data. Machine learning techniques play a crucial role in enabling such systems by analyzing large volumes of healthcare data and identifying meaningful patterns.

This project focuses on developing a Personalized Medical Recommendation System that predicts diseases and suggests appropriate treatments using machine learning algorithms. The system aims to improve diagnostic accuracy, reduce human errors, and provide efficient healthcare solutions.

1.1 OVERVIEW OF THE PERSONALIZED HEALTHCARE SYSTEMS

Personalized healthcare represents a paradigm shift from traditional treatment approaches to a more patient-centric model, where medical decisions are tailored to the specific needs and characteristics of individuals. This approach recognizes that each patient is unique and that effective treatment requires a comprehensive understanding of various factors, including genetic makeup, lifestyle habits, environmental influences, and medical history. Unlike conventional systems that apply the same treatment protocols to all patients with similar symptoms, personalized healthcare focuses on delivering targeted therapies that are more likely to produce positive outcomes.

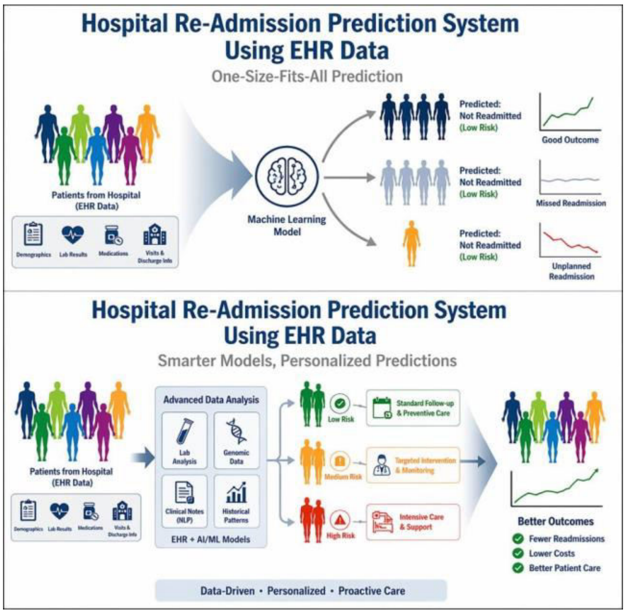


Fig 1.1: The Evolution of Personalized Medicine

The concept of personalized healthcare has gained significant importance with advancements in medical research and technology. The availability of large-scale healthcare data has enabled researchers and practitioners to analyze patterns and correlations that were previously difficult to identify. This has led to the development of systems that can predict disease risks, recommend preventive measures, and provide customized treatment plans. Personalized healthcare systems are particularly beneficial in managing chronic diseases, where long-term treatment strategies need to be tailored to individual patients.

In the context of this project, the personalized healthcare system is designed to analyze inputs such as symptoms, smoking habits, alcohol consumption, and genetic predisposition. These factors play a crucial role in determining the health status of an individual and can significantly influence the effectiveness of treatments. By incorporating these parameters, the system provides a more comprehensive analysis and generates recommendations that are specific to each patient.One of the key advantages of personalized healthcare systems is their ability to improve diagnostic accuracy.

systems can identify subtle patterns that may not be apparent through manual analysis. This leads to early detection of diseases and timely intervention, which can significantly improve patient outcomes. Additionally, personalized systems help in reducing the risk of adverse drug reactions by recommending medications that are suitable for the individual's condition.

However, the implementation of personalized healthcare systems also presents certain challenges. The need for large and high-quality datasets, concerns related to data privacy and security, and the complexity of integrating various data sources are some of the major issues that need to be addressed. Despite these challenges, personalized healthcare is considered the future of medicine, and ongoing advancements in technology are expected to overcome these limitations.

1.2 MACHINE LEARNING IN HEALTH CARE

Machine learning has become an essential component of modern healthcare systems, providing powerful tools for analyzing complex data and supporting decision-making processes. It involves the use of algorithms that can learn from data and improve their performance over time without being explicitly programmed. In healthcare, machine learning is used to analyze large datasets, identify patterns, and make predictions that can assist in diagnosis, treatment, and patient care.

Healthcare data is highly complex and diverse, consisting of structured data such as patient records and laboratory results, as well as unstructured data such as clinical notes, medical images, and sensor data. Traditional data analysis methods are often insufficient for handling such complexity. Machine learning algorithms, on the other hand, are capable of processing large volumes of data and extracting meaningful insights that can be used to improve healthcare outcomes.

One of the key applications of machine learning in healthcare is disease prediction. By analyzing historical patient data, machine learning models can identify patterns and correlations that indicate the presence of certain diseases. This enables early detection and timely intervention, which can significantly improve patient outcomes. In addition to disease prediction, machine learning is also used for treatment recommendation, medical imaging analysis, drug discovery, and patient risk assessment.

In this project, machine learning is used to develop a system that predicts diseases based on patient inputs. The system uses an advanced algorithm known as XGBoost, which is an ensemble learning method that combines multiple decision trees to produce a highly accurate predictive model. XGBoost is widely used in machine learning applications due to its efficiency, scalability, and ability to handle large datasets. It is particularly effective in classification tasks, making it suitable for disease prediction.

Another important aspect of the system is the use of natural language processing techniques to handle textual data. Symptoms provided by users are often in text format, which needs to be converted into numerical form for processing by machine learning models. Techniques such as TF-IDF are used to transform text into feature vectors, enabling effective analysis. This allows the system to interpret user inputs accurately and generate relevant predictions.

The system also uses cosine similarity to compare user inputs with historical data and identify similar cases. This helps in providing personalized recommendations by suggesting treatments that have been effective for similar patients. The combination of machine learning and natural language processing techniques enhances the overall performance of the system and ensures accurate and relevant outputs.

Despite its numerous advantages, the application of machine learning in healthcare also faces several challenges. One of the major challenges is the quality and availability of data. Inaccurate or incomplete data can lead to incorrect predictions and reduce the reliability of the system. Another challenge is the lack of transparency in machine learning models, which can make it difficult to understand how decisions are made. Ensuring the privacy and security of patient data is also a critical concern.

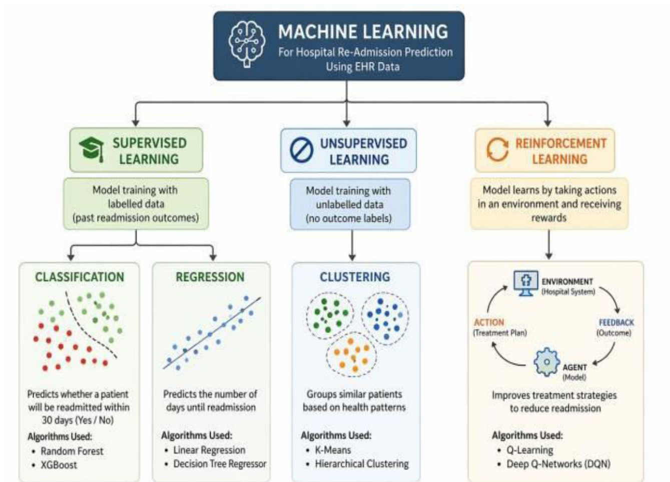


Fig 1.2: Machine Learning in Hospital Re-Admission

1.3 PROBLEM STATEMENT

The healthcare sector is currently facing several challenges that limit the efficiency and effectiveness of medical services. One of the primary issues is the reliance on traditional methods of diagnosis and treatment, which are often based on generalized assumptions rather than individual patient characteristics. This approach does not take into account the variations in patient conditions, leading to inconsistent treatment outcomes and, in some cases, adverse effects.. Another significant challenge is the rapid growth of healthcare data. With the increasing use of digital technologies, a vast amount of patient data is being generated, including electronic health records, diagnostic reports, and data from wearable devices. While this data has the potential to provide valuable insights into patient health, it is often underutilized due to the lack of efficient analysis tools. Manual analysis of such large datasets is not only time-consuming but also prone to errors.

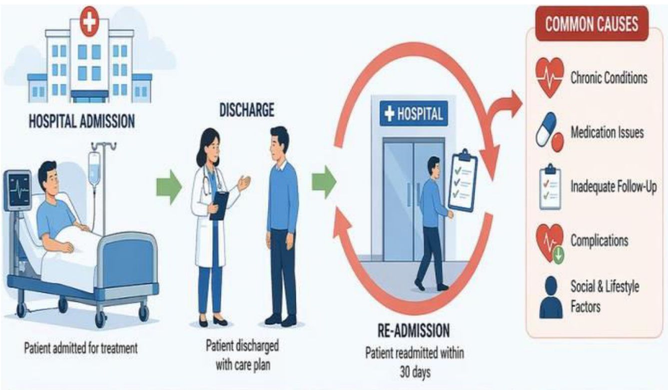


Fig 1.3: Role of Ensemble Learning in Hospital Re-admission

Human errors in diagnosis and treatment further complicate the situation. Factors such as fatigue, lack of experience, and misinterpretation of symptoms can lead to incorrect decisions. These errors can have serious consequences, including delayed treatment, incorrect medication, and deterioration of patient health. The absence of intelligent systems to assist healthcare professionals in decision-making increases the likelihood of such errors.

Existing healthcare systems also lack the ability to provide personalized recommendations. Most systems are designed to handle general cases and do not consider individual factors such as lifestyle habits, genetic predisposition, and environmental influences. This limits their effectiveness in providing accurate and relevant treatment recommendations.The problem addressed in this project is the development of a system that can overcome these limitations by leveraging machine learning techniques.

The system aims to analyze patient data, predict diseases accurately, and provide personalized treatment recommendations. By incorporating advanced algorithms and data analysis techniques, the system seeks to improve the accuracy and efficiency of healthcare service.

1.4 OBJECTIVES OF THE PROJECT

The primary objective of this project is to develop an intelligent medical recommendation system that utilizes machine learning techniques to enhance the quality and efficiency of healthcare services. The system is designed to analyze patient data and provide accurate disease predictions and personalized treatment recommendations One of the main objectives is to develop a predictive model that can accurately identify diseases based on patient inputs. This involves training the model on a comprehensive dataset and optimizing its performance to achieve high accuracy. The use of advanced algorithms ensures that the system can handle complex data and produce reliable predictions.

Another important objective is to provide personalized treatment recommendations that are tailored to individual patients. The system takes into account various factors such as symptoms, lifestyle habits, and genetic predisposition to generate recommendations that are specific to each patient. This approach improves the effectiveness of treatments and reduces the risk of adverse outcomes. The project also aims to integrate natural language processing techniques to handle textual data efficiently. By converting text inputs into numerical features, the system can analyze user inputs more effectively and provide accurate results. The use of cosine similarity further enhances the system by identifying similar cases and recommending appropriate treatments. In addition to improving accuracy and personalization, the project aims to develop a user-friendly interface that allows users to easily interact with the system. The interface is designed to be simple and intuitive, enabling users to input data and receive results without any technical complexity. This ensures that the system can be used by a wide range of users, including patients and healthcare professionals.

The project also focuses on improving the overall efficiency of healthcare services by reducing dependency on manual processes. Furthermore, the system is designed to be scalable, allowing it to handle large datasets and accommodate future enhancements.

2. LITERATURE SURVEY

The prediction of hospital readmissions has become a critical area of research in healthcare analytics due to its significant impact on patient outcomes and healthcare costs. With the rapid growth of digital healthcare systems, large volumes of patient data are now available in the form of Electronic Health Records (EHR). These records contain both structured data such as laboratory values and unstructured data such as clinical notes.

Over the years, various techniques have been proposed to analyze EHR data for predicting hospital readmissions. These approaches range from traditional statistical models to advanced machine learning and deep learning techniques. However, each method has its own limitations in terms of accuracy, scalability, and interpretability.

2.1 TRADITIONAL MACHINE LEARNING APPROACHES

Used for prediction tasks. These methods primarily rely on structured numerical data extracted from EHR.

Logistic Regression

Logistic Regression is one of the most commonly used statistical models for binary classification problems. In hospital readmission prediction, it is used to estimate the probability of a patient being readmitted based on input features such as age, medical history, and lab results.

Advantages:

- Simple and easy to interpret
- Computationally efficient
- Work Well For Small Dataset

Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy.

Advantages:

- Reduces overfitting
- Provides better accuracy than single decision trees
- Handles large datasets effectively

2.2 DEEP LEARNING APPROCHES

Deep learning has significantly improved the performance of prediction systems by enabling automatic feature extraction and modeling complex patterns. Artificial Neural Networks (ANN)

Artificial Neural Networks consist of interconnected layers of neurons that process input data and generate predictions.

Advantages:

- Capable of modeling complex relationships
- Handles large datasets
- Automatic feature learning Limitations:
- Requires large training data
- Computationally expensive
- Difficult to interpret

Recurrent Neural Networks (RNN)

RNNs are designed to handle sequential data by maintaining memory of previous inputs. This makes them suitable for time-series healthcare data.

Advantages:

- Captures temporal dependencies
- Suitable for sequential data Limitations:
- Vanishing gradient problem
- Difficulty in learning long-term dependencies

2.3 NATURAL LANGUAGE PROCESSING IN HEALTHCARE

A significant portion of healthcare data exists in the form of unstructured text, such as clinical notes and discharge summaries. Natural Language Processing (NLP) techniques are used to extract meaningful information from this text. Importance of Clinical Text

Clinical text contains:

- Patient symptoms
- Doctor observations
- Treatment details
- Medical history

Ignoring this data leads to incomplete analysis.

Traditional NLP Techniques

Earlier NLP methods include:

- Bag of Words (BoW)
- TF-IDF Limitations:
- Cannot capture context

- Limited semantic understanding

Topic Modeling

Topic modeling is used to identify hidden themes in text data. One advanced method is BERTopic, which uses transformer-based embeddings.

Advantages:

- Captures semantic meaning
- Provides interpretable topics
- Enhances feature representation

NLP Challenges in Healthcare

- Complex medical terminology
- Abbreviations and jargon

2.4 LIMITATIONS OF EXISTING APPROACHES

Recent research has focused on combining structured and unstructured data to improve prediction accuracy.

Advantages of Hybrid Models

- Better feature representation
- Improved accuracy
- Enhanced interpretability

Challenges

- Data integration complexity
- High computational cost
- Feature selection issues

CLASS IMBALANCE PROBLEM

Healthcare datasets are often imbalanced, with fewer readmission cases compared to nonreadmission cases.

Impact of Imbalance

- Biased predictions
- Poor model performance

Solution: SMOTE

SMOTE is used to generate synthetic samples for minority classes.

Advantages:

- Balances dataset
- Improves model performance

3. PROPOSED METHODOLOGY

The proposed system is designed to overcome the inefficiencies and limitations of traditional healthcare approaches, which often depend heavily on manual assessment and generalized treatment protocols. Conventional methods may fail to capture complex relationships within patient data, leading to delayed or inaccurate risk evaluations. By integrating machine learning techniques, the system introduces an intelligent framework that enhances the accuracy and reliability of healthcare decision-making.

A key foundation of the system is the use of Electronic Health Records (EHR), which provide structured and comprehensive patient information. These records include demographic details, medical history, clinical observations, and treatment data. By leveraging this rich dataset, the system moves beyond surface-level analysis and enables deeper insight into patient conditions. This data-driven approach supports more informed and personalized clinical decisions.

One of the major strengths of the system lies in its high predictive accuracy. By employing the XGBoost algorithm, the model can detect complex, nonlinear patterns within patient data that may not be easily identified through manual review. XGBoost's ensemble learning capability improves model performance by combining multiple weak learners, resulting in robust and precise readmission risk predictions.

Another important aspect of the system is its ability to provide personalized insights. Instead of applying generalized healthcare guidelines, the model evaluates individual patient characteristics such as age, length of hospital stay, prior diagnoses, and clinical indicators. This personalized assessment ensures that risk predictions are tailored to each patient, leading to more targeted and effective interventions.

The system also enhances operational efficiency within healthcare settings. Automated data processing significantly reduces the workload of medical professionals by minimizing the need for manual analysis. Additionally, it decreases the possibility of human error that may arise from fatigue, oversight, or subjective judgment, thereby improving overall reliability in patient evaluation.

Finally, the system offers real-time clinical decision support. It generates instant risk scores along with actionable insights, allowing healthcare providers to quickly identify high-risk patients. Early identification enables timely interventions, which can reduce hospital readmission rates and improve patient outcomes. Overall, the system represents a significant advancement toward intelligent, proactive, and patient-centered healthcare management.

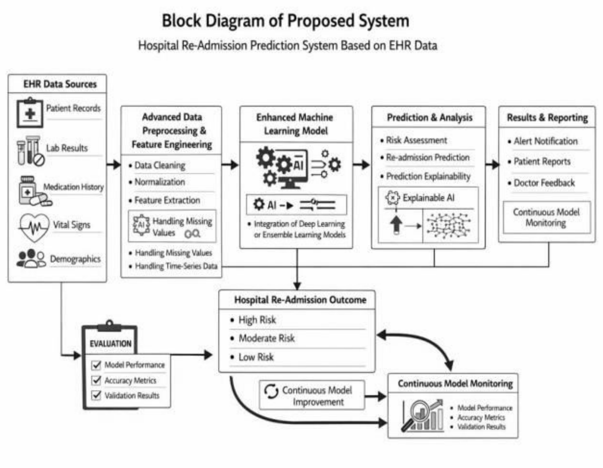


Fig 3. Block Diagram of Purposed System

The system also incorporates natural language processing techniques to handle textual data. Symptoms provided by users are often in text format, which needs to be converted into numerical form for analysis. Techniques such as TF-IDF are used to transform text into feature vectors, enabling effective processing by machine learning models Another important component of the proposed system is the use of similarity-based techniques to identify similar cases from historical data. By comparing user inputs with past records, the system can recommend treatments that have been effective for similar patients. This enhances the reliability of the recommendations and improves the overall performance of the system. The proposed system also provides real-time recommendations, allowing users to receive instant results based on their inputs. This reduces the dependency on manual consultations and improves the efficiency of healthcare services.

Overall, the proposed system leverages advanced technologies to provide a more efficient, accurate, and personalized approach to healthcare. It addresses the limitations of the existing system and offers a comprehensive solution for disease prediction and treatment recommendation.

3.1 SYSTEM ARCHITECTURE

The proposed system architecture adopts a layered design paradigm to ensure modularity, scalability, and efficient data flow across components. This structured organization enables independent development and maintenance of each layer while maintaining seamless integration throughout the predictive pipeline. The layered approach also enhances system flexibility, allowing future upgrades or model enhancements without disrupting the overall framework.

The first layer is the User Interface Layer, implemented using Streamlit. This interactive frontend enables healthcare professionals to input patient-specific information, including demographic details, laboratory results, and discharge summaries. The interface is designed to be intuitive and user-friendly, ensuring accurate data entry while providing real-time responsiveness. By facilitating structured input collection, this layer serves as the primary communication bridge between users and the predictive system.

The second layer is the Data Processing Layer, which prepares raw Electronic Health Record (EHR) data for analysis. This layer performs essential preprocessing operations such as handling missing values, removing inconsistencies, encoding categorical attributes, and normalizing numerical variables. These steps are critical for ensuring data quality and consistency, thereby improving the reliability and performance of subsequent machine learning models.

The third layer constitutes the Machine Learning and Prediction Layer, which functions as the core analytical engine of the system. In this stage, preprocessed data undergoes feature engineering to extract meaningful predictive attributes. The transformed data is then fed into trained ensemble learning models, such as XGBoost or Random Forest, which generate probability-based predictions of patient readmission risk. This layer is responsible for identifying complex patterns and relationships within structured healthcare data.

Following prediction, the Recommendation Engine provides context-aware insights by comparing the current patient profile with historical medical records. Using similarity measures and historical outcome analysis, the system generates personalized treatment suggestions and preventive strategies. This layer enhances the system's clinical utility by moving beyond risk prediction to actionable recommendations.

The final layer is the Output and Clinical Support Layer, which presents prediction results through a comprehensive dashboard. The output includes risk classification (e.g., Yes/No), probability scores, and key contributing factors influencing the model's decision. By delivering clear and interpretable insights, this layer supports healthcare professionals in making informed, data-driven clinical decisions, thereby improving patient management and reducing readmission rates.

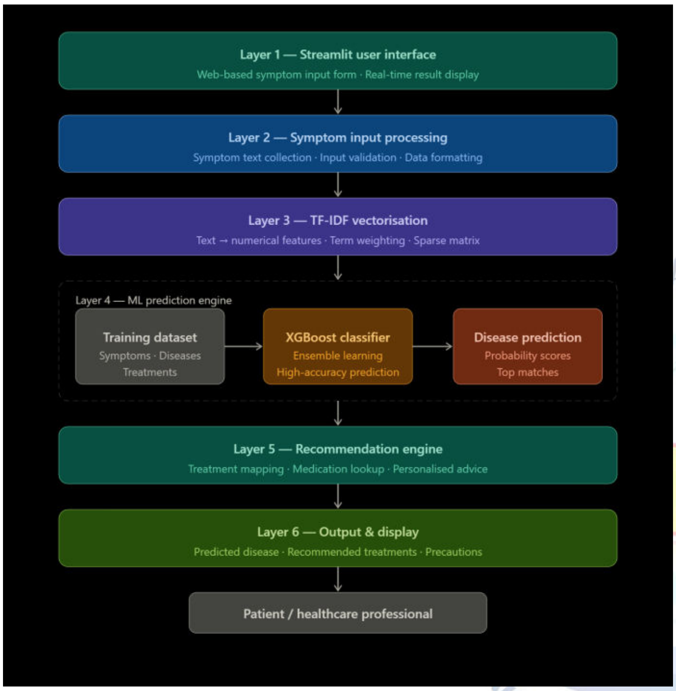


Fig 3.1 Proposed System Architecture

The proposed Personalized Medical Recommendation System follows a multi-layered architecture designed to ensure modularity, scalability, and efficient data processing. The architecture is structured into six logical layers, each responsible for a specific stage in the predictive and recommendation pipeline.

The first layer consists of the Streamlit-based User Interface, which serves as the interaction point between the user and the system. It provides a web-based symptom input form and enables real-time display of prediction results. This layer ensures accessibility and ease of use for both patients and healthcare professionals.

The second layer performs Symptom Input Processing, where textual symptom descriptions are collected, validated, and formatted. This layer ensures that user-provided inputs are clean, structured, and suitable for further computational processing.

The third layer implements TF-IDF vectorisation. In this stage, textual symptom data is transformed into numerical feature representations using term frequency–inverse document frequency weighting. This conversion enables machine learning algorithms to process unstructured text data effectively by representing it in a sparse matrix format.

The fourth layer represents the Machine Learning Prediction Engine. It utilizes a pre-trained XGBoost classifier, trained on a dataset containing symptoms, diseases, and treatments. The classifier generates probability scores and identifies top disease matches based on learned patterns from historical data.

The fifth layer is the Recommendation Engine, which maps predicted diseases to corresponding treatments, medications, and precautionary measures. The final layer handles Output and Display, presenting predicted diseases, recommended treatments, and preventive advice to the end user. Together, these layers create a seamless predictive pipeline that supports clinical decision-making.

In the layered organization, the architecture is designed with clear separation of responsibilities between components, enabling independent development, testing, and optimization of each layer. This modular design ensures that improvements in one module—such as upgrading the classification algorithm or enhancing preprocessing techniques—can be implemented without affecting other system components. Such flexibility is essential for healthcare applications, where models and datasets may evolve over time.

The architecture also emphasizes scalability and adaptability. As the volume of medical data increases or new diseases and treatment protocols are introduced, the system can be extended by updating the training dataset or integrating additional machine learning models. The TF-IDF and classification components can be retrained with expanded corpora, allowing the system to adapt to emerging medical knowledge while maintaining consistent performance and reliability.

Furthermore, the end-to-end pipeline supports real-time processing and decision support. From symptom entry to final recommendation output, each layer contributes to minimizing latency while preserving predictive accuracy. By combining efficient text vectorization, a high-performance ensemble classifier, and a structured recommendation mechanism, the system delivers timely, data-driven insights that assist healthcare professionals and patients in making informed medical decisions.

3.2 USE CASE DIAGRAM

The Use Case Diagram represents the functional behavior of the Heart Disease Prediction System and illustrates how external users interact with the system. It provides a high-level view of the system's capabilities by identifying the actors involved and the services or functions the system offers. The primary purpose of the use case diagram is to define the system's functional requirements in a clear and structured manner.

In this system, the main actor is the User, who may be a doctor, healthcare professional, or patient. The user interacts with the system to assess the risk of heart disease based on medical parameters. The diagram focuses only on external interactions and does not describe internal processing or technical implementation details.

One of the primary use cases is entering medical details. In this functionality, the user provides required health-related information such as age, blood pressure, cholesterol level, chest pain type, and other clinical attributes. This use case ensures that the system receives necessary input data for prediction.

Another important use case is validating and preprocessing input data. The system verifies whether the entered information is complete and within acceptable ranges. Although preprocessing is handled internally, from a functional perspective it ensures that the data is properly prepared before prediction is performed.

The core use case of the system is predicting heart disease. In this functionality, the system analyzes the processed input data using the stacked ensemble learning model and determines whether the individual is at high or low risk of heart disease. This represents the central objective of the entire system.

The final use case involves displaying the prediction result and providing recommendations. The system presents the outcome in a clear and understandable format and may offer precautionary suggestions if a high risk is detected. These use cases collectively define how the user interacts with the system and outline the functional scope of the application

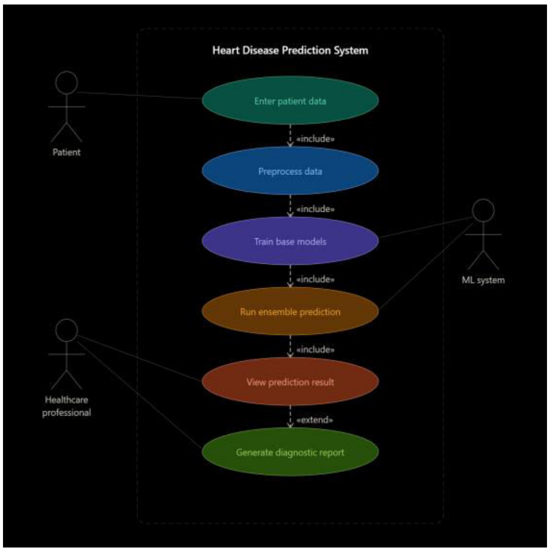


Fig 3.2 Use Case Diagram

The Use Case Diagram provides a functional overview of the Personalized Medical Recommendation System by illustrating the interactions between external actors and the system. It defines the system boundary and identifies the services that the system offers to its users. This diagram is essential for understanding the functional requirements and overall operational scope of the application.

The primary actors identified in the system are the Patient and the Doctor. These actors represent the end users who directly interact with the application. The Machine Learning System operates as an internal computational component that supports prediction and recommendation processes but does not act independently outside the system boundary.

One of the fundamental use cases is Enter Symptoms, where patients or doctors provide textual descriptions of symptoms through the user interface. This use case initiates the system workflow and serves as the primary input mechanism. The accuracy and completeness of symptom input directly influence the quality of prediction results.

The Enter Symptoms use case includes internal operations such as preprocessing and TF-IDF vectorization. Although these processes are not directly visible to users, they are essential system functions that convert textual data into structured numerical features suitable for machine learning analysis. These included use cases ensure proper preparation of data before prediction.

Another central use case is Predict Disease, which activates the trained XGBoost classification model. In this process, the system analyzes the processed feature vectors and generates probability-based predictions. The outcome includes the most likely disease along with associated confidence scores.

Following disease prediction, the View Prediction Results use case allows users to access the generated output. The system presents predicted diseases, probability values, and recommended treatments in a structured format. For doctors, this functionality supports informed clinical assessment and decision-making.

An extended use case, View Treatment History, enables doctors to review previously generated medical records and recommendations. Additionally, the Manage Disease Database use case allows administrative control over disease-related data, including updating treatment mappings and precautionary guidelines. Together, these use cases define the complete functional interaction model of the system and ensure comprehensive healthcare support.

3.3 CLASS DIAGRAM

The static structural design of the system is represented through a set of well-defined classes, each responsible for a specific functional component within the predictive framework. This object-oriented design ensures modularity, maintainability, and clear separation of concerns. Each class encapsulates its data attributes and operational methods, enabling structured communication between different components of the system.

The StreamlitApp class functions as the primary interface layer of the application. It manages user interactions, renders dashboards, and coordinates communication between frontend components and backend services. This class is responsible for capturing user inputs, displaying prediction results, and presenting evaluation metrics in an intuitive format suitable for healthcare professionals.

The UserInputHandler class manages the collection and validation of user-provided data. It ensures that demographic details, laboratory results, and discharge summaries are complete and consistent before further processing. By implementing validation checks and error-handling mechanisms, this class maintains data integrity and prevents invalid inputs from affecting model predictions.

The DataProcessor class handles data preprocessing tasks required before model execution. It includes methods for cleaning raw Electronic Health Record (EHR) data, handling missing values, encoding categorical variables, and normalizing numerical attributes. This class ensures that the dataset conforms to the structured format expected by machine learning algorithms.

The FeatureEngineer class is responsible for transforming processed data into meaningful predictive attributes. It derives additional features such as readmission frequency, aggregated clinical indicators, and computed risk-related metrics. By generating informative features, this class enhances the predictive capability of the machine learning models.

The PredictionEngine class encapsulates the core machine learning models, including XGBoost and Random Forest. It contains methods for loading trained models, performing predictions, and generating probability scores. This class acts as the analytical core of the system, translating processed patient data into readmission risk predictions.

Finally, the ModelEvaluator class monitors and assesses model performance using evaluation metrics such as accuracy, precision, recall, and F1-score. It provides mechanisms for validating model reliability and ensuring consistent predictive quality. Together, these classes form a cohesive and extensible structural design that supports accurate, efficient, and scalable healthcare risk prediction.

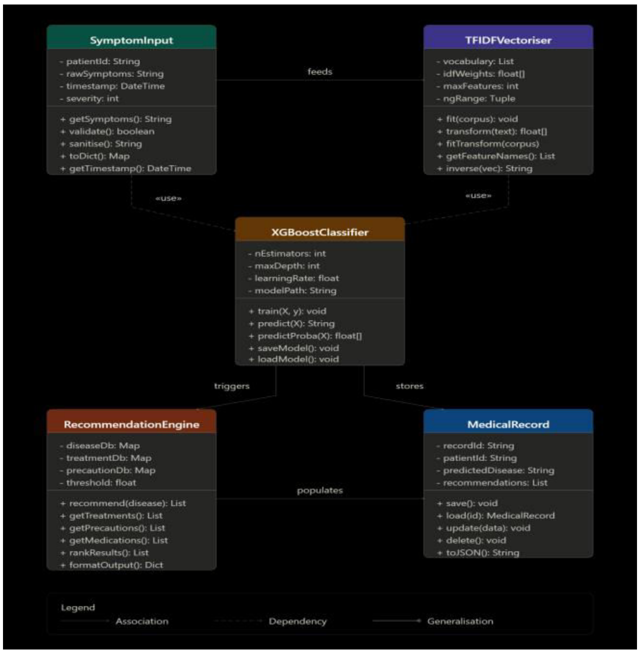


Fig 3.3 Class Diagram

The Class Diagram illustrates the static structural framework of the Personalized Medical Recommendation System by defining the major classes, their attributes, methods, and the relationships among them. It provides a blueprint of how system components are organized and interact at a structural level. By representing the system using object-oriented principles, the diagram enhances clarity, maintainability, and scalability of the software architecture.

The SymptomInput class represents the initial data entity within the system. It encapsulates patient-provided information, including attributes such as patientId, rawSymptoms, timestamp, and severity level. The methods within this class are responsible for validating user input, sanitizing textual entries, and converting unstructured symptom descriptions into a structured format. This ensures that all incoming data adheres to predefined standards before further processing.

The TFIDFVectoriser class functions as a feature transformation component. It converts textual symptom data into numerical feature vectors using term frequency–inverse document frequency weighting. This class maintains essential attributes such as vocabulary terms, inverse document frequency values, and configuration parameters. Its core methods—fit(), transform(), and fitTransform()—enable training on textual data and transformation into sparse numerical matrices suitable for machine learning algorithms.

The XGBoostClassifier class represents the primary predictive engine of the system. It encapsulates model configuration parameters such as the number of estimators, maximum tree depth, and learning rate. Key methods include train() for model training, predict() for disease classification, predictProba() for probability estimation, and saveModel() and loadModel() for model persistence. This class depends on outputs generated by the TFIDFVectoriser and SymptomInput classes, demonstrating dependency relationships within the architecture.

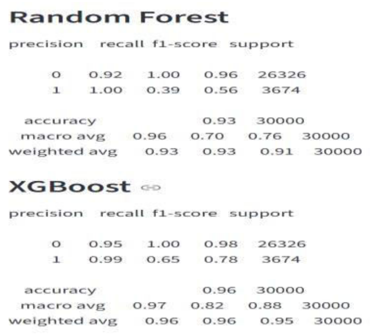
The RecommendationEngine class operates after disease prediction is completed. It maintains mappings between predicted diseases and associated treatments, medications, and precautionary measures. Based on prediction probabilities and defined threshold values, it generates ranked recommendations tailored to the patient's condition. This class enhances the system's functionality by translating predictive results into actionable medical guidance.

The MedicalRecord class is responsible for storing and managing patient prediction histories. It includes attributes related to patient identification, predicted diseases, probability scores, recommended treatments, and timestamps. Its methods support operations such as saveRecord(), updateRecord(), loadRecord(), and deleteRecord(), ensuring systematic storage and retrieval of medical information.

The relationships among these classes define the system's internal data flow. The SymptomInput class provides validated input to the TFIDFVectoriser, which generates feature vectors consumed by the XGBoostClassifier. The classifier produces predictions that trigger the RecommendationEngine, and the resulting outcomes are stored within the MedicalRecord class. These associations represent structured communication pathways between components.

Overall, the class diagram reflects a modular and object-oriented design approach. Each class has clearly defined responsibilities, reducing coupling and improving cohesion within the system. This design not only facilitates easier maintenance and debugging but also allows seamless extension, such as integrating additional classifiers or expanding the recommendation database, thereby ensuring long-term adaptability of the healthcare recommendation framework.

RESULTS


Random Forest

	precision	recall	f1-score	support
0	0.92	1.00	0.96	26326
1	1.00	0.39	0.56	3674
accuracy			0.93	30000
macro avg	0.96	0.70	0.76	30000
weighted avg	0.93	0.93	0.91	30000

XGBoost

	precision	recall	f1-score	support
0	0.95	1.00	0.98	26326
1	0.99	0.65	0.78	3674
accuracy			0.96	30000
macro avg	0.97	0.82	0.88	30000
weighted avg	0.96	0.96	0.95	30000

Fig 4.1: Metrics Performance

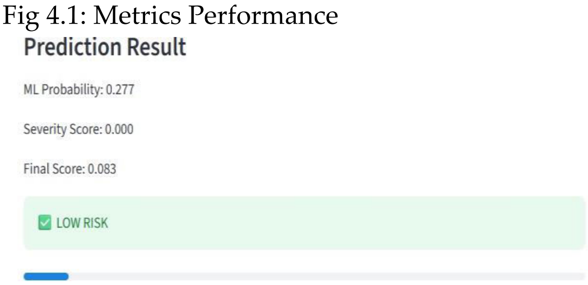


Fig 4.2: High Risk

5. CONCLUSION

The Hospital Readmission Prediction System Based on Electronic Health Record (EHR) Data was developed to address one of the major challenges faced by modern healthcare institutions, which is the increasing rate of unplanned patient readmissions. Hospital readmissions not only increase the financial burden on healthcare organizations but also affect patient satisfaction, hospital resource utilization, and the overall quality of care. Predicting readmission risk before patient discharge can help healthcare providers take preventive actions and improve clinical outcomes. This project successfully demonstrates how machine learning techniques can be applied in the healthcare domain to solve such a critical issue.

The main objective of this project was to design a system capable of analyzing patient medical data and predicting the likelihood of hospital readmission. The system uses structured patient information collected from Electronic Health Records, including demographic details, diagnosis history, medication records, laboratory results, previous admissions, and treatment information. By processing these datasets, the model identifies hidden patterns that may not be easily recognized through traditional clinical observation. These patterns are then used to classify patients into different risk categories such as low risk, medium risk, and high risk.

The implementation of the system involved several important stages. Initially, the dataset was collected and carefully preprocessed to remove missing values, duplicate records, and inconsistent data entries. Since healthcare data often contains noisy and incomplete information, preprocessing played a crucial role in improving the quality of the dataset. Feature engineering was then performed to identify the most significant factors influencing readmission. This step improved the model's ability to make meaningful predictions by selecting relevant clinical variables from the dataset.

FUTURE ENHANCEMENT

Although the current system provides effective predictions, there are several areas where enhancements can be made to further improve its performance and usability.In the future, the system can be extended by incorporating real-time patient monitoring data such as IoT-based health sensors and wearable devices. This would allow continuous tracking of patient health and improve prediction accuracy. Additionally, advanced machine learning models such as deep learning and ensemble techniques can be explored to achieve better performance.

Another potential improvement is the integration of natural language processing (NLP) techniques to analyze unstructured data such as doctors' notes, discharge summaries, and clinical reports. This will provide deeper insights into patient conditions and enhance prediction capabilities. The system can also be deployed as a cloud-based application, enabling easy access for multiple hospitals and healthcare providers. Mobile application support can further improve usability by allowing doctors to access predictions on the go.

Moreover, incorporating explainable AI (XAI) techniques will help in understanding the reasoning behind predictions, thereby increasing trust among healthcare professionals.

Security and privacy measures can also be strengthened to ensure compliance with healthcare regulations. In conclusion, with continuous improvements and integration.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[2] Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems, 29, 3504–3512. https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79Paper.pdf

[3] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[4] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from electronic health records. Scientific Reports, 6(1), 26094. https://doi.org/10.1038/srep26094

[5] Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics, 56, 229–238. https://doi.org/10.1016/j.jbi.2015.05.016

[6] Krumholz, H. M. (2013). Post-hospital syndrome — an acquired, transient condition of generalized risk. New England Journal of Medicine, 368(2), 100–102. https://doi.org/10.1056/NEJMp1212324

[7] Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. JAMA, 306(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515

[8] Zhou, H., Della, P. R., Roberts, P., Goh, L., & Dhaliwal, S. S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: An updated systematic review. BMJ Open, 6(6), e011060. https://doi.org/10.1136/bmjopen-2016-011060

[9] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

[10] Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208. https://doi.org/10.1093/jamia/ocw042

[11] Amarasingham, R., Velasco, F., Xie, B., & Clark, C. A. (2010). Electronic medical record-based multicondition predictive model for inpatient readmission risk. Journal of the American Medical Informatics Association, 17(5), 504–507. https://doi.org/10.1136/jamia.2010.003855

[12] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101