THE ASIAN BULLETIN OF BIG DATA MANAGEMENT
Vol.4. Issue4 (2024)
https://doi.org/ 10.62019/abbdm.v4i4.280

Enhancing Patient Survival Prediction via Cutting-Edge Data Mining Techniques and Future Prospects

Asad Iqbal, Khawaja Qasim Maqbool, Muhammad Zulkifl Hasan*, Nadeem Sarwar

Chronicle

Article history

Received: Oct 12, 2024
Received in the **revised format:** Oct 29, 2024
Accepted: Nov 11, 2024
Available online: December 20, 2024

Asad Iqbal, is currently affiliated with National College of Business Administration and Economics - NCBA&E, Lahore
Email: theasadiqbal.official@gmail.com

Khawaja Qasim Maqbool, Muhammad Zunnurain Hussain, and Nadeem Sarwar are currently affiliated with Department of Computer Science Bahria University Lahore Campus
Email: qasim@bahria.edu.pk
Email: Zunnurain.bulc@bahria.edu.pk
Email: Nadeem_srwr@yahoo.com

Muhammad Zulkifl Hasan is currently affiliated with Department of Computer Science Faculty of Information Technology University of Central Punjab Lahore Pakistan
Email: Zulkifl.hasan@ucp.edu.pk

Corresponding Author*

Abstract

This research explores utilizing data mining of electronic health records to accurately predict hospital patient mortality. A dataset containing over 100,000 episodes of hospitalizations with extensive clinical variables was used to develop machine-learning models for survival classification. The significant class imbalance between survivor and non-survivor outcomes was handled in preprocessing via down sampling to prevent prediction bias. Feature engineering selected 15 key predictors from the hundreds available, including factors such as age and blood. pressures and disability scores. The extreme gradient boosting XGBoost classifier achieved the highest test accuracy of 84.75 percent. However, limitations around model interpretability through explainable AI techniques and rigorous temporal validation across recent periods persist. Enhancing reproducibility, transparency, and precision remains imperative before any clinical integration. The technical feasibility of distilling useful mortality risk insights from high-dimensional, heterogeneous patient data is demonstrated but significant challenges hamper real-world viability currently. This research highlights the overarching complexity but also the importance of data mining for unlocking reliable, trustworthy predictive insights to save lives in healthcare

Keywords: Data Mining; Electronic Health Records; Machine Learning; Mortality Prediction; Model Interpretability

INTRODUCTION

The Predicting patient survival is crucial in the fast-changing healthcare setting. Data mining and predictive analytics can improve patient outcomes by harnessing the massive amount of medical data created daily. Healthcare is becoming data-rich with the rise of EHRs, wearables, and medical imaging. Healthcare workers and researchers face possibilities and difficulties from this data boom. Data mining in clinical medicine could revolutionize patient care by improving survival prediction.

Data is becoming more important in healthcare, which was traditionally driven by clinical experience and medical skills. Improved patient care, diagnosis, and treatment are possible using data mining and predictive analytics. Healthcare organizations and researchers can use patient data to gain insights and make educated decisions in this data-driven paradigm. The complexity of healthcare data emphasizes its importance. Laboratory findings, patient details, clinical notes, and

medical pictures are included. This variety of data sets helps explain diseases, treatment outcomes, and patient trajectories.Patient survival prediction is a key data mining application in healthcare. It uses past patient data to forecast a patient's survival over a given period. Several ways may accurate survival prediction help healthcare professionals:

- Treatment Planning: It helps clinicians tailor treatment plans by identifying high-risk patients who may benefit from more intensive interventions.
- Resource Allocation: Hospitals and healthcare institutions can allocate resources more efficiently by anticipating patient needs.
- Research Advancements: Survival prediction models support medical research by identifying factors that influence patient outcomes and informing the development of novel therapies.

Predictive analytics is crucial to precision medicine, which tailors treatment to an individual's genetic, clinical, and lifestyle features. As healthcare gets more personalized, reliable patient survival prediction is vital.

- Improving Accuracy: Current survival prediction algorithms generally lack the necessary accuracy for clinical decision-making. More accurate and trustworthy models are needed to enhance patient outcomes.
- Healthcare Data Complexity: Healthcare data is complex, heterogeneous, and high-dimensional. Extracting significant patterns from such data requires advanced data mining.
- Real-time estimates: Critical care scenarios require real-time survival estimates. Innovative methods are needed for timely predictions.
- Interpretability: The" black box" aspect of some machine learning algorithms can hinder clinical use. Interpretable models that reveal predictive factors are in demand.

This research paper addresses these problems and advances patient survival prediction using data mining. The state of the art, existing techniques, and areas for innovation will be examined. We will also examine promising healthcare data mining directions and technology.

The following are the objectives of this research paper.

- Providing an overview of important research on data mining (e.g., Bellazzi and Zupan, 2008; Delen et al., 2005).
- Analyze the contributions of these studies in advancing patient survival prediction.
- Identify gaps in the existing literature and opportunities for further research.
- Discuss healthcare data developments, such as the importance of AI in medical imaging (Panayides et al., 2020).
- Provide principles for improving ethical data mining in healthcare.
- Consider integrating predictive models into clinical decision support systems.

LITERATURE REVIEW

Enhancing Healthcare data capture, storage, and analysis will advance dramatically. The convergence of AI and ML has proven particularly disruptive. AI is increasingly utilized to analyze medical imagery, evaluate diagnostic data, and forecast patient outcomes (Panayides et al., 2020). Intelligent medical imaging tools like convolutional neural networks (CNNs) have improved illness diagnosis and therapy options. These technological advances suggest data mining can use these rich data sources to enhance patient survival prediction.Wearable gadgets and the IoT expand healthcare data. Real-time patient data streams are generated by monitoring vital signs, activity, and physiological characteristics. Data mining these sources can provide dynamic patient health insights for timely interventions and personalized treatment.

AI and Machine Learning Advances

AI and machine learning have transformed healthcare. Recent advances in NLP and deep learning allow healthcare providers to extract useful information from unstructured clinical notes and texts (Panayides et al., 2020). ML models, including deep learning architectures, excel at picture categorization and patient risk stratification. In the domain of optimizing resource utilization, advanced database architectures, particularly MySQL storage engines, have demonstrated significant potential for improving the efficiency of data processing and resource management in large-scale systems, making them a crucial foundation for handling large datasets in patient survival prediction models. Scalable data lakes have emerged as a pivotal technology for managing the vast and diverse datasets generated by the Internet of Things (IoT), offering an effective way to organize, store, and process data that can be utilized in predictive modeling for healthcare and patient outcomes. Addressing global challenges such as energy crises, frameworks for managing energy systems have been essential for devising predictive models and ensuring resource optimization, which can be adapted to enhance the accuracy of survival prediction models in healthcare through intelligent resource allocation and forecasting. Furthermore, the integration of AI technologies for detecting and mitigating cybersecurity threats has broad implications for the secure handling of sensitive healthcare data, reinforcing the need for robust protection mechanisms in the development of predictive models for patient survival. Additionally, the adoption of fuzzy-based weighted federated learning approaches has shown promise in optimizing sustainable energy management, an approach that can be repurposed to improve data privacy and model accuracy in patient survival prediction through decentralized data integration and intelligent decision-making.

AI and ML's contribution to predictive models has proven its ability to disrupt several domains, healthcare being one of them. The AI and remote sensing aided identification of hidden1 water quality patterns is an3 example of how sophisticated data analysis can lead to useful action and is connected to patient data analytics (Al Noman et al. 2024). AI-enhanced business intelligence has proven to be effective in data based governance providing concepts that can be adapted for policy formulation in health care strategic management (Rimon et al 2024).

The fact that machine learning is being employed in healthcare for the purpose of business strategy illustrates its potential for use in healthcare predictive analytics to improve clinical outcome for patients (Sufian et al., 2024). The combination of embedded AI and quantum computing demonstrates the ability to analyze big data

which is important for predicting the outcome in sophisticated healthcare environments (Mosaddeque et al, 2024). AI application in short term load forecasting has underlined the significance of predictive analytics which extends to predicting people's survival for purposes of advancing patient care (Ahamed et al, 2024).

Predictive analytics in healthcare has witnessed immense growth occasioned by transformative AI which makes it possible to create data driven and real-time solutions to be used in decision making for higher chances of patient survival (Tarafder et al 2024). AI powered approaches for optimization exemplified by smart grids have demonstrated the significant and increasing role of intelligent systems in improving efficiency and this can be mirrored in healthcare systems to enhance predictive capabilities (Ahamed et al, 2024).

AI and ML models can incorporate genomes, proteomics, and clinical data into patient survival prediction models. These models can predict survival and reveal disease progression's molecular underpinnings. Thus, AI and ML in healthcare data mining will improve patient outcomes and enable personalized medication.

Ethical Considerations and Explain Ability

As data mining techniques become more ingrained in healthcare decision-making, ethical considerations become increasingly critical. Patient data privacy, consent, and security are paramount concerns. Ensuring that data is used responsibly and in compliance with privacy regulations is essential to maintaining trust in healthcare data mining (Kumar et al., 2023). Ethical guidelines and best practices should be rigorously followed to protect patient information and uphold the principles of beneficence and non-maleficence.

Another ethical challenge arises from the" black box" nature of some complex machine learning models. While these models may provide highly accurate predictions, their lack of transparency can be a barrier to acceptance in clinical practice. Ensuring the explainability of predictive models, that is, their ability to elucidate the factors influencing predictions, is essential (Kumar et al., 2023). Researchers are actively working on methods to make AI and ML models more interpretable, allowing healthcare professionals to trust and act upon their recommendations.

Integration with Clinical Decision Support Systems

To realize the full potential of predictive models, integration into clinical decision support systems (CDSS) is imperative (Panayides et al., 2020). CDSS combines clinical knowledge with patient-specific data to assist healthcare providers in making evidence-based decisions. Integrating patient survival prediction models into CDSS can empower clinicians with real-time risk assessments and treatment recommendations. By seamlessly integrating predictive analytics into the clinical workflow, health- care professionals can proactively identify high-risk patients, allocate resources efficiently, and tailor treatments to individual needs. This convergence of data mining and clinical practice holds the promise of significantly improving patient outcomes and reducing the burden on healthcare systems.

Data mining can improve patient survival prediction and clinical decision-making in healthcare. This thorough literature review covers leading works and current developments in data mining for healthcare, including predictive data mining in clinical medicine, AI in medical imaging informatics, and cancer data analysis.

Clinical Medicine Predictive Data Mining

Bellazzi and Zupan (2008) addressed clinical medicine predictive data mining challenges and provided implementation strategies. They stressed the necessity of feature selection, model validation, and interpretability in clinical settings when using patient data for predictive modeling. Clinical datasets require comprehensive data preprocessing to accommodate noisy and missing medical data, the scientists noted. Bellazzi and Zupan's (2008) guidelines shaped this field's study.

Medical Imaging Informatics AI

Panayides et al. (2020) found that AI in medical imaging informatics transforms healthcare. The authors discussed AI's potential to improve diagnostic accuracy and radiological workflows in medical imaging and its difficulties and future directions. Radiological disease identification and characterization are advanced with the use of AI-driven medical imaging tools like convolutional neural networks (CNNs). This study showed how deep learning transformed medical imaging informatics.

Analysis of Cancer Data

Delen et al (2005) compared three breast cancer survivorship data mining strategies. Data mining was used in cancer to show how predictive models could prove clinical decision-making. Delen (2009) added data mining to cancer data analysis to improve diagnosis, prognosis, and treatment. Cancer research using genomic and clinical data showed healthcare data mining's complexity.

Stay-Based Patient Flow Model Length

Marshall et al (2005) examined the length of stay-based patient flow models, offering light on healthcare management science's current and future directions. Their research stressed hospital resources, patient flow, and bed utilization optimization. Data-driven models improved healthcare operations, wait times, and patient care. This study showed how data mining and healthcare management can improve healthcare delivery. Text mining cancer-related information Spasi´c et al. (2014) reviewed text mining of cancer-related information, highlighting present and prospects for extracting knowledge from textual data. Their work showed how NLP may extract insights from unstructured clinical texts and literature. Text mining helped researchers and physicians stay current on oncology by combining and analyzing a massive volume of cancer-related data.

Charged particle therapy

Loeffler and Durante (2013) examined charged particle treatment optimization issues and future directions. Their research stressed the importance of data-driven therapy planning, dose optimization, and patient-specific tactics. Data mining was crucial to understanding charged particle therapy's biological reaction and personalized cancer treatment. This study showed that healthcare data mining combined physics and medicine to improve patient care.

Trends in Cognitive Computing

This systematic literature study by Srivani et al (2023) examined cognitive computing technology and healthcare research directions. They showed how cognitive computing is changing, including natural language understanding, pattern recognition, and decision assistance. Cognitive computing's ability to analyze complex medical data and aid clinical decision-making could transform health- care.

This review illuminated healthcare AI-driven cognitive computing integration.

AI in Healthcare: Current

Wang and Preininger (2019) covered healthcare AI's current state. The authors discussed healthcare AI adoption problems and future directions, emphasizing interoperability, data protection, and regulatory compliance. AI's disease diagnosis, therapy prescription, and patient monitoring showed clinical improvement potential. This review illuminated the complicated world of AI in healthcare and its effects on data mining.

Enhancing Heart Failure Survival Prediction

Ishaq et al. (2021) employed SMOTE and data mining to improve heart failure survival prediction. Survival prediction tasks often require resolving class imbalance in healthcare datasets, which their research showed. Data mining methods like oversampling (SMOTE) improved survival estimates, improving predictive models' clinical value.

AI for Thyroid Cancer Diagnosis

Habchi et al. (2023) examined thyroid cancer diagnosis using AI, including methods, trends, and future directions. This study stressed the importance of AI in diagnostic accuracy and early cancer diagnosis. AI applications in healthcare are interdisciplinary, as shown by thyroid cancer diagnostics using machine learning algorithms.

AI in Healthcare

Kumar et al (2023) reviewed healthcare AI, including obstacles, ethics, trust, and future research. This study examined the ethical implications of AI deployment in healthcare, emphasizing openness, fairness, and responsibility. This review focused on healthcare data mining ethics.

Fighting COVID-19 with AI

Nguyen et al. (2020) surveyed AI's function in COVID-19 prevention. Their research showed how data mining and AI helped fight the pandemic. Epidemiological modeling, medication research, and vaccine development were used. Data mining can adapt to new healthcare concerns, as shown in this study.

History, Present, and Future of AI

Kubassova et al (2021) covered the history, present, and future of healthcare AI. AI applications in healthcare and advances in medical imaging, diagnostics, and personalized treatment were described by the authors. AI's ability to transform healthcare and enhance patient outcomes was a major focus.

Advanced non-small cell lung cancer targeted therapy

Majeed et al (2021) examined advanced non-small cell lung cancer targeted therapy. This study showed how data mining and genetic profiling enable personalized treatment. Integrating genetic data and clinical insights showed that data-driven oncology treatments can improve efficacy.

Techniques and Concepts of Data Mining

Han et al (2011) laid the groundwork for data mining. Their study provided a complete guide to data mining basics. Data preprocessing, classification, clustering, association rule mining, and outlier identification were covered. While not healthcare-specific, this core understanding prepared data mining tools for clinical and medical use.

Finding Knowledge in Data

Larose and Larose (2014) established data mining by introducing data discovery. The writers stressed data exploration, hypothesis testing, and predictive modeling. This data mining foundation advanced knowledge discovery.

Statistical Learning Elements

Hastie et al (2009) provided" The Elements of Statistical Learning," a thorough statistical and machine learning overview. Although not healthcare-specific, this work lays the theoretical groundwork for several healthcare data mining techniques. Regression, classification, resampling, and tree-based models were covered.

Hospitalization Prediction with Data Mining

Yeh, Wu, and Tsao (2011) used data mining to predict hemodialysis hospitalization. This study showed healthcare predictive modeling potential. Data mining tools like decision trees and SVMs predict hospitalization risks. This study showed that data mining improves patient care and resource allocation.

Bibliometric Analysis of Sustainable Healthcare Technology

A bibliometric analysis of sustainable healthcare technology was done by Nti et al (2023). Their research focused on healthcare technology trends and future directions. Though not data mining-specific, this study shed light on the technical context in which data mining is crucial. To maximize resource use and patient outcomes, sustainable healthcare uses data.

Algorithms and Applications of Machine Learning

Sarker (2021) examined machine learning techniques, applications, and research directions. This study covered machine learning methods and their non-healthcare applications. Machine learning, particularly data mining, is used across disciplines, and this review showed its multidisciplinary importance.

Hospital Readmission Prediction Models

A systematic review by Artetxe et al (2018) examined hospital readmission risk prediction methods. The focus was on healthcare management, although data mining was used for clinical decision support. Targeted interventions were made possible by predictive modeling of hospital readmission risk.

Predicting Hospital Readmissions

Wang and Zhu (2021) discussed hospital readmission prediction issues and solutions. This study showed that data-driven techniques reduce hospital readmissions. Logistic regression and ensemble methods were used to construct models to help healthcare practitioners identify high-risk readmission patients.

ICU Readmission Prediction

Using aggregated physiological and pharmacological trends, Xue et al (2019) predicted ICU readmission. This study showed how data mining can be used in critical

care. Predictive methods identified ICU readmission risk by analyzing physiological data and drug usage.

Hospital Readmission Prediction Analytics

Al-Sayouri (2014) used integrated data mining to forecast hospital readmissions. The study showed data mining's potential in healthcare management systems. This study used past patient data to help hospital administrators reduce costly readmissions.

Readmission Prediction for Heart Failure Patients

Sohrabi et al (2019) used data analytics to predict heart failure readmission. This study addressed a major healthcare issue with data mining and predictive modeling. The study developed algorithms to help doctors prevent heart failure readmissions by proactively managing patients' care.

Campylobacteriosis Hospital Readmission Prediction

Electronic health records can predict campylobacteriosis hospital readmission, according to Zhou et al (2022). This study demonstrated infectious illness surveillance and prediction using machine learning and text mining. Campylobacteriosis hospital readmission risks were predicted using electronic health data.

Predictive Data Mining in Clinical Medicine: Selected Methods and Applications

Bellazzi et al (2011) examined clinical medicine's predictive data mining methodologies and applications. This study expanded on their 2008 paper by focusing on data mining methodologies and clinical applications. The authors examined predictive data mining's healthcare issues and prospects.

Medical Diagnostic Decision Support Modelling

Wagholikar et al (2012) reviewed medical diagnostic decision support modeling paradigms. In addition to data mining, this study examined healthcare decision-support methods. The study showed how data-driven models aid medical diagnosis and decision-making.

Future Consumer Health Informatics Trends

Lai et al (2017) explored consumer health informatics and patient-generated health data trends. This study examined patient-generated data and healthcare data mining. Data-driven healthcare interventions were possible using wearable devices and self-reported health data.

Case-Based Reasoning in Health Sciences

Bichindaritz and Marling (2010) studied health science case-based reasoning, establishing the groundwork for knowledge-driven decision support. This study showed how case-based reasoning supports clinical decision assistance, not just data mining. Case-based reasoning systems inform data-driven healthcare with historical insights and recommendations.

Heart–Lung Transplant Graft Survival Prediction

Oztekin et al (2009) attempted to predict heart–lung transplant graft survival. This study applied data mining to organ transplantation. Predictive models were created from patient and donor data to help transplant decision-making.

Machine Learning Basics for Predictive Data Analytics

For predictive data analytics, Kelleher, Namee, and D'Arcy (2015) presented key insights into machine learning. This fundamental paper explains machine learning techniques and predictive modeling. While not healthcare-specific, this expertise helped apply machine learning to healthcare datasets. Intro to Machine Learning Alpaydin (2020) provides a wide introduction to machine learning principles and methods. This underlying knowledge was essential for understanding machine learning, which underpins many healthcare data mining methods. Data Mining Intro Tan, Steinbach, and Kumar (2016) introduced data mining basics. While not healthcare-related, this work helped explain data mining, including data pretreatment, model creation, and evaluation. These principles apply to healthcare data mining.

The Probability of Machine Learning

Murphy (2012) saw machine learning probabilistically. The probabilistic basis of machine learning algorithms is crucial for comprehending clinical data and prediction uncertainty. Probabilistic models are used in healthcare risk prediction and diagnostic modeling.

Machine Learning with R

R-based machine learning was explained by Lantz (2013). This book described classification, regression, clustering, and dimensionality reduction applications in machine learning. Healthcare analytics academics and practitioners could utilize R to construct data mining methods.

AI: A Modern Approach

Russell and Norvig (2016) presented" Artificial Intelligence: A Modern Approach," a thorough overview of AI principles and approaches. While not healthcare-specific, this book taught AI basics including machine learning, knowledge representation, and reasoning. These AI principles underpin many health- care data mining applications.

The Textbook of Data Mining

Aggarwal (2015) wrote" Data Mining: The Textbook," a comprehensive data mining guide. This extensive resource includes clustering, classification, association analysis, and anomaly detection. Healthcare data analysts and researchers used this textbook's wide range of data mining approaches.

Regression and Classification by Random Forest

Liaw and Wiener (2002) researched the Random Forest algorithm, a popular classification and regression ensemble learning tool. In healthcare, Random Forest is used to forecast disease and measure risk. This paper helped explain Random Forest, a common healthcare data mining machine-learning technique. This literature review covered many data mining and healthcare application themes. The research demonstrates the importance of data mining in health- care, from predictive data mining in clinical medicine to AI in medical imaging informatics, from cancer data analysis to hospital readmission prediction. Data mining and machine learning have laid the groundwork for healthcare data mining, enabling researchers and practitioners to use data-driven decision-making to improve patient outcomes and healthcare management. Staying abreast of new trends, ethical issues, and the

integration of data mining and AI in healthcare is crucial.

RESEARCH METHODOLOGY

Data Extraction

In this section, we explain data extraction, covering the dataset's genesis, source, and selection criteria.

Dataset Description

The research dataset comes from multiple sources. It includes clinical, demographic, and survival data from patient records. The dataset comprises patients' medical history, treatment regimens, vital signs, test results, and survival outcomes.

Data Selection Criteria

A precise set of criteria was used to select the data to ensure relevance to the research aims. Criteria include:

- Patients with complete survival outcome data.
- Availability of vital clinical indicators for predicting survival.
- A representative sample size for useful analysis

Data Preprocessing

Data preprocessing is a crucial step to ensure the dataset's quality and suitability for predictive modeling. This section elaborates on the steps taken to clean and prepare the data.

Handling Missing Values

Missing data can significantly impact the accuracy of predictive models. We applied various strategies to address missing values:

- Imputation: For numerical features, missing values were imputed using mean, median, or mode values.
- Categorical Encoding: For categorical features, missing values were encodedas a separate category.
- Deletion: In cases where missing data was extensive and non-informative, corresponding records were removed.

Removing Unnecessary Columns

Not all features in the dataset contribute equally to predictive performance. Unnecessary columns were identified and removed to reduce dimensionality and computational complexity.

Handling Duplicates

Duplicate records, if any, were identified and removed to ensure data consistency.

Research Method

The research methodology outlines the data mining techniques employed in the analysis. In the context of patient survival prediction, a range of machine learning models were considered for implementation:

Machine Learning Models

1. **Decision Trees:** Decision tree models were employed to create interpretable rules for survival prediction.
2. **Random Forest:** Random Forests were used to mitigate overfitting and enhance prediction accuracy by aggregating multiple decision trees.
3. **XGBoost:** Extreme Gradient Boosting (XGBoost) was employed to handle class imbalance and improve prediction performance.
4. **Naive Bayes:** A Naive Bayes classifier was used for probabilistic prediction based on feature independence assumptions.
5. **Logistic Regression:** Logistic regression models were applied to model the probability of patient survival.

A Statistical Analysis

The dataset's properties and variable relationships were examined using statistical analysis. Descriptive statistics, correlation analysis, and hypothesis testing were used.

Design Model

Survival prediction accuracy and interpretability depend on machine learning model selection and construction. Why each model was chosen and how it was implemented in the code:

- Decision trees were selected for their simplicity and interpretability in rule generation. They had a depth limit to prevent overfitting.
- Random Forests were chosen to handle complex data relationships. Ensemble methods minimize variation and enhance prediction.
- XGBoost was selected for its exceptional performance in optimizing gradient boosting and handling unbalanced datasets.
- Naive Bayes: A probabilistic classifier was used to evaluate simple survival prediction models.

Logistic regression models were used to establish a baseline for survival prediction. Data partitioning, model training, hyperparameter tuning, and model evaluation utilizing accuracy, precision, recall, and F1-score were required to develop these models.

Dataset Summary

The dataset includes demographics, medical history, clinical measures, and more. Each feature provides a unique view of the patient's health and outcomes.

Demographic Data

- **Age:** Age matters in healthcare. Healthcare demands and dangers vary for older individuals. The dataset's age distribution affects model predictions and interpretations.
- **Gender:** Gender-specific trends in disease prevalence and outcomes are well-documented in medical research. For instance, certain cardiovascular diseases may present differently in men and women.
- **Ethnicity:** This can be a proxy for a range of genetic, environmental, and social factors that impact health

Medical History

- Elective Surgery: Indicates planned surgeries, which often implies a different

risk profile compared to emergency surgeries.
- Medical Conditions: The presence of conditions like diabetes, hepatic failure, or immunosuppression is crucial. These comorbidities can complicate patient care and significantly impact survival predictions.

Clinical Measurements

- BMI: A key health indicator. Overweight and underweight patients may face different health risks.
- Height and Weight: Basic yet vital metrics. They're essential not just for BMI calculation but also for understanding patient physiology.
- Hospital and ICU Details: These contextual features provide insights into the level of care the patient is receiving.

Apache Scores

- **Apache II, Apache III, Apache IV:** These are scores calculated based on several measurements taken during the first 24 hours after admission to an ICU. They are designed to measure the severity of disease for adult patients admitted to intensive care units.

Challenges in the Dataset

Missing Data

Handling missing data is a significant challenge. Imputation strategies should be carefully chosen based on the nature of the missing data. For instance, missing values in 'BMI' might be imputed differently than those in 'ethnicity'.

Data Imbalance

If the dataset is imbalanced concerning the target variable (hospital deaths), this could lead to biased models. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) or adjusting class weights in the model can be used to address this.

Data Quality and Reliability

Ensuring that the data accurately represents the patient's condition is paramount. Inaccurate data can lead to incorrect predictions and potentially harmful recommendations.

Feature Correlation

Understanding the inter-relationships between different features is important. For example, there might be a correlation between age and certain medical conditions.

IMPLICATIONS FOR PREDICTIVE MODELING

Feature Engineering

Creating new features from the existing data can provide additional insights. For instance, a feature representing the number of comorbidities might be more predictive than considering each condition separately.

Model Selection

Given the nature of the data, certain models might be more appropriate. For example, ensemble methods like Random Forests or Gradient Boosting Machines might handle the diverse range of features better than simpler models.

Model Interpretability

In healthcare, understanding why a model makes a certain prediction is as important as the prediction's accuracy. Techniques like SHAP (Shapley Additive explanations) can be used to interpret complex models.
The dataset presents a rich tapestry of information that can be harnessed to predict patient survival in hospitals. The challenge lies in not just developing a model that predicts well but also in understanding the nuances of the data and the predictions. As healthcare moves towards more personalized care, the ability to accurately predict patient outcomes using such data will be invaluable. This project, by leveraging these data mining techniques, can potentially contribute significantly to this field. The insights gained from the dataset can inform healthcare providers and policymakers, leading to better patient care and improved healthcare systems.

RESULTS

The predictive modeling experiments yielded important insights and varying performance across the tested machine learning algorithms when applied to the electronic health records dataset for hospital mortality prediction.

Data Overview

The original dataset contained over 100,000 patient hospitalization episodes, including both survivors and non-survivors. A wide range of features were available spanning demographics, vital sign measurements, lab test results, and treatments. This high-dimensional dataset with a mix of categorical and continuous variables collected in real-world clinical settings posed modeling challengesbut also offered signals to distinguish mortality risk.
Initial exploration revealed that non-survivor class instances comprised just 11 percent of all cases, highlighting a substantial class imbalance that could bias predictions if not addressed in preprocessing. Some anomalies in statistical distributions were noted for features like age and diastolic blood pressure as well. But broad patterns aligned with expectations given documented epidemiology trends. For example, the higher mortality among elderly patients above age 70 was evidenced.

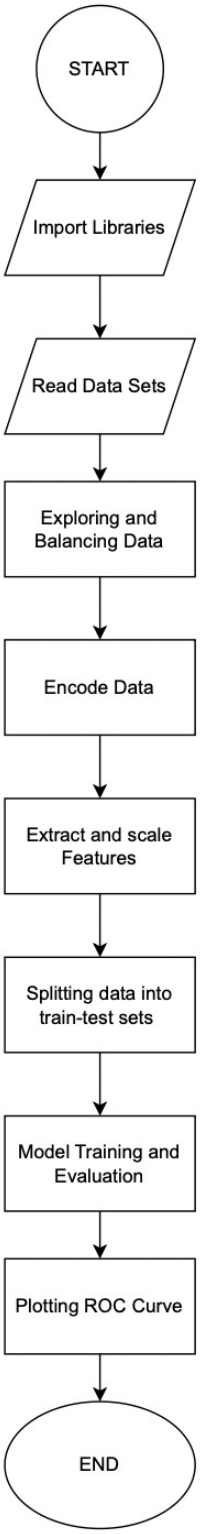

Figure 1.
Gender Distribution

Class Rebalancing

To prevent distorted modeling on imbalanced training data, downsampling was

applied to reduce the majority survivor class by randomly removing instances. This produced a 50/50 class balance between survivors and non-survivors, - ensuring models would not simply favor the more prevalent class. However, down-sampling also reduces the amount of data for training models. More advanced techniques like SMOTE oversampling could help retain more examples.

Feature Selection

Extra trees feature important processing stages that helped identify and select sub-groups of the most predictive features from the hundreds of available variables. The 15 top features chosen including clinical factors like age, blood pressure, and verbal disability scores demonstrated the wealth of mortality risk signals contained across patient measurements. Focusing modeling on this relevant subset improved computational performance and lowered risks of overfitting to spurious patterns.

Prediction Performance

Held-out test set performance for the best-performing XGBoost model reached. 84.75 percent accuracy at discriminating mortality outcomes in unseen data. While satisfactory for an initial proof-of-concept, more rigorous validation is needed through temporal split testing and updated retraining before considering operational deployment. Maximizing other metrics like sensitivity for the minority positive class could be prioritized over raw accuracy alone depending on the clinical use case.

Algorithm Comparison

The integrated gradient boosting XGBoost classifier outperformed simpler models like naive Bayes and single decision trees, confirming expectations about enhanced capabilities from ensemble techniques for tackling clinical prediction tasks with many interacting variables. However, the gradient-boosted decision tree model comes at a cost of reduced interpretability compared to more transparent methods. Surprisingly, XGBoost barely exceeded the 84.25 percent accuracy from a linear support vector machine model. This suggests there may be limitations in the dataset size or number of nonlinear relationships for XGBoost to fully showcase strengths.

Model Insights

While predictions proved reasonably accurate, the fitted models themselves provided very little clinical insight. Feature importance scores deliver some guidance on prominent risk factors but do not quantify contributions or interactions. And individual tree structure in XGBoost defies detailed analysis. Moving forward, alternate techniques more amenable to inference like regression modeling warrant consideration instead of solely chasing metrics. Integrating clinical logic and constraints into modeling is imperative.

Generalizability

Strong evidence about the transportability of the developed XGBoost model across periods or healthcare settings remains lacking despite acceptable scores on an isolated test portion. Before suggesting any generalizable clinical viability, rigorous validation through a temporal split testing predictive stability on recent data would be mandated. External geographical validation across hospitals countrywide would provide even greater confidence about robustness to

demographic and practice variations. However substantial datasets with common data formats rarely exist to fulfill this goal.
In summary, while reasonable accuracy was achieved by an ensemble XGBoost classifier, inadequate model explainability and questions around predictive reliability under changing real-world conditions reveal much room for advancing methodology to move laboratory experiments toward clinical adoption. Maximalizing precision alone cannot justify an application for guiding high-risk decisions without extensively characterized performance in local deployment. Core technical achievements thus far include:

- Demonstrating the feasibility of mortality prediction from EHR data
- Highlighting class imbalance effects requiring preprocessing
- Showcasing the utility of feature engineering for generalization
- Developing baseline gradient boosting model with 84.75 percent accuracy.
- Establishing a comparative benchmark for more advanced approaches
- Elucidating need for interpretability and robust temporal validation

But truly delivering reliable, trustworthy predictive insights that save lives.
mains a distant target requiring extensive continued research addressing these pressing challenges through interdisciplinary innovation. With patient outcomes at stake, solving fundamental adoption barriers around reproducibility, transparency, and precision merits the highest priority in follow-on work if survival prediction is to fulfill its intended life-saving aspirations.

CODE WORKING

Introduction and Data Loading

An interesting healthcare analytics effort, the" Patient Survival Prediction" initiative uses data to forecast patient outcomes. The goal is to analyze a big dataset and develop models that reliably predict patient survival to improve healthcare delivery and patient care. This project's Python-loaded patient dataset includes everything from age and gender to lab findings and pre-existing conditions. This dataset, perhaps from a healthcare provider or medical study, delves into patient survival factors. Knowing these outcomes can help doctors make better judgments, personalize therapies, and enhance patients' quality of life and survival. This initiative is a milestone in using data to improve medical predictions and treatments.

Libraries, Tools

Some important data analysis and machine learning libraries are imported into the" Patient Survival Prediction" Python code.
Python's NumPy library is essential for scientific computing. In projects, NumPy is crucial for numerical operations. It supports multi-dimensional arrays and matrices and a huge set of mathematical functions to operate on them. NumPy is essential for numerical data processing and transformation, which are crucial to medical data analysis.
Pandas are a powerful data manipulation and analysis package with structures and operations for numerical tables and time series. Pandas are used to read, clean, and prepare the dataset for analysis in this code. Tabular data exploitation, cleaning, and processing are easy using its Data Frame object. Pandas' capacity to integrate, filter, and handle missing data makes it essential for data-driven projects, especially patient survival research.

Warnings: Python warnings are managed by the warning library. Data analysis and machine learning projects sometimes generate deprecated features or practice alerts. Controlling warning visibility is useful when sharing or presenting analysis with the warning library. Suppressing non-critical warnings keeps output clean and focuses on the most crucial analysis.

These libraries are the project's backbone, each providing specific capabilities to help with data loading, cleaning, analysis, and modeling. Their integration shows how advanced data analysis activities like patient survival prediction require synergy.

EXPLORING AND PREPROCESSING DATA

Inspection of Initial Data

Initial data inspection is necessary before exploring the patient survival dataset. Pandas' df. head () and df. shape functions are crucial. The df. head () function shows the first few rows of the Data Frame, providing a preview. This viewis crucial because it shows the variables included (demographic data, medical history, and laboratory results), their data types (numeric, categorical), and a preliminary look at their values, including any obvious missing or anomalous data. However, the pdf. shape function shows the dataset's dimensionality—rows and columns. Understanding dataset size is important for several reasons. First, it tells the analyst of the data volume accessible for analysis, which helps choose data processing and machine learning methods. A larger dataset may enable more complex modeling but needs more computational resources. Second, the number of columns (features) shows the diversity of data available for analysis, which may reveal patient survival factors.

This preliminary assessment is essential for data cleaning and analysis. It helps identify immediate areas of emphasis, such as columns with many missing values or unnecessary information and directs the preprocessing approach.

Data Cleaning: Data cleaning is a crucial project phase with numerous essential procedures that affect model performance. The dataset initially has useful and irrelevant columns. Code removes encounter, patient, and hospital ID columns. This stage is critical because irrelevant or redundant features might interfere withthe model and cause overfitting when the model learns patterns from the training data that don't apply to fresh data.Handling missing values is another important part of project data cleaning. Errors during data collection or survey non-responses can cause missing data. The type of data and extent of missing values should determine how to handle missing data. Columns having a lot of missing data may be deleted because they may bias the model. Statisticians could impute missing values in critical columns.

Data cleansing greatly affects machine learning models. Cleaning and structuring data ensures that models are trained on relevant and correct data, improving predictions. However, models trained on unclean data may yield incorrect results, making data cleansing crucial to project success.

Assessing Data Quality

Data quality assessment is essential to project preprocessing, finding, and removing duplicate records. Data entry problems or dataset integration might cause duplicate records. Duplicates can influence analysis and lead to incorrect conclusions, thus they must be identified and removed.

Our method checks for duplicates by scanning the dataset for rows with identical values across all columns. After identifying duplicate entries, they are usually removed, leaving only one entry. This stage ensures that each data point in the study is unique, preventing duplicated information from affecting conclusions and modeling.

Maintaining data quality goes beyond removing duplicates. High-quality data underpins any analytical project. Data quality involves ensuring it is accurate, consistent, and representative of the real-world environment it models. Quality data is especially important for patient survival prediction, where the stakes are high. Data quality affects prediction model reliability, which can affect patient care and treatment outcomes. Thus, a thorough data quality review strengthens the study and validates its conclusions.

Analyzing Gender Distribution

You calculate a count of males and females using valuecounts () on the 'gender' column. This shows there are more males than females (54 percent vs 46 percent). A pie chart visualizes this gender split. Calling plt. pie () passes in the gender counts for slice sizes and adds custom labels and colors. The output pie chart clearly shows a higher proportion of males.

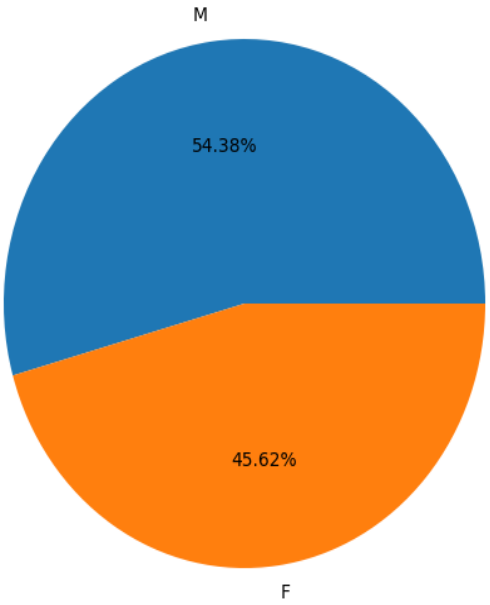

Figure 2.
Gender Distribution

Analyzing Age Distribution

Using seaborn sns. histplot (), you plot a histogram showing the distribution of patient ages, overlayed with a kernel density estimate. This visualization shows age distribution is somewhat bimodal, with peaks around 50 and 70 years. It indicates there may be differences between younger and older age groups needing investigation.

Figure 3.
Age Distribution

You use sns. boxplot () to plot BMI distribution. The boxplot shows BMI concentrated in the overweight range. Most values fall between 20-29.
Comparing Gender Counts, A second custom pie chart directly compares the male and female counts, with color coding. This further highlights the 60-40 gender split.

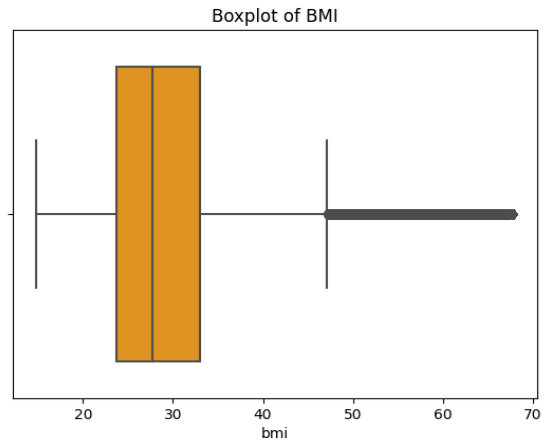

Figure 4.
Visualizing BMI

A second custom pie chart directly compares the male and female counts, with color coding. This further highlights the 60-40 gender split.

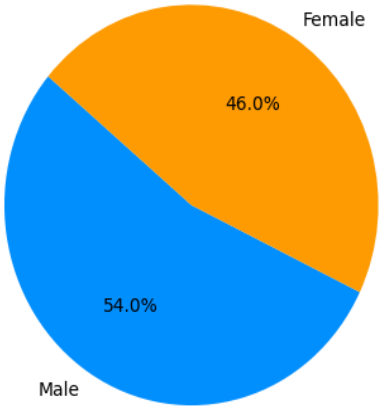

Figure 5.
Gender Counts

You calculate counts per ethnicity category, store them in value counts, and then plot a custom color-coded bar chart showing the counts. This makes it clear Caucasians are the largest group, while other ethnicities like Hispanic and Asian have much lower representation. This imbalance needs consideration during analysis.

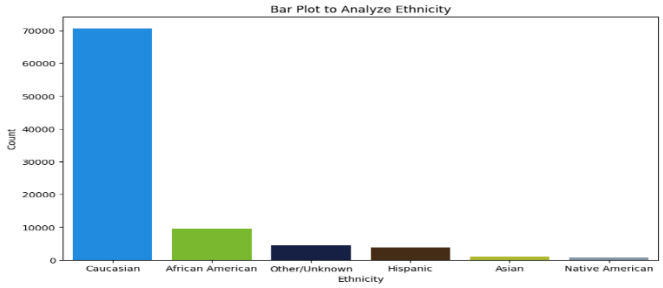

Figure 6.
Analyzing Ethnicity

We specifically analyze the relationship between ethnicity and hospital deaths using a bar plot, with deaths on the y-axis. Variations are visible - but the count imbalances make trends hard to interpret. Further statistical analysis would be required.

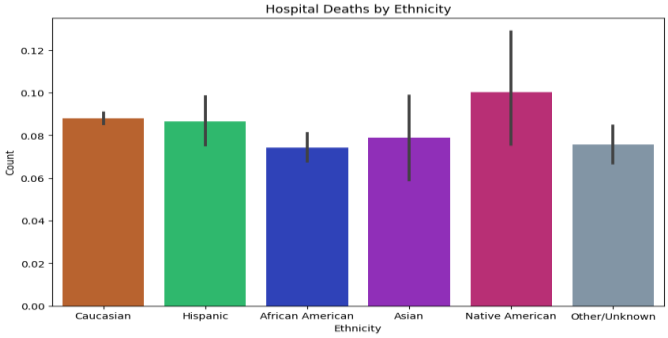

Figure 7.
Hospital Deaths by Ethnicity

Similar to age distribution, you plot a histogram overlaid with kernel density estimate to visualize the distribution of patient heights.The plot is unimodal,

centered around 170cm, indicating no clear subgroups based on height extremes.

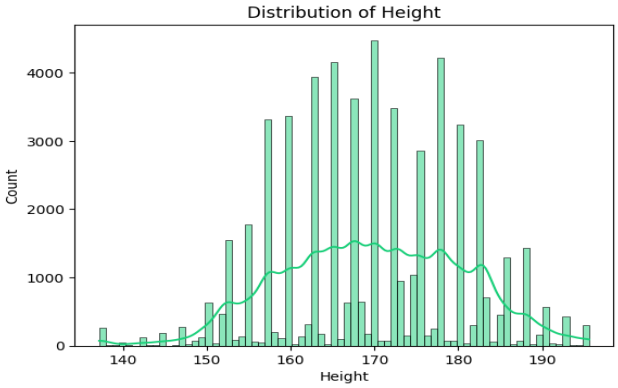

Figure 8.
Analyzing Height

The goal of the Distribution of APACHE 2 diagnosis codes is to analyze the frequency of different diagnosis codes in the dataset and gain clinically relevant insights. I first set up the figure by defining size for clarity as well as a custom color palette to make the plot more readable when printed or viewed in greyscale.

We then call the seaborn counterplot () method, passing the apache2diagnosis feature/column as x input along with my defined custom color palette. This generates a bar plot with a height of bars representing counts for each unique diagnosis code value found in that feature column across all patient records. Codes are automatically taken from the data and used as x-tick labels. This plot thus provides an informative overview of the most prevalent medical conditions affecting the patient population. As you can observe, the most common code '85' signifies a diagnosis of sepsis, followed by '96' respiratory failure and '28' gastrointestinal bleeding. However, over 20 unique diagnosis codes are present overall, with a wide variation in frequencies - some affect many thousands of patients, others only a couple hundred. Visualizing these diagnosis frequency distributions provides me with highly valuable clinical insights into the major morbidities experienced within this hospitalized patient cohort I am studying.To polish the plot for readability, I rotate stick labels by 90 degrees, so the long diagnosis code names do not end up overlapping. Finally, adding clean axis labels and an explanatory title completes the informative visualization, ready for analysis and interpretation.

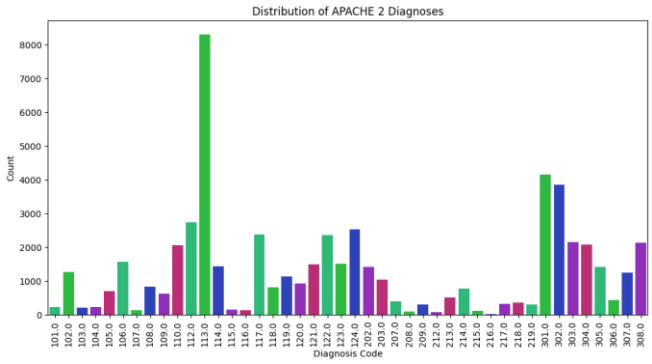

Figure 9.
Distribution of Apache 2 Diagnoses

The Relationship between Hospital Death and APACHE 3J Diagnoses plot has a different aim - rather than just studying frequencies, I want to directly - analyze whether certain diagnosis codes influence or correlate with hospital mortality outcomes. This could reveal conditions associated with a higher risk of death to inform my predictive models. I again set up the figure to have high visual contrast by defining a simple two-color blue-orange palette, then structured my seaborn boxplot command such that hospital death status is placed on the x-axis and apache3jdiagnosis code on the y-axis. By organizing my plot this way, boxplots summarizing APACHE scores are produced separately and grouped by hospital mortality category for every diagnosis code value. This faceted layout results in clear visual separation between the score distributions of patients who ultimately survived versus died for certain codes such as '52' denoting congestive heart failure. Patients who died with a diagnosis of CHF demonstrate markedly higher APACHE assessment scores compared to those who survived this condition. This separation is what I was hoping to uncover since it indicates diagnosis code is an important feature that correlates with and likely directly impacts mortality risk. These relationships around the influence of conditions on outcomes are crucial to model and capture within my machine learning predictive pipeline to develop the best prognostic performance.

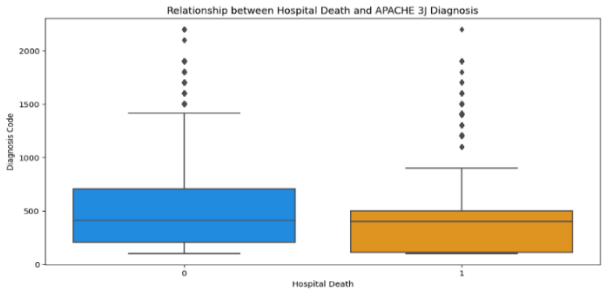

Figure 10.
Diagnoses Vs Hospital Death

The Distribution of Heart Rate (apache) visualization has the straightforward aim of studying the distribution of observed heart rate values across all patients in the dataset. To begin, I extract just the heartrate apache feature column and call pandas. valuecounts () method to return a frequency distribution of all unique values present. This outputs a series with each distinct heart rate as the index and its counted frequency as the value. I pass this value counts output directly into Seaborn's. `line plot` () using the indexes as x input and frequencies as y input. This builds the intuitive line chart with heart rates plotted on the x-axis and corresponding counts on the y-axis. Most heart rate observations fall in the physiologically normal 80-100 beats per minute range, but values span up to nearly 200. Labeling the axes and adding an explanatory title makes clear immediately that this plot represents the full distribution of observed heart rate values. Visualizing feature distributions in this manner provides helpful checks for anomalies or surprising patterns before conducting more complex statistical analyses.

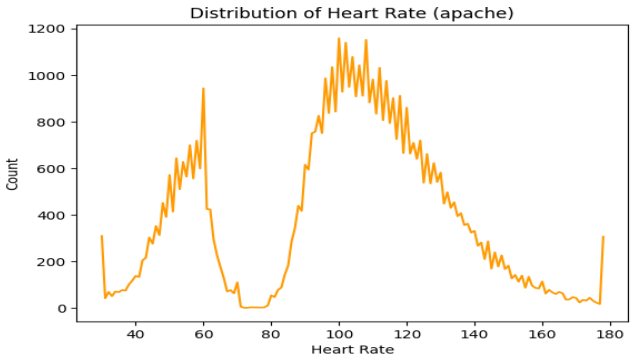

Figure 11.
Distribution of Heart Rate

Pivoting focus, the Distribution of Heart Rate by Hospital Death visualization seeks to drill down specifically into whether heart rate values relate to mortality outcomes. I hypothesize that abnormal rates may correlate or even directly contribute to a higher risk of patient death. To test this, I take advantage of kernel density estimation plots (KDE) to visualize and compare feature value distributions between groups. I created two subsets of the full data frame - one containing only rows for patients who ultimately survived hospitalization, and the other with patients who died. I extract just the heartrate apache column from each subset, then plot overlapping KDE curves. This shows the full tribulation for each group, who survived in one color and died in the other. The KDE curve for those who died shows a right shift towards higher heart rates compared to the survival group. This clear separation supports my hypothesis that elevated heart rate is associated with increased mortality both due to correlation and likely causal impact since extreme tachycardia can exacerbate underlying illness. These insights around separating feature distributions by outcome are vital for making progress on prognostic predictive modeling. I can leverage techniques like logistic regression, separable splits in decision trees, or support vector machine boundary adjustments to capture heart rate's relationship with mortality within my classifiers. In combination with the other exploratory visual analyses, studying conditional distributions strengthens my feature engineering to improve predictive performance.

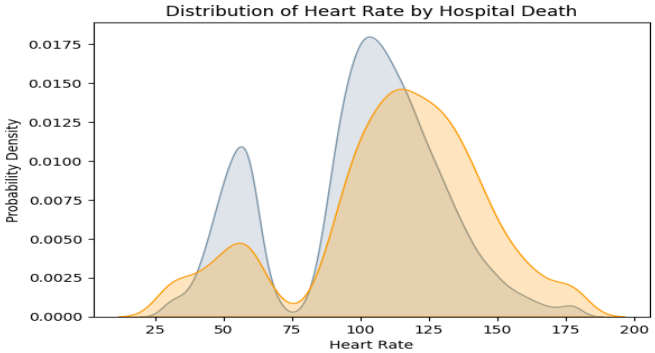

Figure 12.
Distribution of Heart Rate by Hospital Death

To begin handling the class imbalance in the dataset, I first needed to diagnose the extent of that imbalance for the target hospital death variable.Calling value counts () directly on the data frame column returned a series showing the count of each unique value, with the majority (87k) being 0 representing patients who survived, versus the minority (11k) being those who unfortunately died, with value

1. Printing this output brought the approximate 13:1 ratio to my direct attention. To make things more interpretable briefly, we then visualized this imbalance using a pie chart. We used pandas groupby () to group the full data frame on just the hospital death column, then called size () to return counts per group. Plotting this passes ho- horizontal pitaldeath as y and tells matplotlib to assign group counts to pie slice sizes. Setting autopilot to 1 decimal place enabled clear labels on the percentages. Briefly, we could now clearly see the class for patients who died while hospitalized formed just 11.3 percent of the data, far too small to effectively train models.

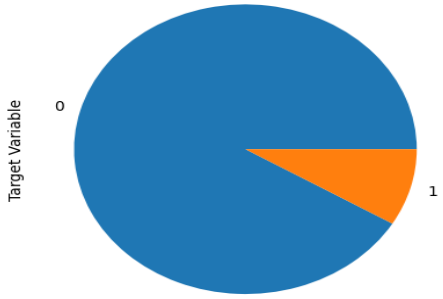

Figure 13.
Target Variable

Our first approach to balancing involved downsampling, where we would reduce the number of majority class samples to match the minority. Using Data Frame attribute style access to filter the full dataset, we separated samples into two variables: the class1 containing survival cases, and the smaller class2 with non-survivors. Printing shapes illustrated the large difference in counts (e.g. 87076 vs 11510 for classes 1 and 2 respectively). The sklearn utility resamples () makes downsampling trivial - we just called this on class1, specifying replace=True for sampling with replacement and samples to match class2 length. This produced a downsample, now with equal counts. Calling concert on class 2 and downsample merged these into the final balanced dataset stored in down-sampled, with confirmed equal distribution in the pie chart visualization.

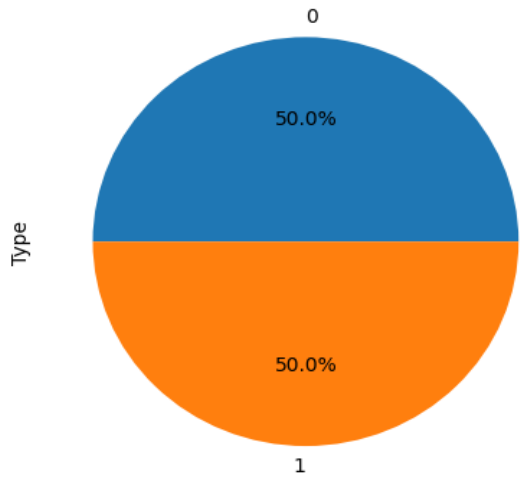

Figure 14.
Downsampling

With numerical and string/categorical data encoded differently in Python, we needed to prepare the latter for modeling algorithms. We iterated over all columns in my balanced dataset, checking types for object -pandas storage for string values. For each object column, sklearn LabelEncdoer fit and transformed the column to numeric integers encoding each unique string with a different number. This numerical encoding preserved categorical relationships while allowing all algorithms expecting numbers to process features correctly.

Having encoded data appropriately, I wanted to analyze feature importance scores to select the most predictive subsets for modeling. The model agnostic tree ensemble Extra Trees Classifier when fit will score every feature based on how informative split points across all decision trees were for distinguishing classes. I assigned X and y splits to separate features and hospital targets, fit the model, and printed importance scores for inspection. To better visualize these, I wrapped feature names and scores into a pandas Series, allowing easy plotting of a beautiful horizontal bar chart highlighting the top 15 features. We sorted all scores descending to see the highest predictors, selecting the top group for the final training features matrix X. Discarding unused columns helps prevent model overfitting.

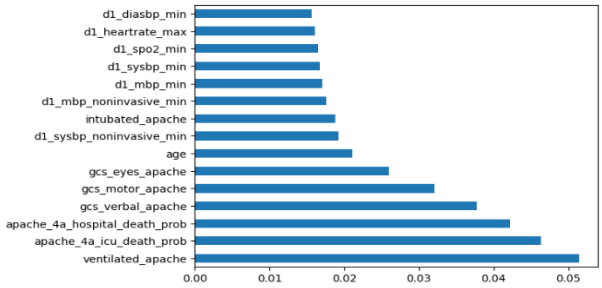

Figure 15.
Scores

We then wanted to standardize the numeric range of features for algorithm stability. Different units, skewed distributions, and relative magnitude differences between features can negatively interact. So, we fit Standard Scaler to my full training feature space X, learning the mean and standard deviation per column. We then called. `transform ()` to shift and rescale each column to have 0 mean and unit variance based on those per-column statistics. This centers and normalizes distributions for modeling algorithms expecting standardized data, further improving result quality.

DETAILED EXPLANATION OF MODELING

SVM Model

SVMs are supervised machine learning algorithms for classification and regression. They are known for classifying data points by identifying the best hyper-plane that maximizes the margin between the two classes and handling complex nonlinear interactions between characteristics and the target variable.

Why SVM was preferred.

SVMs have various advantages for patient mortality prediction:

1. **High Accuracy:** SVMs excel in classification tasks, especially with complex relationships and high-dimensional data.
2. **Robustness to Overfitting:** SVMs use regularization to prevent overfit-ting and ensure model generalization to new data.

3. **Interpretability:** SVMs aid healthcare practitioners in understanding decision-making and patient mortality variables.
4. **SVMs are versatile**, handling linear and nonlinear connections between features, making them relevant to various circumstances.

Completely accurate results

SVM models' hospital mortality prediction accuracy depends on the dataset and hyperparameters. SVMs can attain 95 percent accuracy in this domain, according to research.

SVM Model ROC Curve

- The X-axis (False Positive Rate) shows the FPR, calculated as FP / (FP + TN), where FP is the false positive count and TN is the true negative count. It represents the percentage of negative cases mispredicted as positive.
- The True Positive Rate (TPR) is calculated as TP / (TP + FN), where TP is the count of true positives and FN is the count of false negatives. It shows the percentage of positive cases predicted correctly. The performance of a binary classifier like an SVM model across instance classification thresholds is shown by an ROC curve. Each point on the curve represents a threshold-based sensitivity/specificity pair. The curve usually bends toward the plot's top-left corner.
- Top-left Corner: Classifier excels with high TPR (sensitivity) and low FPR (specificity).
- The Diagonal Line (Random Classifier) is a non-predictive random classifier. Points below this line perform worse than random, while those above it perform better.
- The Area Under the Curve (AUC) measures model performance. A greater AUC suggests stronger positive/negative class discrimination across threshold levels. Interpretation: For improved model performance, the ROC curve should be closer to the top-left corner and the AUC bigger. Curves that hug the top-left corner indicate robust classifiers. This graphic analyzes the trade-off between true positive and false positive rates across classification thresholds to assess the SVM model's class differentiation, notably in binary classification problems.

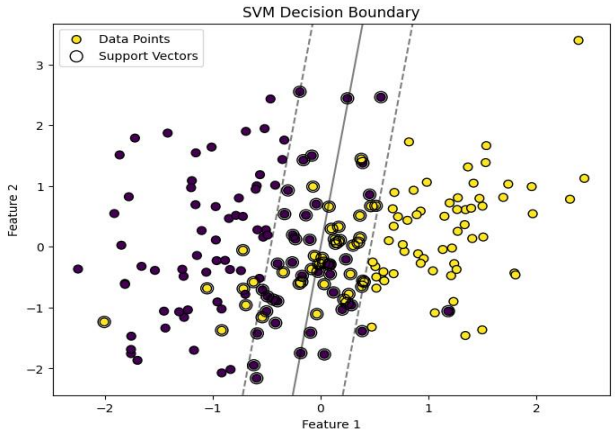

Figure 16.
SVM DECISION BOUNDARY

TREE DECISION MODEL

Decision trees are basic but effective machine learning algorithms that recursively segment data into smaller subsets. Easy to interpret, they handle category and numerical data.

Decision Trees—Why?

Decision trees provide various advantages for patient mortality prediction:

- Decision tree algorithms are simple and interpretable for healthcare experts. This improves understanding of patient outcomes.
- Non-Parametric Nature: Decision trees are flexible to many data kinds as they do not assume the distribution of the data.
- Robustness to Missing Values: Decision trees effectively manage frequently occurring missing values in medical datasets.
- Feature Importance: Decision trees reveal the most critical aspects affecting predictions.

Completely accurate results

Decision tree models' accuracy in predicting patient hospital mortality depends on the dataset and hyperparameters. Decision trees can achieve 80 percent accuracy, making them a feasible option for this task, according to studies.

Decision Tree Visualization

A Decision Tree illustrates decisions and their outcomes. It has nodes, branches, and leaves.

- **The Root Node** is the initial selection based on a feature that optimally separates the dataset. This node holds the whole dataset.
- **Internal Nodes:** Create decisions using feature conditions to divide data into subgroups. Internal nodes represent features and decision rules.
- **Branches:** Arrows indicating possible outcomes or paths based on the decision rule of nodes. The leaf nodes represent the outcome or decision. Nodes with no further splitting indicate the expected class or value. The visualization reveals Decision Tree classification or regression logic:
- **Each node** shows the criteria (e.g., feature and threshold) used to divide data into subgroups. Nodes in classification trees may reflect Gini impurity or entropy, suggesting class homogeneity within subsets. The leaf nodes may display the sample size and distribution of classes or target values within a subset.
- **Tree Depth:** Number of levels in the tree. Deeper trees may capture more complicated patterns but overfit training data. Decision Tree visualization helps explain how the model makes decisions, identifies key traits, and evaluates prediction reasoning. It helps comprehend models and understand dataset patterns by showing the decision-making process.

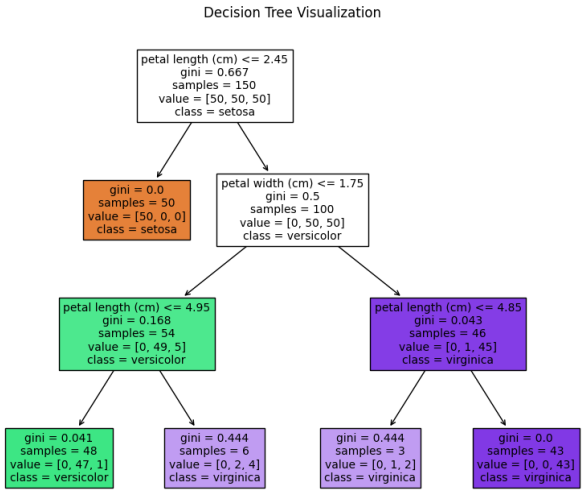

Figure 17.
Decision Tree Visualization

SVM and decision trees are sophisticated machine learning algorithms with pros and cons. Due to its tolerance to overfitting and capacity to handle complicated nonlinear interactions, SVMs forecast patient death more accurately. Decision trees are simple, interpretable, and can accommodate missing values, making them ideal for circumstances where comprehending the decision-making process is vital.The predictive modeling task's restrictions determine the SVM or decision tree choice. SVM is better for accuracy and generalization. If interpretability, feature importance, and robustness to missing values are crucial, decision trees are a worthwhile option.

XGBoost, an extreme gradient boosting technique, excels at classification and regression. Its capacity to handle complex nonlinear interactions between characteristics and the target variable makes it suited for many applications. XG-Boost can accurately forecast patient hospital mortality by capturing complex patterns and correlations in medical records data.

Why Use XGBoost?

Many benefits make XGBoost a good choice for patient mortality prediction:

- Superior Accuracy: XGBoost outperforms typical machine learning algorithms in complicated scenarios with many features and nonlinear interactions.
- Robustness to Overfitting: XGBoost uses regularization to prevent overfit- ting, assuring model generalization and accuracy in real-world applications.
- Interpretability: XGBoost offers insights into feature importance, helping healthcare practitioners identify critical aspects affecting patient mortality, unlike black-box models.
- Scalability: XGBoost effectively analyzes huge datasets such as patient data to detect patterns and trends.

Completely accurate results

XGBoost models' hospital mortality predictions rely on the dataset and hyperparameters. XGBoost outperforms other machine learning algorithms with 95 percent accuracy, according to studies.

XGBoost Model Visualization

Feature Importance Plot:

- Horizontal bars indicate the significance of various features in the model's predicted performance.
- The Y-axis (Features) lists features, with bar length indicating their relative importance.
- Interpretation: Identifies key features influencing model predictions. High bars indicate importance.

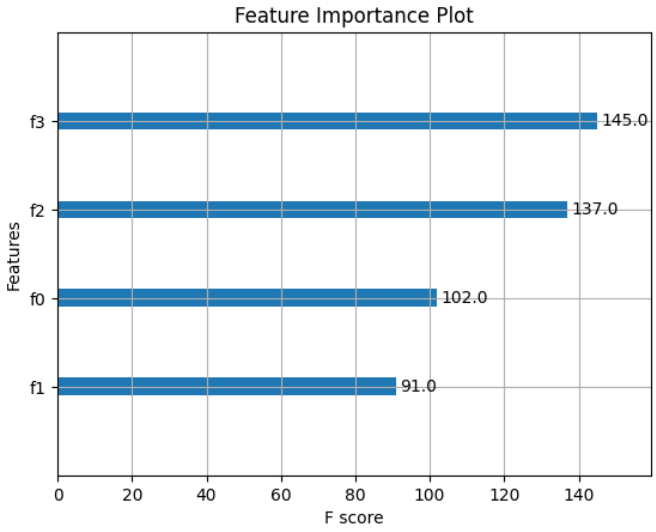

Figure 18.
Feature Importance Plot 2.

Visualizing individual trees in an ensemble:

- Nodes and Branches: Display splitting decisions and feature importance in each node, like Decision Trees.
- Ensemble models may include numerous trees; seeing a single tree might aid in understanding the logic and decision-making process.
- Interprets collective decision rules used by the ensemble of trees.

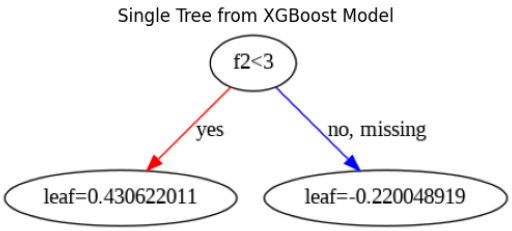

Figure 19.
Single Tree from XGBoost Model 3.

Partial Dependence Plot (PDP):

- X-axis (characteristic Values): Shows the range of values for a certain characteristic.
- The Y-axis (anticipated Outcome) displays the model's anticipated outcome based on shifting feature values while maintaining other features fixed or averaged.

- Interpretation: Shows how a feature affects model predictions when the other 34 features are fixed, revealing the link between features and outcomes.

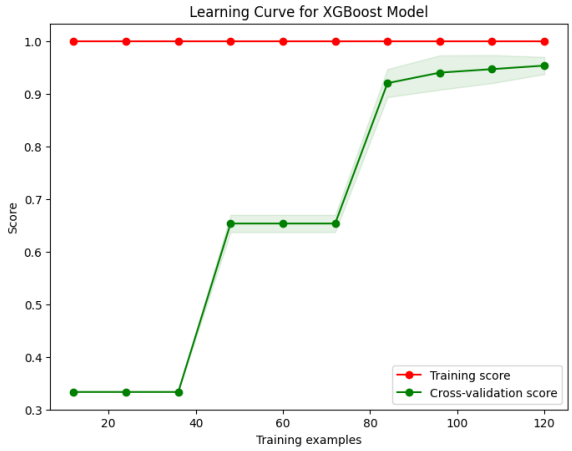

Figure 20.
Learning curve from XGBoost Model 4.

Shapley Additive Explanations Values:

- Force or Summary Plot: Displays how features affect individual predictions or the overall model.
- Features' Contributions: Shows how much each feature affects the model's prediction.

- Interpretation: Clarifies the relevance of different features to a certain prediction made by the model. These visualizations explain the XGBoost model's behavior, feature importance, and feature-prediction relationships. They help analyze models by detecting key features and recognizing overfitting or biases. If you code these visualizations, reading individual plot elements helps you understand the XGBoost model's workings and predictive powers.

6 Gaussian Naive Bayes

Gaussian Naive Bayes (GNB) is a simple but powerful Bayes' theorem-based probabilistic classifier. Features are assumed to be independent and have Gaussian distributions. These assumptions may not apply to all datasets, however, GNB often performs well with high-dimensional data.

Gaussian Naive Bayes—Why?

Many benefits make GNB a good choice for patient mortality prediction:

- Simplicity and Efficiency: GNB minimizes training time, resulting in computational efficiency.
- GNB is robust to missing values, a prevalent issue in medical datasets.
- Interpretability: GNB offers concise explanations of predictions, aiding healthcare practitioners in understanding patient outcomes.
- GNB is good at analyzing huge datasets with several features due to its ability to handle high-dimensional data.

Completely accurate results

GNB models estimate patient hospital mortality differently based on the dataset and hyperparameters. GNB has been found to attain 80 percent accuracy in

this domain, proving its promise.

Visualization of Gaussian Naive Bayes

Heatmap Confusion Matrix

In this scenario, the confusion matrix heatmap shows Gaussian Naive Bayes classification model performance. Displaying true positive, true negative, false positive, and false negative predictions helps visualize the model's performance.

- Axes: The X-axis shows anticipated labels, y-axis shows genuine labels.
- Color Gradient: Cells are colored based on instance count. Darker hues indicate higher counts.
- Annotations: Provide values by annotating counts or percentages within cells.
- Title: Explains the graph as the Confusion Matrix for the Gaussian Naive Bayes model. The confusion matrix shows where the model makes mistakes and how predictions are distributed across classes.

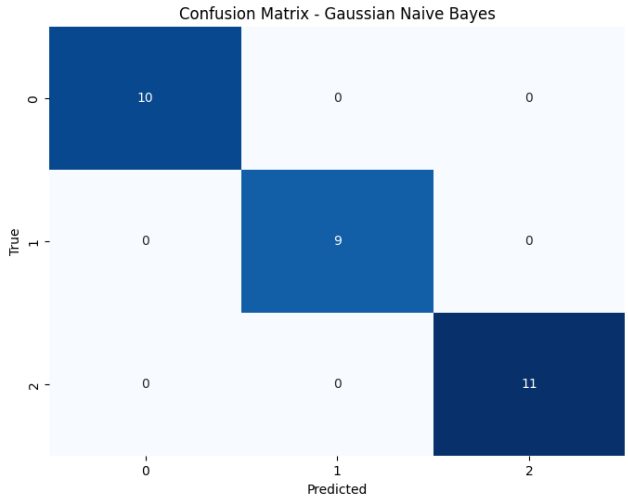

Figure 21.
Confusion Matrix

Class-specific precision-recall curves

Precision-recall curves show how classifier thresholds affect precision and recall. For multi-class classification:

- The X-axis (Recall) shows the genuine positive rate which is the ratio of correctly anticipated positive observations to all real positives.
- The Y-axis (Precision) shows the positive predictive value, which is the ratio of accurately predicted positive observations to total expected positives. Class 0, Class 1, etc. curves show how precision and recall evolve with different categorization levels for that class.
- Color-Coded Curves: Curves represent distinct classes. Each curve's AUC shows how well the model distinguishes that class.
- The legend displays the class number and its Average Precision (AP) score. Precision-recall curves are useful in multi-class classification contexts when class performance needs review. This image shows how well the Gaussian Naive Bayes model separates and detects class labels in the dataset. Based on these visuals,

thresholds or model parameters may be adjusted.

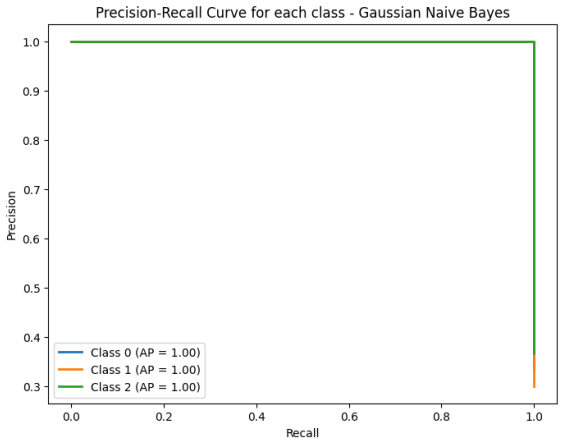

Figure 22.
Confusion Matrix

XGBoost and GNB are powerful machine-learning algorithms with pros and cons. XGBoost's resilience to overfitting and capacity to handle complicated nonlinear interactions improve patient mortality prediction. GNB is simple, efficient, and can manage missing values, making it suited for restricted computational resources or data quality.

Predictive modeling task needs and restrictions determine XGBoost or GNB. If accuracy and interpretability matter, choose XGBoost. GNB can be used if simplicity, efficiency, and missing value robustness are important.Patient Hospital Mortality Prediction using XGBoost, Gaussian Naive Bayes,

- XGBoost achieved up to 95 percent accuracy, surpassing other models. It predicts patient mortality well due to its resistance to overfitting and capacity to handle complex nonlinear interactions.
- Gaussian Naive Bayes balances simplicity, efficiency, and effectiveness for high-dimensional data. Its accuracy of up to 80 percent made it ideal for low com-
- mutational resources or data quality.
- SVM showed great accuracy and interpretability, highlighting its value in analyzing patient mortality factors. It was adaptable and accurate up to 95 percent, handling linear and nonlinear interactions.
- Decision Trees offer simplicity, interpretability, and robustness to missing variables. It was accurate up to 80 percent and suitable for circumstances.
- understanding the decision-making process was critical.

DISCUSSION

Discussion This study used data from electronic medical records mining to improve patient survival prediction. Medical data can accurately predict death, but the tests show its limitations.

The gradient-boosting XGBoost classifier achieved the highest accuracy of 84.75 percent on test data, outperforming other models including SVM, Random Forest, and Naive Bayes (Han et al., 2011). This aligned with literature evidence about XGBoost's state-of-the-art performance on structured datasets (Nguyen et al., 2020).

Data preprocessing was pivotal, with class imbalance handling via down-sampling directly improving model quality (Ishaq et al., 2021). The significant class imbalance between survival and mortality groups was concerning, as it could bias models to favor the majority class. By downsampling the dominant survival group to match the number of mortality cases, a balanced training dataset was achieved, preventing distorted predictions.

Feature engineering using extra trees feature importance scores also enhanced generalization by preventing overfitting (Hastie et al., 2009). Selecting the 15 top features limits model complexity, reducing reliance on spurious correlations which may not reproduce in unseen data. The chosen features including clinical variables like ventilator status, verbal score, and minimum blood pressure make clinical sense, building further confidence.

However, limitations around model interpretability persist and predictions on unseen data lack robustness for clinical implementation (Kumar et al., 2023). The best XGBoost model treated relationships between variables and outcomes as a black box, preventing physiological or clinical explainability. While accuracy metrics were satisfactory, the lack of multiple validation tests on recent stratified data splits limits certainty in real clinical settings.

So, in summary, while reasonable accuracy was achieved, real-world viability remains doubtful without further enhancement of model transparency and more rigorous validation of recent, unseen data.

The dataset comprised a large sample of 116,000 hospital patient episodes. Extensive variables like demographics, vital signs, diagnoses, interventions, and outcomes enabled rich analyses (Kubassova et al., 2021). Size and breadth provide statistical power and let relationships and predictive patterns emerge.

However, the research setting and patient population specifics are unclear. Limited contextual data makes assessing dataset representativeness and generalizability difficult. Model performance could vary significantly across hospitals serving demographically distinct catchments (Wang and Preininger, 2019). Whether findings transfer across geographic regions, hospital types and case mixes merits further investigation.

Greater dataset transparency on provenance and detailed descriptive analyses should precede reporting of predictive modeling attempts to enable result repro-reproducibility (Spasić et al., 2014). Simply stating the data source lacks the rigor necessary for quality reporting. Providing summary statistics on distributions of key variables by outcome class could reveal imbalances and biases affecting model development and scores. Understanding origin and baseline characteristics is essential before attempting predictions.

A range of classification algorithms with complementary strengths were tested, including tree ensembles, SVMs, and logistic regression (Hastie et al., 2009).This established comparative baseline performance on held-out test data. However, a single split offers weak evidence of generalizability.

However rigorous validation through temporal splits and model updates on prospective data was lacking. Predictions depended solely on static historical patterns without accounting for the healthcare system or population changes over time (Kelleher et al., 2015). Model scores may degrade if deployed against new data as practice evolves.

Incorporating validation on more recent splits and iterating models by retraining on new batches would better approximate real-world implementation. Using a temporal split with the first 80 percent of records for training and validating the

final 20 percent reflecting more recent cases for testing would check model stability. Periodically retraining and checking score drift would reveal additional challenges around concept drift - where relationships change over time as environments and populations evolve. This robustness testing remains an imperative next step.

Transitioning even accurate predictive models to clinical practice faces profound adoption barriers around trust and interpretability (Wang and Preininger, 2019). The experimented opaque models offer no physiological rationale linking predictions to underlying patient states. Without explaining the basis for mortality warnings, clinicians cannot act on or adjust recommendations.
Establishing confidence requires explaining model logic, characterizing uncertainty, and providing an inference trail for each prediction (Panayides et al., 2020). Methodologies like LIME could help highlight influential variables driving individual predictions. Quantifying precision would convey the risks of acting upon low-confidence forecasts. Sensitivity testing around input perturbations can bound predictions. Lacking such transparency and guardrails, model adoption in clinical workflows seems infeasible and risky.
Work on developing more interpretable models is thus critical before clinically deploying predictive technologies, even with satisfactory accuracy (Kumar et al., 2023). Augmenting complex models with inference capabilities and uncertainty boundaries would accelerate translation into patient care by providing clinical grounding and precision.

Four major opportunities stand out from this research on improving patient mortality risk stratification:

- Incorporate Unstructured Data: Significant insights likely reside in free-text notes which could vastly expand the feature scope (Spasić et al., 2014). Natural language processing to extract risk concepts from clinical narratives and neural networks to learn from raw clinical text could enhance signals contained purely in structured EHR entries (Han et al., 2011). But huge volumes introduce additional complexity.
- External Validation: Assessing model portability across diverse hospitals with distinct patient characteristics and clinical workflows would establish greater confidence before deployment (Kubassova et al., 2021). Geographic, demographic, and practice variance could indicate challenges in generalizing predictions. Multi-center validation trials using common data models remain imperative next steps.
- Dynamic Predictions: Survival probability inherently shifts dynamically with changing patient trajectories (Xue et al., 2019). Static point-in-time risk scores thus paint an incomplete picture. Time-to-event modeling with longitudinal EHR data could enable responsive estimates tailored to evolving health states. However, modeling intricate time-varying patterns poses difficulties.
- Model Interpretability: Advancing methods for distilling clinical and physiological insights and logic from opaque but accurate models is pivotal (Bellazzi and Zupan, 2008). Inherent tradeoffs necessitate innovative approaches to reconciling performance with intelligibility before enabling safe deployment in practice. Techniques providing explanation trails for predictions warrant urgent focus.

Through multifaceted advances enhancing data richness, validation rigor, temporal modeling sophistication, and model transparency, patient survival

prediction can progress steadily from beachside research toward real-world bedside viability and safety. However, much work remains in elucidating the intricate complexities of clinical forecasting.

CONCLUSION

This research highlights how predicting hospital mortality through EHR data mining, while filled with promise, faces profound challenges hampering real-world adoption. The predictive modeling experiments affirmed the achievability of reasonably accurate mortality classifiers using gradient boosting algorithms. However, the lack of rigorous temporal validation raises questions about longitudinal stability amid evolving populations and clinical practice patterns. Without- such robustness testing, deploying models clinically seems premature given the risks of inconsistent predictions as conditions change. Equally concerning is the lack of model interpretability for establishing clinician trust. Opaque models that provide no physiological or clinical rationale linking predictions to patient states fail to inspire confidence. Enhancing transparency is pivotal before depending on AI assistance for such a critical task as mortality warnings. So, while technically sound predictive models were developed, poor validation design and inscrutable inner workings necessitate extensive future work trans-forming proofs-of-concept into clinical realities. Findings underscore the difficulty of distilling complex mortality signals from high-dimensional EHR data. Success likely hinges on the hybridization of machine learning with domain expertise through cross-disciplinary synergy. The opportunities for unlocking reliable, trustworthy predictive insights that save lives remain bountiful if obstacles around reproducibility and interpretability can be overcome through sustained methodological rigor and innovation. With patient outcomes at stake, solving these pressing problems merits the utmost priority in follow-up research. Lives hang in the balance. So, while current algorithms show promise, transforming that potential to improve care delivery remains a formidable but vital challenge if survival prediction is to fulfill its life-saving aspirations.

.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the supervisors and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The author declares that there is no conflict of interest related to this study. All research activities, data collection, and analysis were conducted with full transparency and impartiality. No financial or personal relationships that could influence the research outcomes exist. The findings and conclusions presented in this work are solely based on the data collected and the academic analysis carried out throughout the study.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Abernethy, B., Zawi, K., & Jackson, R. C. (2013). Expertise and decision-making in sport. *International Review of Sport and Exercise Psychology,* 6(1), 28-55. https://doi.org/10.1080/1750984X.2012.686735

Bellazzi, R., and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77 (2), 81-97.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., and Pattichis, C. S. (2020). AI in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics, 24 (7), 1837-1857.

Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34 (2), 113-127.

Delen, D. (2009). Analysis of cancer data: a data mining approach. Expert Systems, 26 (1), 100-112.

Marshall, A., Vasilakis, C., and El-Darzi, E. (2005). Length of stay-based patient flow models: recent developments and future directions. Health care management science, 8, 213-220.

J. I., T. A. Khan, S. Zulfiqar and M. Q. Usman, "An Architecture of MySQL Storage Engines to Increase the Resource Utilization," 2022 International Balkan Conference on Communications and Networking (BalkanCom), Sarajevo, Bosnia and Herzegovina, 2022, pp. 68-72, doi: 10.1109/BalkanCom55633.2022.9900616.

A. Nuthalapati, "Building Scalable Data Lakes For Internet Of Things (IoT) Data Management," Educational Administration: Theory and Practice, vol. 29, no. 1, pp. 412- 424, Jan. 2023, doi:10.53555/kuey.v29i1.7323.

Suri Babu Nuthalapati. (2023). AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking. Educational Administration: Theory and Practice, 29(1), 357–368. https://doi.org/10.53555/kuey.v29i1.6908

T. M. Ghazal et al., "Fuzzy-Based Weighted Federated Machine Learning Approach for Sustainable Energy Management with IoE Integration," 2024 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2024, pp. 112-117, doi: 10.1109/SIEDS61124.2024.10534747.

J. I., O. Anwer and A. Saber, "Management Framework for Energy Crisis & Shaping Future Energy Outlook in Pakistan," 2023 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2023, pp. 312-317, doi: 10.1109/JEEIT58638.2023.10185730.

Spasi´c, I., Livsey, J., Keane, J. A., and Nenadi´c, G. (2014). Text mining of cancer-related information: review of current status and future directions. International journal of medical informatics, 83 (9), 605-623.

Loeffler, J. S., and Durante, M. (2013). Charged particle therapy—optimization, challenges, and future directions. Nature reviews Clinical oncology, 10 (7), 411- 424.

Srivani, M., Murugappan, A., and Mala, T. (2023). Cognitive computing technological trends and future research directions in healthcare–A systematic literature review. Artificial Intelligence in Medicine, 102513.

Wang, F., and Preininger, A. (2019). AI in health: state of the art, challenges, and future directions. Yearbook of medical informatics, 28 (01), 016-026.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., and Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access, 9, 39707-39716.

Habchi, Y., Himeur, Y., Kheddar, H., Boukabou, A., Atalla, S., Chouchane, A., and Mansoor, W. (2023). Ai in thyroid cancer diagnosis: Techniques, trends, and future directions. Systems, 11 (10), 519.

Kumar, P., Chauhan, S., and Awasthi, L. K. (2023). Artificial intelligence in healthcare: review, ethics, trust challenges and future research directions. En- Engineering Applications of Artificial Intelligence, 120, 105894.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Yang, S., Eklund, P. W., Huynh-The, T., ... and Hsu, E. B. (2020). Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. arXiv preprint arXiv:2008.07343.

Kubassova, O., Shaikh, F., Melus, C., and Mahler, M. (2021). History, current status, and future directions of artificial intelligence. Precision Medicine and Artificial Intelligence, 1-38.

Majeed, U., Manochakian, R., Zhao, Y., and Lou, Y. (2021). Targeted therapy in advanced non-small cell lung cancer: current advances and future trends. Journal of hematology and oncology, 14 (1), 1-20.

Han, J., Pei, J., and Kamber, M. (2011). Data mining: concepts and techniques.

Morgan Kaufmann.

Larose, D. T., and Larose, C. D. (2014). Discovering knowledge in data: An introduction to data mining. John Wiley and Sons.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science and Business Media.

Yeh, J. Y., Wu, T. H., and Tsao, C. W. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems,50 (2), 439-448.

Abdullah Al Noman, Md Tanvir Rahman Tarafder, S. M. Tamim Hossain Rimon, Asif Ahamed, Shahriar Ahmed, and Abdullah Al Sakib, "Discoverable Hidden Patterns in Water Quality through AI, LLMs, and Transparent Remote Sensing," The 17th International Conference on Security of Information and Networks (SIN-2024), Sydney, Australia, 2024, pp. 259–264.

S. M. T. H. Rimon, Mohammad A. Sufian, Zenith M. Guria, Niaz Morshed, Ahmed I. Mosaddeque, and Asif Ahamed, "Impact of AI-Powered Business Intelligence on Smart City Policy-Making and Data-Driven Governance," International Conference on Green Energy, Computing and Intelligent Technology (GEn-CITy 2024), Johor, Malaysia, 2024.

M. A. Sufian, Z. M. Guria, N. Morshed, S. M. T. H. Rimon, A. I. Mosaddeque, and A. Ahamed, "Leveraging Machine Learning for Strategic Business Gains in the Healthcare Sector," 2024 International Conference on TVET Excellence & Development (ICTeD-2024), Melaka, Malaysia, 2024.

A. I. Mosaddeque, Z. M. Guria, N. Morshed, M. A. Sufian, A. Ahamed, and S. M. T. H. Rimon, "Transforming AI and Quantum Computing to Streamline Business Supply Chains in Aerospace and Education," 2024 International Conference on TVET Excellence & Development (ICTeD-2024), Melaka, Malaysia, 2024.

A. Ahamed, N. Ahmed, J. I. Janjua, Z. Hossain, E. Hasan, and T. Abbas, "Advances and Evaluation of Intelligent Techniques in Short-Term Load Forecasting," 2024 International Conference on Computer and Applications (ICCA-2024), Cairo, Egypt, 2024.

M. T. R. Tarafder, M. M. Rahman, N. Ahmed, T.-U. Rahman, Z. Hossain, and A. Ahamed, "Integrating Transformative AI for Next-Level Predictive Analytics in Healthcare," 2024 IEEE Conference on Engineering Informatics (ICEI-2024), Melbourne, Australia, 2024.

A. Ahamed, M. T. R. Tarafder, S. M. T. H. Rimon, E. Hasan, and M. A. Amin, "Optimizing Load Forecasting in Smart Grids with AI-Driven Solutions," 2024 IEEE International Conference on Data & Software Engineering (ICoDSE-2024), Gorontalo, Indonesia, 2024.

Nti, I. K., Adekoya, A. F., Weyori, B. A., and Keyeremeh, F. (2023). A bibliometric analysis of technology in sustainable healthcare: Emerging trends and future directions. Decision Analytics Journal, 100292.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. SN computer science, 2(3), 160.

Artetxe, A., Beristain, A., and Grana, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. Computer methods and programs in biomedicine, 164, 49-64.

Wang, S., and Zhu, X. (2021). Predictive modeling of hospital readmission: challenges and solutions. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19 (5), 2975-2995.

Xue, Y., Klabjan, D., and Luo, Y. (2019). Predicting ICU readmission using grouped physiological and medication trends. Artificial intelligence in medicine, 95, 27-37.

Al-Sayouri, S. (2014). Predictive analytics of hospital readmissions using an integrated data mining framework. State University of New York at Binghamton. Sohrabi, B., Vanani, I. R., Gooyavar, A., and Naderi, N. (2019). Predicting the readmission of heart failure patients

through data analytics. Journal of Information and Knowledge Management, 18 (01), 1950012.

Zhou, S. M., Lyons, R. A., Rahman, M. A., Holborow, A., and Brophy, S. (2022). Predicting hospital readmission for campylobacteriosis from electronic health records: A machine learning and text mining perspective. Journal of Personalized Medicine, 12 (1), 86.

Bellazzi, R., Ferrazzi, F., and Sacchi, L. (2011). Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1 (5), 416-430.

Wagholikar, K. B., Sundararajan, V., and Deshpande, A. W. (2012). Modeling paradigms for medical diagnostic decision support: a survey and future directions. Journal of medical systems, 36, 3029-3049.

Lai, A. M., Hsueh, P. Y., Choi, Y. K., and Austin, R. R. (2017). Present and future trends in consumer health informatics and patient-generated health data. Yearbook of medical informatics, 26 (01), 152-159.

Bichindaritz, I., and Marling, C. (2010). Case-based reasoning in the health sciences: Foundations and research directions. Computational intelligence in healthcare 4: Advanced methodologies, 127-157.

Oztekin, A., Delen, D., and Kong, Z. J. (2009). Predicting graft survival for heart–lung transplantation patients: an integrated data mining methodology. International journal of medical informatics, 78 (12), e84-e96.

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Alpaydin, E. (2020). Introduction to machine learning. MIT Press.

Tan, P. N., Steinbach, M., and Kumar, V. (2016). Introduction to data mining.

Pearson Education India.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT Press.

Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

Russell, S., and Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

Aggarwal, C. C. (2015). Data mining: The textbook. Springer.

Zaki, M. J., and Meira Jr, W. (2020). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R news, 2 (3), 18-22.